

Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Human-Computer Interaction

Clone Testing: An Investigation of Trust and Suspicion in Identity Recognition During Human-Machine Dialogue

Master's Thesis

Shriya Singh

1. Referee: Prof. Dr. Benno Stein
2. Referee: Prof. Dr. Maria Wolters

Submission date: October 5, 2026

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work. Generative AI and agentic automation were used as assistive tools for grammatical correction, language refinement, structural revision, LaTeX and diagram support, and code development and debugging. All generated outputs were critically reviewed, verified, and revised by the author, who remains responsible for the thesis.

Weimar, Thuringia, October 5, 2026


.....
Shriya Singh

Abstract

Socially capable language-model agents can sustain plausible personal conversations and can be used for impersonation, manipulation, and fraud, making it important to understand how people judge the authenticity of online partners [Schmitt and Flechais, 2024]. Prior Turing-style studies prominently report final classification outcomes [Jannai et al., 2023, Jones and Bergen, 2024, 2025], leaving less evidence about the probing strategies and trust dynamics through which people reach and revise those judgments.

We conducted a two-phase Turing-style study in which seven trainers roleplayed personas that informed conversational clones, followed by paired human-and-clone conversations with ten independent evaluators who recorded suspicion throughout each interaction. Evaluators correctly classified 19 of 20 conversations; suspicion trajectories often reversed before the final decision, while excessive regularity, uniformly polished responses, and weak contextual adaptation were recurring signs of artificiality. These findings show that conversational identity is inferred through relations among cues across turns, and that the consistency used to sustain a persona can also expose its artificial construction.

Contents

1	Introduction	1
2	Related Work	4
2.1	AI-Mediated Deception and Online Impersonation	4
2.2	Turing-Style Identity Detection	5
2.3	Interactional Evidence in Identity Judgment	6
2.4	Personas, Role-Play, and Thesis Positioning	8
3	Conceptual Framework	10
3.1	Conversational Identity as Interactional Construction	10
3.2	Suspicion Trajectories and Conversational Recovery	11
3.3	Behavioral Elicitation and Trainer-Informed Clones	14
4	The Conversational Signal Table	17
4.1	Purpose and Definition	17
4.2	Signal Domains and Trajectories	18
4.3	Use and Evidential Status	21
5	Study Design and Methods	23
5.1	Study Design	23
5.2	Pilot Study and Instrument Refinement	26
5.3	Phase 1: Trainer Study	26
5.4	Clone Construction and Trainer Feedback	31
5.5	Phase 2: Evaluator Study	32
5.6	Data Preparation and Analysis	37
5.7	Ethical and Data-Protection Considerations	40
6	System Design and Prototype Development	41
6.1	System Architecture	41
6.2	Phase 1 Prototype	43
6.3	Trainer-Informed Clone Configuration	43
6.4	Phase 2 Prototype	44

7	Phase 1: Trainer Study Findings	47
7.1	Variation in Trainer Behavior	47
7.2	Role Enactment and Persona Drift	49
7.3	Signals of Artificiality	50
7.4	Consequences for Clone Construction	52
7.5	Limitations	53
8	Phase 2: Evaluator Study Results	54
8.1	Identity-Detection Outcomes	54
8.2	Suspicion Trajectories	56
8.3	Evaluator Strategies and Reported Cues	58
8.4	Variation Across Profiles	60
9	Discussion	63
10	Conclusion	69
A	Study Instruments	72
B	Character Archetypes	75
C	Pass-as-Human Test Specification	77
D	Data Codebook	79
	Bibliography	81

List of Figures

4.1	Conceptual organization of the Conversational Signal Framework.	20
5.1	Overall two-phase study design.	25
5.2	Phase 1 archetype-selection interface.	28
5.3	Phase 1 trainer process.	29
5.4	Annotated Phase 1 trainer-study setup.	30
5.5	Clone construction and trainer-review activity.	33
5.6	Phase 2 evaluator process.	35
5.7	Annotated Phase 2 evaluator-study setup.	36
6.1	Experimental setup and response routing.	42
6.2	Representative states of the Phase 1 prototype.	43
6.3	Evaluator and control interfaces of the Phase 2 prototype.	45
8.1	Evaluator suspicion trajectories by partner condition.	57
A.1	Pilot study setting and presentation.	74

List of Tables

4.1	Conversational signal domains established before Phase 1. . . .	19
5.1	Role archetypes offered in Phase 1.	27
5.2	Planned and executed Phase 2 sampling.	34
7.1	Descriptive characteristics of Phase 1 trainer behavior.	48
7.2	Surface-level Phase 1 trainer descriptors.	48
7.3	Phase 1 findings and consequences for clone construction.	52
8.1	Phase 2 outcomes by partner condition.	55
8.2	Paired Phase 2 condition comparisons.	55
8.3	Themes in evaluators' written explanations.	59
8.4	Descriptive comparison of the four profiles used in Phase 2. Mean words are reported per trainer or partner turn. Confidence and realism are Phase 2 evaluator ratings on their respective study scales.	60
8.5	Cross-case comparison for T1 and T3.	61
8.6	Cross-case comparison for T6 and T7.	62
9.1	Empirical revision of the Conversational Signal Table.	66
B.1	Available archetypes and Phase 1 selection.	76
C.1	Formative pass-as-human test specification, cases 1–8.	78
C.2	Formative pass-as-human test specification, cases 9–15.	78
C.3	Examples of profile-specific target response styles in the pass-as-human workbook.	78

Acknowledgements

I thank the seven trainers, five pilot participants, and ten evaluators who gave their time to this study. Together, these twenty-two participants made the empirical work possible. I am grateful to Prof. Dr. Benno Stein and Prof. Dr. Maria Wolters for their scholarly guidance and for serving as the thesis referees. I am especially grateful to Marcel Gohsen, whose sustained day-to-day supervision helped carry the project from its early methodological questions through prototype development and the two empirical phases. My sincere regards to Prof. Dr. Eva Hornecker and Margarita Osipova for offering valuable advice right around the time I started planning the thesis.

Chapter 1

Introduction

Large language models are increasingly capable of sustaining conversations that appear socially responsive, personal, and human. In text-based interaction, people cannot rely on physical presence or conventional identity cues. They instead interpret language, response timing, conversational continuity, personal specificity, affect, and reciprocity. Recent Turing-style experiments demonstrate that prompted language models can be difficult to distinguish from human interlocutors in short conversations [Jannai et al., 2023, Jones and Bergen, 2024, 2025]. The relevant question is therefore no longer only whether a machine can produce plausible language, but how a person decides who or what is producing it.

This uncertainty has consequences beyond experimental demonstrations. Generative AI can support realistic, personalized, and automated forms of social engineering across text, image, voice, and video. In messaging platforms and social media, a deceptive conversational persona may imitate a person, develop credibility over multiple exchanges, and adapt its responses when the recipient becomes hesitant or suspicious [Schmitt and Flechais, 2024]. Experimental phishing research likewise shows that language models can create personalized deceptive messages while reducing the effort required to produce them [Heiding et al., 2024]. Impersonation, phishing, fraud, and manipulative relationships all depend in part on whether a recipient accepts the apparent identity of the other party. Understanding how people form, question, lose, and recover that acceptance is therefore relevant to digital safety.

The Turing test provides the traditional experimental paradigm for studying whether people can distinguish machine behavior from human behavior [Turing, 1950]. Contemporary studies commonly report the participant’s final classification and the proportion of conversations in which an agent was judged to be human [Jannai et al., 2023, Jones and Bergen, 2024]. Some studies also examine reported strategies, but the developing behavior of the human judge

remains comparatively underexplored. A final answer does not show which questions the participant used as probes, which responses produced suspicion, whether later evidence reversed an earlier impression, or whether conversational repair restored trust after a breakdown. These processes require the interaction itself to be studied as a sequence of evolving judgments.

This thesis addresses that gap through a two-phase experiment grounded in Turing-style identity testing. In Phase 1, seven trainers role-played selected personas while interacting with an AI investigator. Their conversations, think-aloud observations, and retrospective accounts provided stylometric, semantic, and pragmatic evidence for constructing trainer-informed conversational clones. In Phase 2, ten independent evaluators each completed one conversation with a human trainer and one with the corresponding clone. The human condition was implemented through a Wizard-of-Oz arrangement in which the trainer replied through a separate operator interface while the evaluator continued to use the same participant-facing chat. At the boundary between rounds, response authority changed between the trainer and the clone according to the assigned condition order. The evaluator interface recorded final identity judgments together with turn-level suspicion reports, allowing trust dynamics, authenticity shift events, probing strategies, breakdowns, and conversational recovery to be examined. The term *Clone Testing* refers to this experimental configuration; it does not claim the invention of an alternative to the Turing-test paradigm.

The evaluators correctly classified 19 of the 20 conversations, but their suspicion trajectories frequently changed direction before the final decision. They tested their partners through follow-up questions, references to earlier turns, topic changes, personal questions, and direct challenges. Excessive regularity, uniformly polished responses, and affect that did not adapt to context were repeatedly interpreted as artificial. Uneven response depth, imperfect but relevant recall, situated specificity, and contextual recovery were more often interpreted as human. These findings suggest that conversational identity is inferred through relations among cues across multiple turns rather than through a single reliable tell. They also reveal a central tension in persona construction: consistency can preserve a recognizable character while simultaneously exposing its artificial construction.

Research questions.

The study investigates human behavior during conversational identity inference through the following research questions:

1. Which conversational signals contribute to shifts in trust and suspicion during human–AI identity judgments?

2. Which probing strategies do participants use when attempting to determine whether a conversational partner is human or artificial?
3. How do suspicion trajectories develop, reverse, recover, or stabilize across interactions with human trainers and trainer-informed conversational clones?
4. Which differences between human and clone interactions do evaluators interpret as evidence of authenticity or artificiality?

Contributions.

This thesis makes five contributions:

1. It designs a two-phase experiment, grounded in the Turing-test paradigm, for investigating human behavior during conversational identity judgments.
2. It reports a Phase 1 study that collected conversational, behavioral, think-aloud, and retrospective data from seven human trainers while they role-played personas and were investigated by an AI conversational partner.
3. It develops trainer-informed prototype clones from stylometric, semantic, and pragmatic signals identified in the Phase 1 material.
4. It reports a Phase 2 study in which ten independent evaluators interacted with human trainers and their corresponding conversational clones through a visually consistent interface. A Wizard-of-Oz operator arrangement implemented the human condition, while turn-level suspicion reports and final judgments captured the evaluators' developing identity inferences.
5. It provides an in-depth qualitative analysis of the evaluators' probing strategies, authenticity shift events, reported cues, breakdowns, and conversational recoveries, and organizes the resulting observations through the Conversational Signal Table.

Chapter 2

Related Work

This chapter situates the study within research on AI-mediated deception, Turing-style identity detection, trust dynamics, conversational repair, game-based interaction, and prompt-conditioned personas. The central concern is not how conversational systems should be evaluated in general, but how people behave when they must determine whether a conversational partner is human or artificial.

2.1 AI-Mediated Deception and Online Impersonation

Deception in digital communication predates generative AI. Phishing and other forms of social engineering already rely on misleading representations and relationships. Generative AI changes the scale and adaptability of this problem: it can produce realistic content, personalize messages to a target, and automate parts of an interaction that previously required sustained human labor [Heiding et al., 2024, Schmitt and Flechais, 2024].

These capabilities extend beyond isolated fraudulent messages. A conversational agent can maintain an invented persona, respond to questions, adjust its language when a recipient becomes hesitant, and continue developing credibility across multiple exchanges. Schmitt and Flechais describe realistic content generation, advanced personalization, and automated attack infrastructure as three ways in which generative AI can amplify social engineering. Their analysis also identifies the possibility of conversational systems adapting their approach when a target displays suspicion [Schmitt and Flechais, 2024].

This threat gives conversational identity detection practical importance. Technical detection systems remain necessary, but recipients also make their own judgments about whether an apparent conversational partner is authen-

tic. Understanding those human judgments requires attention to the questions people ask, the cues they interpret, the moments at which suspicion changes, and whether later responses restore an earlier loss of trust.

2.2 Turing-Style Identity Detection

The imitation game introduced by Turing [1950] reframed machine intelligence as an interrogator’s attempt to distinguish machine-generated responses from human responses. Turing predicted that an average interrogator would have no more than a 70minutes of questioning [Turing, 1950]. The paradigm therefore places a human judge, rather than an automated performance metric, at the center of the experiment.

Recent studies have applied this paradigm to contemporary language models. In *Human or Not?*, more than 1.5 million users held anonymous two-minute conversations with either another person or a language model and then classified the partner. Overall identification accuracy was 68enabled participation at a scale that would be difficult to obtain through a conventional laboratory study.

In a controlled and preregistered experiment, Jones and Bergen [2024] found that GPT-4 was judged to be human in 54that judges relied substantially on stylistic and socio-emotional cues rather than only testing factual intelligence. A later three-party formulation reported that some frontier models were selected as human at rates comparable to or exceeding the actual human interlocutor [Jones and Bergen, 2025]. These results do not establish that language models are universally indistinguishable from people. They show that, under particular conversational conditions and prompts, judgments may approach chance or may favor a model over the human interlocutor.

Humanness judgments are also affected by the cues through which a chatbot is presented. Experimental work has shown that visual, identity, and conversational cues can alter perceived humanness, with conversational interactivity contributing to that perception [Go and Sundar, 2019]. This reinforces the need to hold the participant-facing interface constant when comparing human and artificial partners.

The primary reported outcome in these studies is whether the judge’s final classification was correct. Some studies also report the reasons or strategies participants described, but a terminal judgment cannot reconstruct the full process by which the decision developed. It does not show when suspicion first appeared, which question was intended as a test, whether a response temporarily restored confidence, or why a cue changed meaning after being repeated. This distinction between classification outcome and judgment process

motivates the present study.

2.3 Interactional Evidence in Identity Judgment

Games as structured interaction settings.

Games can create a setting in which concealment, inference, persuasion, and strategic questioning arise from the activity rather than from an isolated instruction. Social-deduction environments are particularly relevant because participants must form and revise beliefs about other players while the interaction unfolds. For example, Bailis et al. [2024] study language models through Werewolf tournament play, where turn-taking, persuasion, and concealed roles form part of the interaction.

Narrative environments provide a related example. LIGHT places human and model-controlled characters within a configurable text world in which dialogue, persona, action, and social setting are represented together [Urbanek et al., 2019]. It therefore provides an example of language behavior being studied within a role and interactional context rather than through disconnected questions.

The present study uses a game structure for a different purpose. The game creates conditions in which human participants actively investigate a partner, test competing identity hypotheses, and commit to a final judgment. Its value lies in eliciting observable human behavior under uncertainty rather than in producing a score for general agent performance.

Trust, suspicion, and sequential impression updating.

Trust research distinguishes relatively stable predispositions from attitudes that change as a person gains experience with a particular system [Hoff and Bashir, 2015, Lee and See, 2004]. Much of this literature concerns appropriate reliance on automation. The present study concerns a narrower question: whether a participant provisionally accepts or questions the apparent identity of a conversational partner. The broader literature nevertheless establishes that trust is dynamic and can be revised as new evidence becomes available.

Research on sequential belief updating provides a basis for understanding that revision. The weight assigned to new information depends on its order, the way evidence is processed, and the characteristics of the judgment task [Hogarth and Einhorn, 1992]. New information does not automatically replace a first impression merely because it appears later. However, information perceived as highly diagnostic of a person’s character can substantially revise or reverse an earlier evaluation [Cone and Ferguson, 2015]. Impression updating can also be asymmetric because negative and positive evidence may differ in their perceived diagnostic value [Mende-Siedlecki et al., 2013].

These findings support a temporal account of conversational identity inference. Participants may begin with a provisional impression, encounter a response that increases suspicion, and then revise that suspicion after a later clarification or contextually appropriate reply. A cue may also change meaning through repetition. A typographical error may initially appear human but become artificial when it recurs too systematically. The relevant object of analysis is therefore not only the final classification, but the trajectory through which the classification was reached.

In this thesis, that process is represented through *suspicion trajectories*. Individual movements of the suspicion slider are treated as *authenticity shift events*. Movement toward the artificial endpoint represents increasing suspicion or trust drift, while movement back toward the human endpoint may represent conversational recovery. Establishing that a specific repair caused such a recovery requires the slider event to be examined alongside the corresponding conversational turns.

Conversational breakdown and repair.

Conversation analysis describes repair as the practices through which speakers address problems in speaking, hearing, and understanding. These practices include clarification requests, repetition, correction, reformulation, and self-repair [Schegloff et al., 1977]. Grounding research similarly treats mutual understanding as something interlocutors establish and update together rather than something transmitted successfully in a single message [Clark and Brennan, 1991].

Breakdown and repair are relevant here as possible identity signals. In conversational-agent use, users frequently encounter a gap between the competence they expect and the system’s ability to recognize or recover from a problem [Luger and Sellen, 2016]. A repetitive or generic repair may increase suspicion if it fails to incorporate what was previously said. A situated repair may instead restore confidence by acknowledging the breakdown, recovering relevant context, or adapting the response appropriately.

Repair must therefore be analyzed relationally. The same clarification phrase cannot be classified as human or artificial in isolation. Its meaning depends on the preceding breakdown, the conversational history it incorporates, and whether the subsequent interaction demonstrates recovery. This relationship between signal, context, and consequence informs the Conversational Signal Table presented in Chapter 4.

2.4 Personas, Role-Play, and Thesis Positioning

Persona-based conversational agents can be produced through prompting rather than model fine-tuning. Prompt-conditioned personas are editable and inspectable, but this does not make them complete psychological replicas of the people or groups they represent. Accounts of language-model behavior as role-play similarly caution against treating a generated performance as evidence of human-like internal characteristics [Shanahan et al., 2023]. Research on synthetic users reports interactions such as seeking feedback and testing assumptions, and positions these systems as augmentations to traditional persona representations rather than direct replacements [Gu et al., 2025].

This distinction is particularly important when a persona is derived from one person’s observed behavior. A prompt may preserve recurring features such as response length, vocabulary, repair style, topic preference, or affective expression. It may nevertheless fail to reproduce when that person departs from the pattern, responds unevenly, forgets something, or changes behavior because of the immediate conversational context.

Role-play introduces an additional layer. The Phase 1 trainers were not asked to reproduce their unrestricted everyday identities. They enacted socially recognizable archetypes within a bounded interaction. The resulting clones must therefore be understood as trainer-informed representations of those performances, not as replicas of the trainers as complete individuals.

The present study examines how third-party evaluators interpret these representations. The question is not whether the clones possess the psychological characteristics of the trainers. It is which features of the human and clone conversations evaluators treat as authentic, artificial, consistent, suspicious, or reparative.

Research gap.

The reviewed literature leaves three connected gaps.

First, Turing-style studies primarily report classification outcomes. They show whether a participant identified a partner correctly, but provide less evidence about the participant’s developing behavior during the interaction. The probing questions, hypothesis tests, reactions to breakdowns, and reasons for changing direction remain comparatively underexamined [Jannai et al., 2023, Jones and Bergen, 2024, 2025].

Second, terminal judgments compress a temporally developing process into one answer. They cannot show when trust drift occurred, whether suspicion later weakened, or whether a conversational repair produced recovery. Turn-aligned reports are needed to relate changes in suspicion to the messages and events that preceded them.

Third, most Turing-style comparisons concern whether an agent appears human in general. They do not usually compare an agent with the particular human performance from which its behavioral profile was derived. Such a comparison makes it possible to investigate how people respond to similarities and differences between a source performance and its prompt-conditioned representation [Jannai et al., 2023, Jones and Bergen, 2024, 2025].

Positioning of this thesis.

This thesis addresses these gaps through an experimental design grounded in the established Turing-test paradigm. Phase 1 records human behavior while trainers enact personas and respond to conversational investigation. Those records inform prompt-conditioned prototype clones. Phase 2 then places human trainers and their corresponding clones behind the same evaluator-facing interface, using a Wizard-of-Oz operator arrangement for the human condition.

The study combines final identity judgments with turn-level suspicion reports, interaction logs, and qualitative explanations. This makes it possible to analyze both whether an evaluator reached the correct conclusion and how that conclusion developed. The contribution is therefore not a new replacement for the Turing test. It is an empirical investigation of human probing, impression updating, trust dynamics, breakdown, repair, and identity judgment within Turing-test conditions.

Chapter 3

Conceptual Framework

This chapter defines the concepts used to analyze human behavior during the two Turing-style studies. The framework does not propose an alternative to the Turing test. It provides an analytical vocabulary for examining what happens between the beginning of a conversation and the participant’s final identity judgment.

Four ideas organize the analysis. First, conversational identity is treated as an interpretation developed through interaction rather than as a property revealed by one message. Second, trust and suspicion are treated as dynamic judgments that can change as new evidence appears. Third, breakdown and repair are examined as possible identity signals. Fourth, conversational signals are interpreted relationally, meaning that their significance depends on context, sequence, recurrence, and consequence.

Together, these ideas support the analysis of probing strategies, authenticity shift events, suspicion trajectories, trust drift, interactional repair, and conversational recovery.

3.1 Conversational Identity as Interactional Construction

Conversational identity is not communicated through vocabulary alone. It is constructed through timing, stance, grounding, affect, specificity, repair, and continuity across turns. Sociocultural-linguistic accounts similarly describe identity as something produced through local interaction and positioning rather than merely expressed as a stable internal property [Bucholtz and Hall, 2005].

This perspective is especially relevant in text-based identity detection. A participant cannot directly observe who produced a message. Instead, the par-

ticipant interprets traces of behavior and asks whether they form a coherent account of the partner. A response may appear human because it recalls an earlier detail, adapts its tone, admits uncertainty, or repairs a misunderstanding. The same response may appear artificial when its detail, polish, or emotional expression seems too regular or insufficiently connected to the local exchange.

Turing-style studies ultimately require a categorical decision, but the decision need not emerge categorically. Participants may begin uncertain, develop a provisional interpretation, test it through conversation, and revise it as additional evidence appears. Conversational identity inference is therefore treated here as an evolving interpretative process rather than only as a final classification event.

3.2 Suspicion Trajectories and Conversational Recovery

A final human-or-AI judgment captures the outcome of identity inference, but not the process through which that judgment develops. Preliminary observations suggested that participants rarely reached an immediate and unchanging conclusion. Instead, suspicion accumulated, weakened, recovered, or stabilized as additional conversational evidence became available. Trust and suspicion are therefore treated in this thesis as temporally evolving and interactionally situated rather than as static post-conversation assessments.

In this context, trust refers to the participant's provisional acceptance of the conversational partner's apparent identity. Suspicion refers to the active questioning of whether that apparent identity is authentic. This use of trust is narrower than the broader concept of trust in automation, which may include reliance, competence, predictability, or willingness to delegate a task [Hoff and Bashir, 2015, Lee and See, 2004]. The present study is concerned specifically with how confidence in a partner's apparent conversational identity changes during interaction.

This process can be understood as sequential impression updating. Each response supplies evidence that may reinforce, weaken, or reverse the participant's current judgment. Research on belief and impression updating shows that the effect of new information depends not only on whether it appears early or late, but also on its perceived diagnostic value and its relationship to previously encountered evidence [Cone and Ferguson, 2015, Hogarth and Einhorn, 1992, Mende-Siedlecki et al., 2013]. A later response does not automatically erase a first impression. However, information that appears particularly revealing of the partner's identity can cause an earlier interpretation to be reconsidered.

The study represents this temporal process through *suspicion trajectories*. An individual slider movement is treated as an *authenticity shift event*: a recorded change in the participant’s expressed belief about whether the partner is human or artificial. Movement toward the AI endpoint represents increasing suspicion or trust drift. Movement back toward the human endpoint after an earlier rise in suspicion represents conversational recovery. Where that reversal follows clarification, re-grounding, contextual adaptation, or another response to a preceding breakdown, it may indicate successful interactional repair.

This distinction is analytically important. A failed repair sequence, vague response, contradiction, or contextually inappropriate expression of affect may increase suspicion. A subsequent response that acknowledges the breakdown, recalls relevant context, or adapts appropriately may restore provisional trust in the partner’s apparent identity. Directional movement alone does not establish that repair caused the recovery. That interpretation requires the slider event to be examined alongside the corresponding conversational turns.

Suspicion trajectories therefore describe how identity judgments emerge, intensify, reverse, recover, or stabilize throughout an interaction. They allow conversational identity inference to be studied as a developing interpretative process rather than only as a terminal classification event.

Conversational signals as relational evidence.

A conversational signal is a feature of an interaction that a participant or researcher interprets as evidence about the partner. Potential signals include response timing, repetition, repair, grounding, specificity, affect, typographical variation, conversational continuity, and reactions to probing. These categories provide starting points for analysis, not a checklist in which each feature has a fixed meaning.

The same observable feature can support different judgments. A typographical error may initially appear human because it interrupts polished language, but it may later appear manufactured if the same kind of error occurs at a regular rate. A detailed autobiographical response may appear authentic when it answers the immediate question, yet artificial when every topic produces equally complete and characterful detail. A delayed response may indicate reflection, distraction, technical latency, or deliberately simulated pacing.

The analytical unit is therefore the relationship among the feature, its context, its recurrence, and its conversational consequence. A signal becomes meaningful through questions such as:

- What occurred immediately before the signal?
- Did the signal answer or avoid the participant’s probe?
- Did it incorporate shared conversational history?

- Was it repeated across contexts where different behavior might have been expected?
- Did the participant’s suspicion change after it occurred?

This relational treatment prevents the analysis from reducing humanness to a collection of surface markers. It also allows the same category to contain evidence associated with both humanness and artificiality. The Conversational Signal Table in Chapter 4 organizes these relationships and records where anticipated signals appeared, changed meaning, or remained unsupported.

Breakdown, repair, and conversational recovery.

Conversation analysis describes repair as the practices through which speakers address problems in speaking, hearing, and understanding. Clarification, repetition, correction, reformulation, and self-repair allow interlocutors to re-establish mutual understanding [Schegloff et al., 1977]. Grounding research similarly treats understanding as something participants establish and update together [Clark and Brennan, 1991].

In conversational identity inference, breakdown is relevant because it creates an opportunity for the partner to display how it handles difficulty. The initial problem may matter less than the response that follows it. A partner that acknowledges the misunderstanding, connects the repair to previous turns, and adjusts its subsequent behavior may appear more human. A partner that repeats a generic apology, ignores the problem, or resumes the same pattern without adaptation may increase suspicion. Breakdowns in conversational-agent use frequently become salient when the system cannot meet or recover from the expectations produced by its otherwise capable presentation [Luger and Sellen, 2016].

The thesis distinguishes three related events. A *breakdown* is a moment in which conversational coherence, grounding, or expectation is disrupted. *Interactional repair* is the partner’s attempt to address that disruption. *Conversational recovery* occurs when the subsequent exchange re-establishes sufficient coherence or credibility for suspicion to decrease. This distinction is consistent with trust-repair research, which treats repair as the process of rebuilding trust after a perceived violation [Baker et al., 2018]. Studies of conversational agents further show that repair strategies following imperfect interventions can affect user acceptance and the interaction experience [Do et al., 2023].

Recovery is not inferred from a well-formed response alone. It requires evidence that the participant’s interpretation changed following the repair. In Phase 2, this can be investigated by aligning conversational turns with authenticity shift events. A movement back toward the human endpoint after a repair provides evidence of recovery, although the small sample requires such

patterns to be interpreted descriptively rather than as a general repair success rate.

3.3 Behavioral Elicitation and Trainer-Informed Clones

Role-play as behavioral elicitation.

Phase 1 uses role-play as a bounded means of eliciting conversational behavior. Trainers selected socially recognizable archetypes and interpreted them through their own language, experiences, and interactional choices. The archetypes supplied a common scaffold without prescribing every attitude, memory, or response. This use follows HCI design work that treats role-play as a means of making interactional situations experiential and available for reflection [Simsarian, 2003].

This distinction matters because the study did not attempt to reproduce the trainers as complete individuals. It recorded how particular trainers enacted particular personas under conversational investigation. The resulting evidence therefore combines features of the role, the trainer’s habitual behavior, and the immediate interaction.

Linguistic experience also shaped how at least one trainer approached the role. During the session, the trainer explained that she was thinking through parts of the interaction in Spanish and reflected that the apparent personality of a language model can change when the language of interaction changes. She also questioned whether grammatical gender in another language might affect how an agent describes itself. This observation illustrates that role enactment is mediated not only by the assigned archetype, but also by the trainer’s linguistic repertoire and expectations of language-specific behavior. Because the formal sessions were conducted in English, the study cannot determine whether the same trainer-informed profile would be perceived differently when enacted in another language.

Role-play also created observable tension between archetype adherence and contextual adaptation. A trainer could maintain a recognizable role while responding unevenly across topics, departing from an expected stereotype, or drawing on personal experience. These departures were not necessarily failures. They provided evidence about how human performances preserve identity without applying the same characteristics uniformly to every turn.

Trainer-informed conversational clones.

The conversational clones were implemented as editable, prompt-conditioned profiles rather than through model fine-tuning. This made

their instructions inspectable and allowed trainer feedback to be incorporated during development. It also preserved an important analytical distinction: the clone was a structured representation derived from the Phase 1 performance, not a psychological replica of the trainer.

Computational stylometry demonstrates that measurable textual features can carry information about writing style, while also identifying short and noisy texts as a difficult setting for reliable attribution [Stamatatos, 2009]. The present study therefore uses stylometric descriptors as profile evidence, not as an independent authorship classifier.

Research on synthetic users positions generated personas as augmentations to traditional persona representations rather than direct replacements [Gu et al., 2025]. A prompt-conditioned profile can preserve observed tendencies such as response length, vocabulary, tone, topic preference, repair behavior, and affective style. It may still fail to reproduce when, why, or how strongly the source person departs from those tendencies.

The purpose of the Phase 2 comparison was therefore not to establish that a clone contained the trainer’s identity. It was to investigate how evaluators responded to a human performance and a prompt-conditioned representation derived from that performance. The relevant evidence lies in the evaluators’ questions, shifting suspicions, explanations, and interpretations of the differences between the two conditions.

Relational language as an interpretative influence.

The relational organization of the framework was influenced by Bhartrhari’s account of linguistic meaning. In this tradition, meaning is not adequately represented as an independent property of isolated words. It emerges through relations among linguistic elements and through the larger interpretative whole in which an utterance occurs [Pandey, 2009].

This thesis does not use Bhartrhari’s work as an empirical theory of human–AI interaction. It uses the relational principle as an interpretative influence. Conversational signals are not assigned fixed meanings independent of their context. Their significance depends on how they relate to previous turns, participant expectations, subsequent repair, and the developing judgment of the evaluator.

This influence is most visible in the Conversational Signal Table. The table does not classify a feature such as timing, repetition, or specificity as inherently human or artificial. It records conditional relationships and allows observations to revise the meaning of an anticipated category. The empirical authority remains with the participant records and analyses; the philosophical perspective informs how those records are organized.

Conceptual positioning.

The conceptual framework brings together conversational identity, sequential impression updating, trust dynamics, relational signals, breakdown, repair, role-play, and prompt-conditioned representation. These concepts support analysis of the behavior of human judges within an established Turing-test paradigm.

The framework does not redefine the Turing test and does not turn the study into a general evaluation of conversational systems. Its purpose is to make the human judgment process inspectable. It distinguishes the final classification from the trajectory that preceded it, the observable feature from its contextual interpretation, the breakdown from the attempted repair, and the source performance from its prompt-conditioned representation.

This positioning supports the empirical question developed throughout the thesis: how do people probe conversational partners, update trust and suspicion, and interpret the signals through which a partner appears human or artificial?

Chapter 4

The Conversational Signal Table

4.1 Purpose and Definition

Chapter 3 defined conversational identity as an evolving judgment produced through relations among messages, expectations, and responses. The Conversational Signal Table provides a structured instrument for organizing the evidence used in those judgments.

The Table does not propose a universal taxonomy of human and artificial conversation. It records categories that were considered relevant before data collection and provides a consistent structure through which later observations can be compared. Its purpose is to prevent the analysis from becoming an unrestricted list of cues assembled after the results were known.

The Table supports three questions:

1. Which anticipated conversational signals appeared in the study?
2. How did participants interpret those signals while judging a partner's identity?
3. Which anticipated signals remained absent, weakly observed, or insufficiently instrumented?

These questions concern human interpretation. A signal enters the analysis because a participant responded to it, described it, tested it, or changed a judgment in relation to it. The Table is not a benchmark for assigning an independent quality score to a conversational system.

Relational organization.

Conversational cues do not possess fixed meanings in isolation. Timing, specificity, typographical variation, emotional expression, and repair may support different judgments depending on when they occur, how often they recur,

and whether they fit the surrounding exchange. The Table therefore organizes signals as relational evidence rather than independent indicators.

This organization was influenced by Bhartrhari’s relational account of linguistic meaning. In that account, the meaning of an expression depends on its relation to a larger linguistic and interpretative whole [Pandey, 2009]. The thesis uses this principle as an organizational influence, not as an empirical explanation of human–AI interaction.

A second organizational principle is to retain explicitly open positions in the structure. Conversational signals do not form a complete or periodic classification, and an empty position does not predict a future behavior. Instead, it preserves an unanswered analytical question and prevents the initial categories from appearing complete.

Definition.

The *Conversational Signal Table* is a structured arrangement of observable interactional indicators used to examine how people form and revise judgments about a conversational partner. It organizes cues into related domains, interprets them through their context and recurrence, and preserves unsupported or unobserved categories for later revision.

The Table has two analytical levels. The first identifies an observable feature, such as a delay, repetition, clarification, personal detail, or change in tone. The second examines the relation between that feature and the surrounding interaction. This includes what preceded the feature, whether it answered a probe, whether it incorporated shared context, and whether the participant’s suspicion subsequently changed.

The Table consequently distinguishes between an observation and its interpretation. A typographical error is an observation. Treating it as evidence of humanness, or later treating its repetition as evidence of artificiality, is an interpretation. The same observation may therefore produce more than one authenticity shift event during a conversation.

4.2 Signal Domains and Trajectories

Initial signal domains.

Table 4.1 presents the signal domains established before the formal Phase 1 sessions. The categories were drawn from the literature on timing, grounding, repair, social responses to computers, conversational design, and identity interaction [Bucholtz and Hall, 2005, Clark and Brennan, 1991, Hall, 2018, Nass

Table 4.1: Conversational signal domains established before Phase 1. The anticipated forms represent starting hypotheses. Their meaning remained conditional on context, distribution, and participant interpretation.

Domain	Analytical focus	Anticipated contrasts
Temporal	Response delay, pacing, hesitation, and variation across turns	Situated and irregular timing compared with unusually fast, uniform, or mechanically repeated timing
Linguistic	Grammar, vocabulary, register, typographical variation, and adaptation to the interlocutor	Contextual variation compared with invariant polish, repetition, or conspicuous simulation of informality
Pragmatic	Clarification, self-repair, grounding, topic continuity, and recovery after misunderstanding	Situated repair and gradual recovery compared with generic clarification or failure to incorporate prior turns
Social and affective	Empathy, humor, politeness, reciprocity, boundaries, and emotional adaptation	Context-sensitive modulation compared with generic warmth, invariant politeness, or affect unrelated to the immediate exchange
Narrative	Recall, personal specificity, continuity, and treatment of shared conversational history	Relevant but uneven specificity compared with contradiction, non-retention, or invented shared experience
Meta-conversational	Reflection on the conversation, the partner, and the identity task	Flexible engagement with identity-oriented questions compared with rigid, defensive, or formulaic responses

and Moon, 2000, Schegloff et al., 1977]. They were hypotheses about potentially relevant evidence, not predetermined rules for distinguishing humans from machines.

Figure 4.1 shows the same structure as a UML class diagram. The six domains are parallel specializations rather than stages or ranked branches. Observable features are interpreted within their conversational context, and those interpretations may contribute to recorded authenticity shifts within a suspicion trajectory.

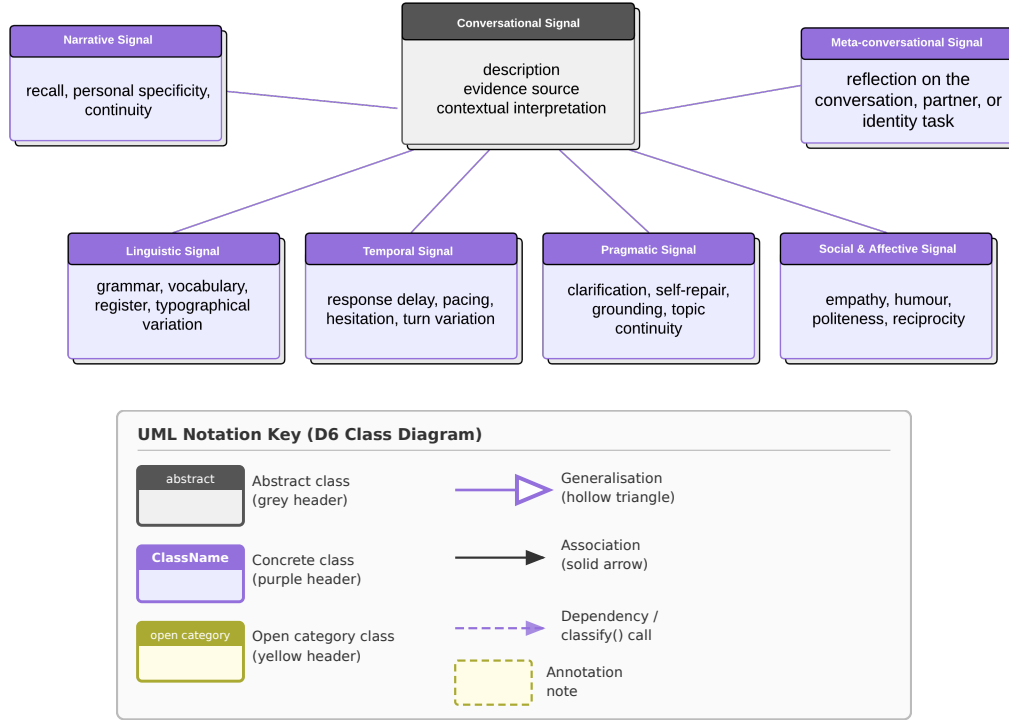


Figure 4.1: Conceptual organization of the Conversational Signal Framework. Six parallel domains classify observable features, whose evidential meaning depends on the surrounding interaction. Contextual interpretations may contribute to authenticity-shift events that form a round-level suspicion trajectory; open analytical positions preserve evidence not captured by the initial domains. The notation key applies to this class diagram.

The original structure also reserved open analytical positions. These did not predict a fixed number of undiscovered signals. Their purpose was to preserve the possibility that participant behavior would reveal recurring phenomena that did not fit the initial domains. An open position could later remain empty, be populated provisionally, or motivate a change to the organization of an existing domain.

From signals to trajectories.

The Table is used together with the temporal concepts defined in Chapter 3. A signal is not analyzed only because it occurred. It is aligned, where possible, with the participant’s probing behavior, think-aloud commentary, written explanation, and suspicion trajectory.

This alignment supports the following sequence of analysis:

1. Identify the conversational event or participant probe.

2. Identify the partner response and the signal domains implicated by that response.
3. Examine whether the participant commented on, challenged, or returned to the response.
4. Determine whether an authenticity shift event followed.
5. Examine whether later interaction reinforced the shift or produced trust drift, interactional repair, or conversational recovery.

The sequence does not establish causality from temporal proximity alone. A slider movement may reflect several accumulated cues rather than the immediately preceding message. Transcript evidence, participant commentary, and post-round explanations are therefore used to interpret the trajectory.

4.3 Use and Evidential Status

Relationship to trainer-informed clones.

The Table organizes evidence involved in human identity judgments. It does not claim to measure whether a clone accurately reproduces the trainer as a complete person. The trainer-informed clones represent behavioral tendencies observed during bounded role-play sessions.

Phase 1 uses the Table to organize signals that trainers noticed while being investigated by the baseline agent and to identify behavioral patterns in the trainers' own performances. Phase 2 uses the same domains to examine what evaluators noticed while comparing a trainer with a prompt-conditioned representation derived from that trainer's performance.

This continuity allows the analysis to ask whether signals anticipated before collection appeared across both phases, changed meaning, or required revision. It does not make the two phases equivalent. Phase 1 trainers commented while producing the source material, whereas Phase 2 evaluators judged concealed partner identities through paired interactions.

Evidential status.

The Conversational Signal Table is an author-developed analytical instrument. Its initial domains were established before the formal trainer sessions, but their later interpretation was informed by the collected data. The Table should therefore be described as empirically revised rather than independently validated.

Several limitations follow. The qualitative coding was conducted by the author, the participant samples were small, and not every anticipated signal

was captured equally well by the prototype. For example, sent-message logs can show response timing but cannot establish composition-level hesitation without keystroke or draft-state data. An unsupported category may therefore indicate absence, insufficient opportunity, or inadequate instrumentation.

The empirical revision of the Table belongs after the Phase 1 and Phase 2 results have been presented. Chapter 9 returns to the initial domains, identifies which were supported or reinterpreted, and discusses the new categories that emerged from the two studies.

Chapter 5

Study Design and Methods

5.1 Study Design

This thesis uses a sequential, mixed-method experimental design grounded in the Turing-test paradigm. The study investigates how people behave when forming and revising judgments about whether a conversational partner is human or artificial. The term *Clone Testing* identifies the two-phase experimental configuration used in this thesis; it does not denote a newly invented alternative to the Turing test.

The empirical comparison is between a human trainer enacting a selected persona and a trainer-informed conversational clone derived from the same role-playing context. The study does not include an unconditioned baseline language-model condition in Phase 2. Its purpose is therefore not to rank language models or estimate the general human-likeness of a model family. It examines how evaluators probe, interpret, and distinguish a source performance and its prompt-conditioned representation.

A *trainer-informed conversational clone* denotes a prompt-conditioned agent constructed from behavioral patterns observed while a trainer enacted a fictional persona. The clone represents selected tendencies in phrasing, response length, timing, repair, topical grounding, specificity, and affective stance. It is not treated as a complete reproduction of either the trainer or the fictional character.

The study proceeded in two linked phases. Phase 1 elicited role-played conversational behavior from seven trainers and used that material to construct and revise trainer-informed profiles. Phase 2 placed four of those profiles in a paired identity-inference study involving ten independent evaluators. Each evaluator interacted separately with a human trainer and the corresponding clone through the same participant-facing interface.

Across Phase 1, piloting, and Phase 2, 22 people participated in three dis-

tinct roles: seven trainers, five pilot participants, and ten evaluators. These groups are not pooled analytically. Phase 1 findings concern the seven trainer sessions, pilot observations concern instrument development, and Phase 2 findings concern the ten paired evaluator sessions.

The first phase produced the behavioral material from which the trainer-informed profiles were constructed. The second phase then made those profiles objects of judgment. This dependency distinguishes the design from a generic human-versus-chatbot comparison: each clone was associated with a particular trainer performance, and each Phase 2 evaluator encountered both the trainer and the corresponding clone.

Phase 2 used a within-participant comparison. Every evaluator completed two rounds, one with the assigned human trainer and one with the associated clone. Condition order was balanced across the executed sample: five evaluators encountered the clone first and five encountered the trainer first.

The human condition was implemented through a Wizard-of-Oz arrangement. Wizard-of-Oz studies conventionally use a human operator to provide or control behavior presented through an apparently unified system interface [Dahlbäck et al., 1993, Maulsby et al., 1993]. In this study, the technique provided a controlled human comparison rather than simulating unfinished automation. The trainer replied through a separate operator interface, while the evaluator continued to use the same participant-facing chat used in the clone condition.

At the boundary between rounds, the server changed response authority according to the evaluator’s assigned condition order. Trainer access was enabled only during the human round and disabled during the clone round. The term *handover* refers to this transfer between rounds. Partner identity did not change within a round, so the study does not support claims about reactions to a covert mid-conversation substitution.

Figure 5.1 summarizes the two linked phases.

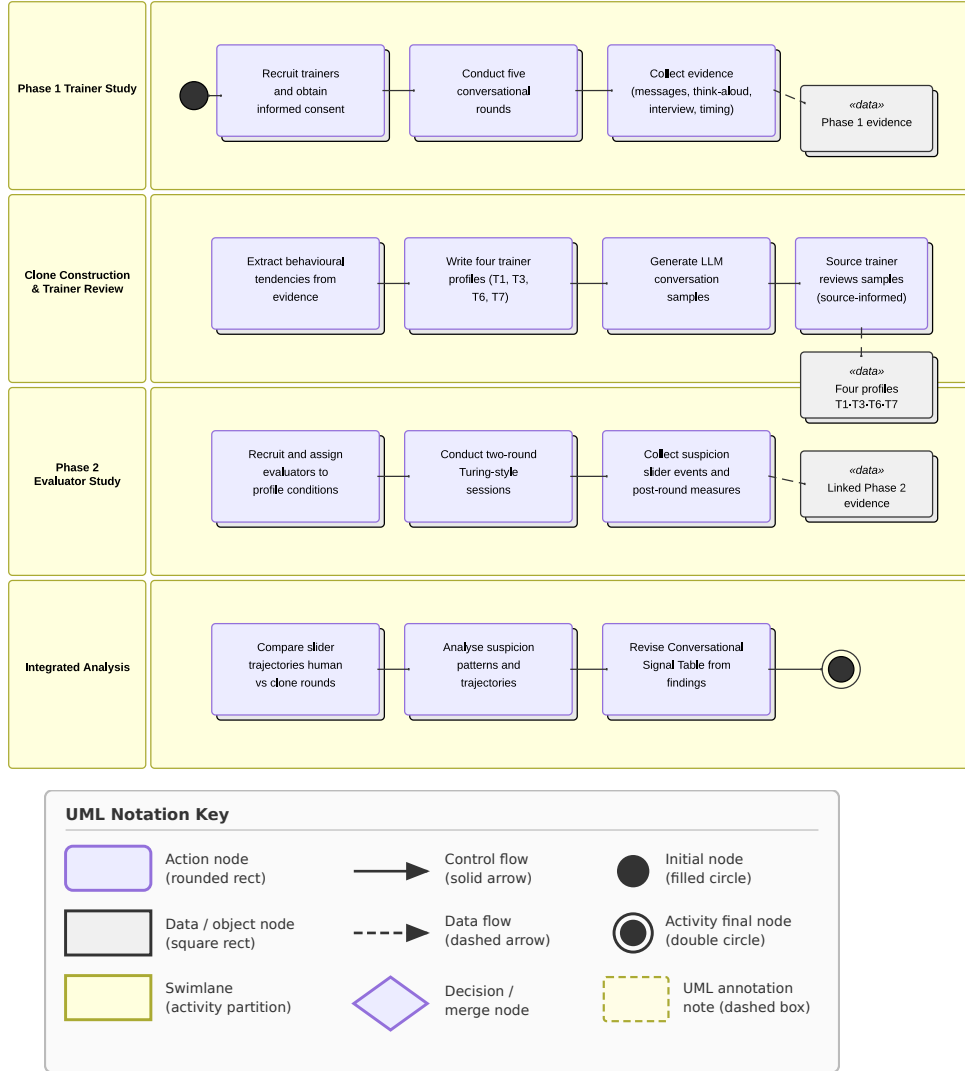


Figure 5.1: Overall two-phase study design. Phase 1 elicited trainer behavior used to construct and review editable human-informed profiles. Phase 2 compared each deployed clone with its associated human trainer, after which the linked judgments, slider trajectories, transcripts, and written explanations were analyzed together. The notation key applies to this activity diagram.

5.2 Pilot Study and Instrument Refinement

Before the formal trainer and evaluator studies, an in-person pilot was conducted on 28 March 2026 with five participants. The pilot used an early version of the conversational identity-detection interface and was intended to examine whether participants understood the task, could record changes in suspicion, and could complete the final partner judgment. The pilot showed that participants revised their judgments during a conversation rather than forming only a terminal classification. Changes were associated with response timing, apparent empathy, conversational specificity, and the partner’s ability to respond coherently to the developing situation. These observations supported the decision to retain a suspicion slider and to record its movements with timestamps and message-turn references.

The pilot also revealed practical problems involving text-entry visibility, the prominence of the final decision control, the interpretation of the slider direction, and the pressure created by the original audio and timing design. These observations informed subsequent revisions to the prototypes described in Chapter 6. Pilot participants were not included in either the Phase 1 trainer sample or the Phase 2 evaluator sample. Pilot observations were used for instrument refinement and were not combined with the formal results. Photographs of the pilot setting and participant-facing study presentation are included in Appendix A, Figure A.1.

5.3 Phase 1: Trainer Study

Phase 1 collected conversational and reflective evidence from trainers while they enacted selected personas and were questioned by a baseline LLM investigator. The phase served two purposes. It provided behavioral material for constructing the trainer-informed profiles, and it documented how trainers themselves responded to being conversationally investigated.

The phase did not attempt to clone unrestricted personal identities. Trainers enacted bounded fictional roles, allowing behavioral variation to emerge within a shared role-play structure.

Participants and role archetypes.

Seven trainers completed the formal study. Each selected one of nine school-based archetypes. The roles were designed to be recognizable but incomplete, providing a shared social orientation without prescribing a full personality or fixed script.

The archetype design was influenced by character selection in role-playing games, where a broad role becomes specific through enactment. The school

setting supplied familiar contrasts in status, responsibility, sociability, authority, and personal interests. The roles were generalized rather than modeled on named fictional characters or real individuals.

Table 5.1: Role archetypes offered in Phase 1. Each participant-facing description specified a broad social orientation while leaving the trainer room to interpret the role.

Archetype	Participant-facing description
Creative Loner	Independent and creative; prefers exploring ideas alone and expressing thoughts through art or imagination.
Athletic Jock	Confident and competitive; thrives in sports and teamwork situations.
Alternative Interests	Curious and unconventional; enjoys niche interests outside the mainstream.
Popular PR Teacher	Charismatic and approachable; engages with students and cultivates a positive school atmosphere.
Strict but Kind Teacher	Disciplined but supportive; encourages hard work while caring about students' success.
Tired Principal	An overworked administrator focused on managing the school rather than everyday classroom interaction.
Hobby Geek	A passionate hobbyist who enjoys exploring comics, technology, games, or another interest in depth.
Class Clown	Playful and mischievous; ready with a joke or prank to lighten the mood.
Confident Class Representative	Responsible and organized; represents classmates and coordinates shared activities.

Each selection card contained a title and a short description. The limited detail was deliberate: the roles needed to be clear enough to support selection while remaining open enough for trainers to introduce their own language, experiences, and assumptions.

Procedure.

Each formal session proceeded through consent and onboarding, archetype selection, five timed conversational rounds, concurrent think-aloud, and a semi-structured retrospective interview. The procedural order was fixed across trainers, while the selected role and the trainer's enactment remained participant-dependent.

During the five rounds, the trainer conversed with a baseline LLM investigator that attempted to learn about the enacted persona indirectly. The interaction was not presented as a conventional personality questionnaire. Trainers

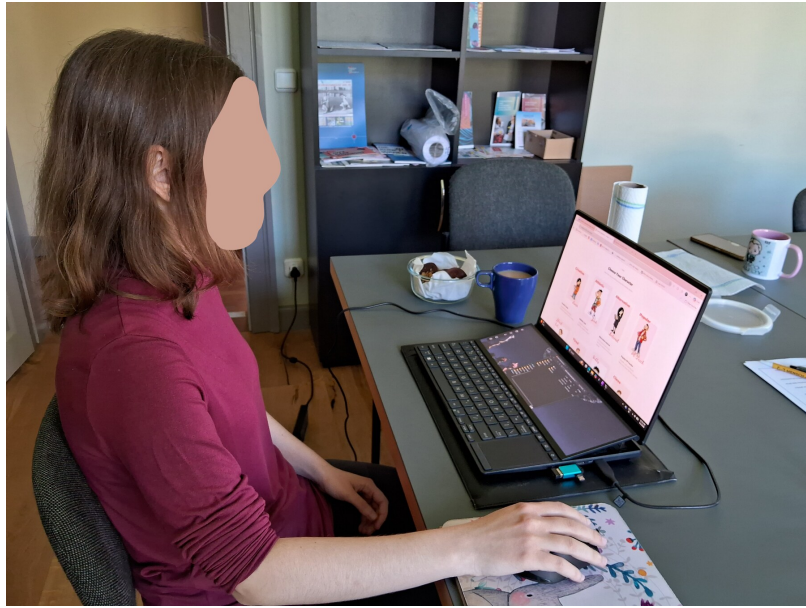


Figure 5.2: A trainer viewing the archetype-selection interface before the timed conversation. The participant’s face was obscured before the image entered the thesis corpus.

were encouraged to remain in character while also verbalizing observations about the questions, the investigator’s behavior, and their own reasoning.

Figure 5.4 complements the activity diagram with an annotated view of the participant-facing setup and the evidence recorded during the session.

Probing-question and investigator configuration.

The Phase 1 investigator followed a researcher-designed probing framework. The question set was designed to elicit the behavioral and situational evidence needed for constructing a trainer-informed profile. It covered everyday habits, social preferences, reactions to pressure or change, attitudes toward rules and responsibility, personal interests, self-expression, and responses to concrete situations.

The investigator settings separated the probing questions from the system instructions governing conversational behavior. This allowed the researcher to revise the elicitation strategy during pilot development without changing the participant-facing interface or the five-round procedure.

The system instructions also specified negative constraints. The investigator was instructed not to ask the trainer to name the selected archetype, not to request direct personality labels, not to reveal the intended inference task, and not to conduct the exchange as a rigid questionnaire. It was expected to behave as a conversational peer and obtain relevant evidence indirectly through

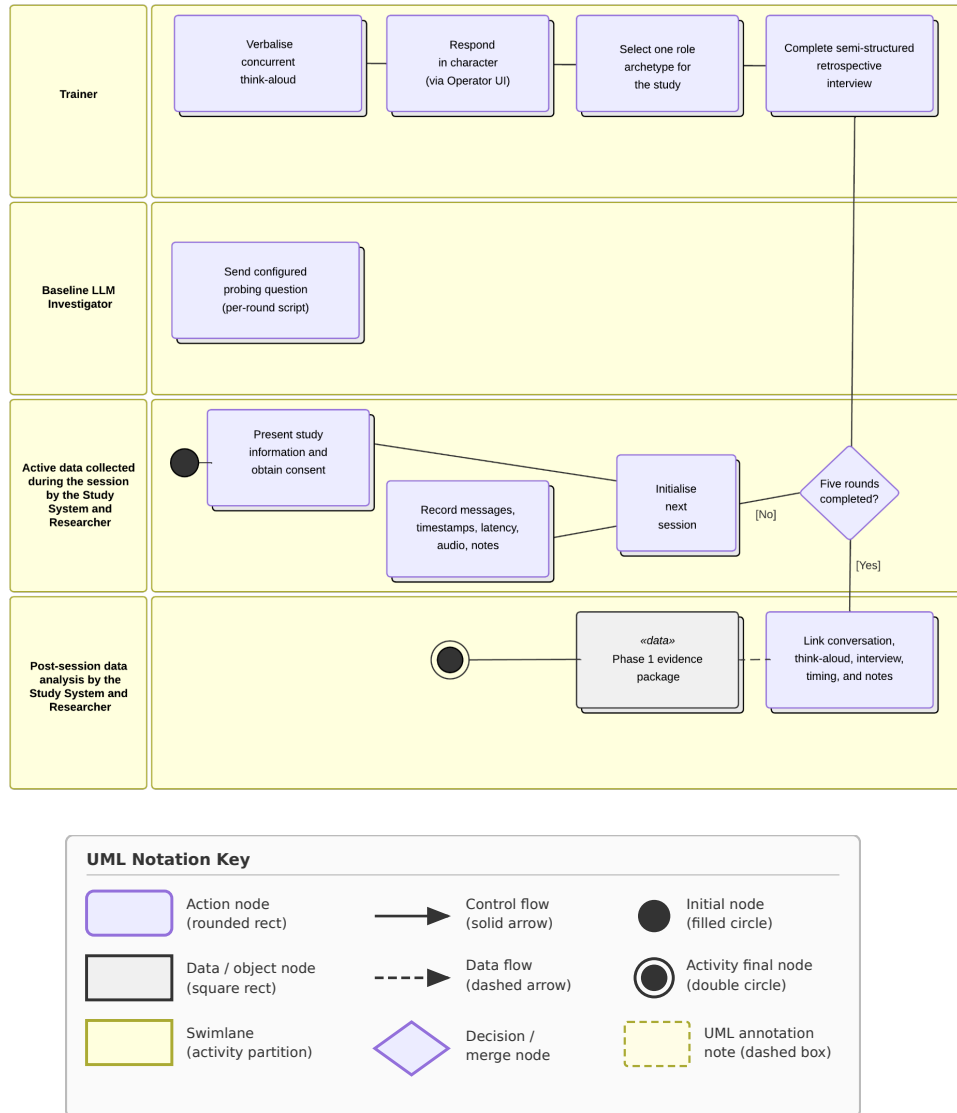


Figure 5.3: Phase 1 trainer-study activity. The baseline LLM investigator and trainer completed a repeated five-round probing sequence while the study system recorded conversational and reflective evidence. A retrospective interview completed the linked Phase 1 evidence package. The notation key applies to this activity diagram.

dialogue.

The probing configuration was frozen before the formal trainer sessions and used consistently across trainers. The investigator could formulate context-sensitive follow-up questions, but its probing remained oriented toward collecting linguistic, pragmatic, social, narrative, and temporal evidence relevant to



Figure 5.4: Rough HCI sketch of the Phase 1 interaction and data channels. Call-outs: (1a) trainer; (1b) typed-message input; (1c) baseline-LLM probe output; (1d) phone audio channel for think-aloud and interview recording; (1e) researcher observation and note-taking channel; (1f) consent and persona materials; (1g) prototype-log output; (1h) recorded-audio output; and (1i) observational-note output. The locally hosted laptop chat was the principal interaction medium. Trainer and researcher remained co-present throughout the five rounds.

later profile construction.

Think-aloud and retrospective accounts.

The concurrent think-aloud protocol captured trainers’ stated interpretations, including moments of uncertainty, discomfort, suspicion, or perceived artificiality. Think-aloud reports can reveal information attended to during a task, but they do not provide direct access to every cognitive process and may alter performance when participants are asked to explain rather than simply verbalize [Boren and Ramey, 2000, Ericsson and Simon, 1993].

The researcher therefore used neutral reminders and treated think-aloud statements as situated participant accounts. They were interpreted alongside the corresponding chat logs rather than as direct transcripts of cognition.

After the five rounds, each trainer completed a semi-structured retrospective interview. These interviews allowed trainers to explain earlier reactions, identify recurring behaviors, and discuss whether their role enactment or their perception of the investigator changed during the session.

Collected data.

The formal Phase 1 corpus contains seven completed sessions and 35 rounds,

comprising 305 messages: 135 trainer messages and 170 baseline-investigator messages. The chat exports preserve sender, round, message text, message order, and timestamps.

Fourteen verbal records were analyzed, consisting of one concurrent think-aloud record and one retrospective interview for each trainer. These were examined alongside digitized observer notes. A separate incomplete Phase 2 pilot involving T1 was used only for instrument development and was not counted as formal Phase 1 evidence.

One workbook required session-level filtering. T1’s combined export contained the completed Creative Loner session and five researcher test sessions. Only the completed trainer session was retained for Phase 1 analysis. The other six trainer workbooks each contained one completed formal session.

Raw names, signed consent forms, and unredacted source materials were excluded from the Overleaf package.

5.4 Clone Construction and Trainer Feedback

The clone-construction process converted the Phase 1 evidence into editable, prompt-conditioned profiles. It did not use model fine-tuning. The profiles were derived from the trainer messages, interaction summaries, think-aloud observations, retrospective accounts, and recurring stylometric, semantic, and pragmatic patterns.

Each profile documented relevant tendencies in areas such as tone, response length, vocabulary, grounding, repair, affect, topic preference, specificity, and behavioral constraints. These tendencies were translated into editable instructions used by the Phase 2 clone condition.

Before the configurations were shared with the source trainers, the researcher prepared and used formative *pass-as-human* test cases for each clone. The test workbook specified prompts, the purpose of each check, desired and undesired response patterns, persona-specific target styles, and pass criteria. The researcher reports that every specified case passed, but the archived workbook does not preserve the generated responses or item-level pass/fail and notes entries. The outcome is therefore a retrospective report of an internal check rather than an independently auditable test record. The name describes the design objective of the cases; they were not blinded Turing tests, participant evaluations, or independent evidence that a clone could pass as human. A condensed version of the specification appears in Appendix C.

Clone construction followed an iterative feedback loop. After an initial profile had been created, the associated trainer reviewed sample persona enactments. Trainers considered whether the enactment remained recognizable,

whether its language and interactional tendencies reflected their Phase 1 performance, and whether any instruction exaggerated or misrepresented the persona. Their feedback informed profile revisions before deployment.

This procedure constituted feedback-based validation, but not independent or blind validation. The trainers had produced the source performances and knew the intended personas. Their review established source-informed recognizability and helped identify obvious misrepresentation; it did not establish that independent observers would perceive the profiles as faithful.

Four profiles were subsequently used in Phase 2: T1 Creative Loner, T3 Responsible Class Representative 2, T6 Tired Principal, and T7 Hobby Geek.

5.5 Phase 2: Evaluator Study

Phase 2 investigated how independent evaluators distinguished a human trainer from a trainer-informed conversational clone associated with the same profile. The object of analysis was the evaluator’s behavior: the questions used to test the partner, the cues interpreted as human or artificial, the development of suspicion, and the final identity judgment.

Planned and executed sampling.

The original plan crossed six trainer profiles with four evaluators per profile, producing 24 planned sessions. The seventh trainer profile was retained as a buffer. This plan proved infeasible because of withdrawals and the difficulty of coordinating evaluators with trainers for the live human condition.

The executed study retained the paired within-participant comparison and balanced condition order but included ten evaluators across four profiles. T1 and T6 were each paired with three evaluators, while T3 and T7 were each paired with two. T7 consequently moved from buffer status into the executed sample.

The final sample was sufficient to complete the paired protocol and produce descriptive evidence from both partner types. It was not sufficient to restore the planned statistical precision or support population-level estimates. The analysis therefore prioritizes participant-level records, exact counts, paired observations, and qualitative explanations.

Procedure and Wizard-of-Oz handover.

Phase 2 was conducted remotely through a browser-based website. Evaluators entered an assigned identifier that resolved the profile, condition order, and session state. Each evaluator then completed two rounds through the same participant-facing interface.

During the human round, the assigned trainer received the evaluator’s messages through a separate mobile-compatible operator interface and returned

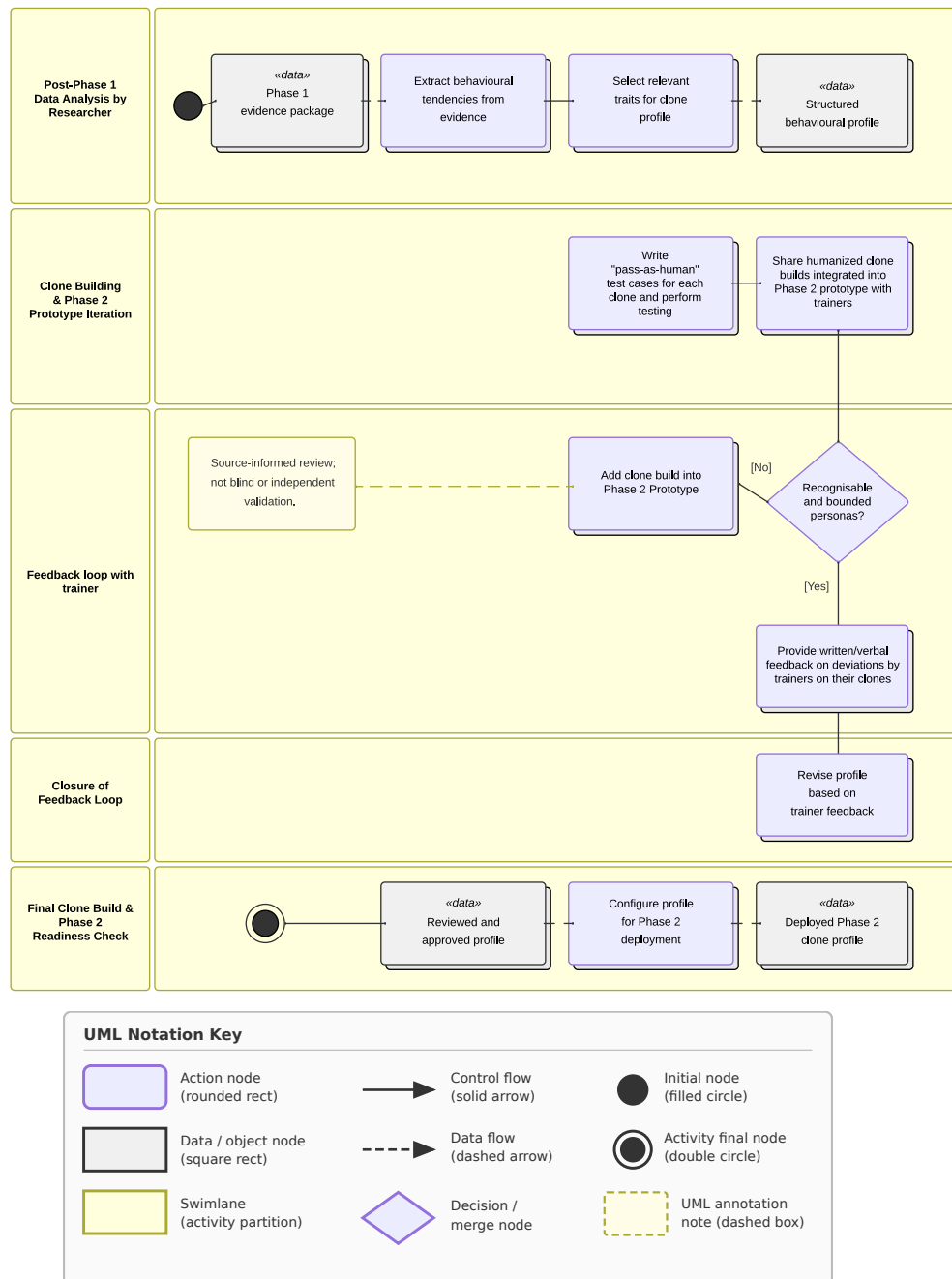


Figure 5.5: Clone construction and trainer-review activity. Phase 1 evidence was translated into an editable behavioral profile and prompt-conditioned configuration. Source trainers reviewed generated enactments; profiles were revised until sufficiently recognizable and bounded for deployment. This review was source-informed rather than blind independent validation. The notation key applies to this activity diagram.

Table 5.2: The executed study preserved the paired comparison and overall condition-order balance, but not the planned number of evaluators or equal profile balance.

Design feature	Original plan	Executed study
Trainer profiles	Six profiles; seventh held as buffer	Four profiles: T1, T3, T6, and T7
Evaluators	Four per profile; 24 total	Ten total; two or three per profile
Condition order	Human-first and clone-first sequences	Five human-first and five clone-first sessions
Pairing	Balanced crossing of profiles and evaluators	Availability-based trainer-evaluator coordination

replies through the same participant-facing chat. During the clone round, trainer access was disabled and responses were generated from the associated trainer-informed profile.

The order of conditions was balanced rather than randomly generated at run time. Five evaluator identifiers were configured as human-first and five as clone-first. At the boundary between rounds, the session controller changed response authority according to this assigned order.

The arrangement preserved a visually consistent evaluator experience while implementing a genuine human comparison. Evaluators were informed that the study involved human and artificial partners, so the Wizard-of-Oz component did not conceal the general presence of human operation. It concealed which response source was active during a particular round.

Figure 5.7 shows the same study from the evaluator’s perspective and distinguishes the visible interface from the two possible response sources.

Suspicion slider and authenticity-shift events.

During each round, evaluators could repeatedly adjust a slider representing their current belief about the partner’s identity. The scale ranged from 1 (more human) to 10 (more AI), with 5 as the neutral starting point.

Each submitted change was timestamped and linked to the session and round. An individual movement is treated analytically as an authenticity shift event. The complete sequence forms the round’s suspicion trajectory.

Movement toward the AI endpoint indicates increasing suspicion or trust drift. Movement back toward the human endpoint may indicate conversational recovery. A movement alone does not establish why the judgment changed. Interpretation requires comparison with the surrounding transcript, participant

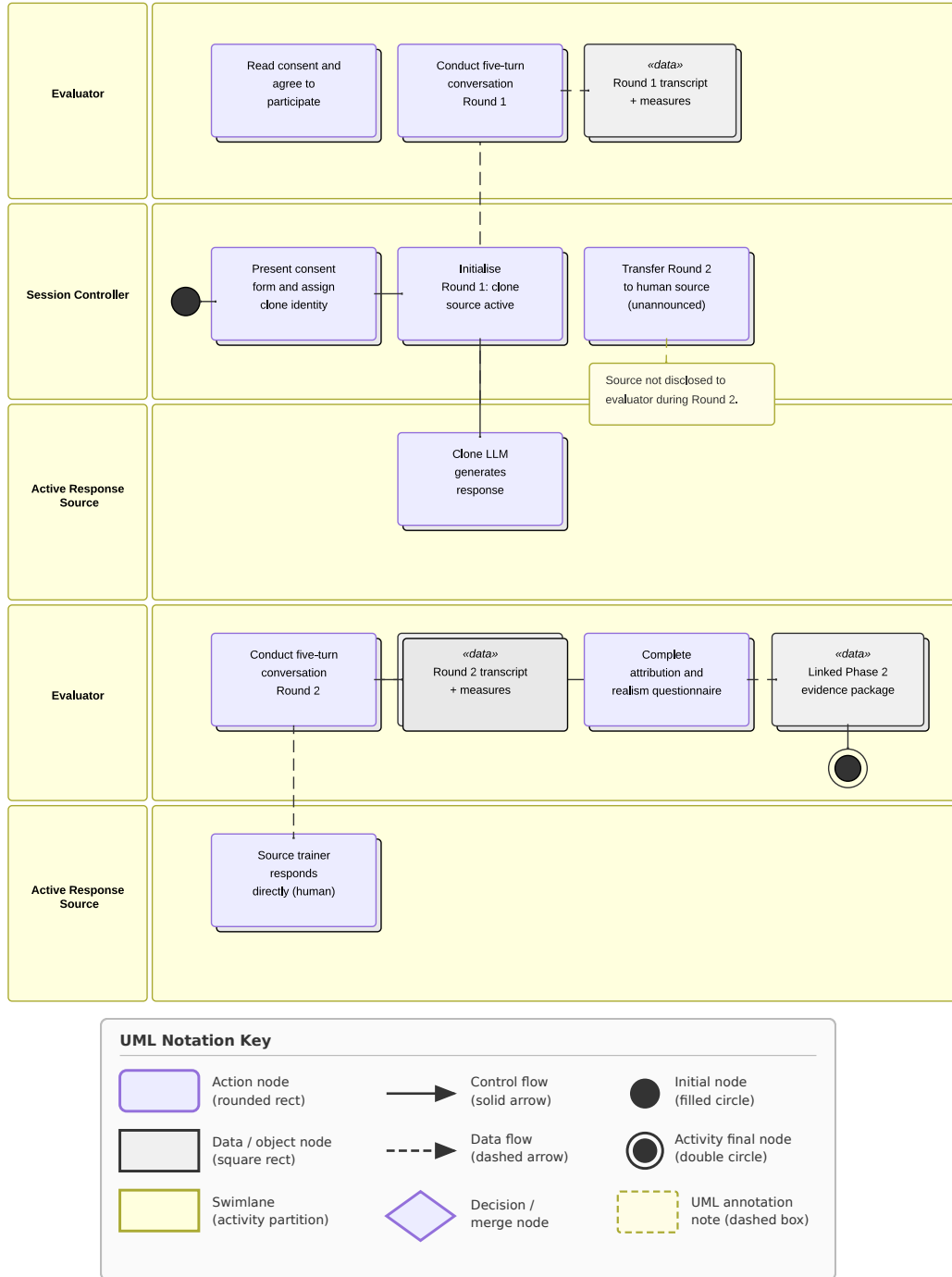


Figure 5.6: Phase 2 evaluator-study activity. Each assigned identifier resolved a linked trainer, clone profile, and balanced condition order. Evaluators completed two rounds through the same visible interface, with response authority transferred only at the boundary between rounds. The notation key applies to this activity diagram.

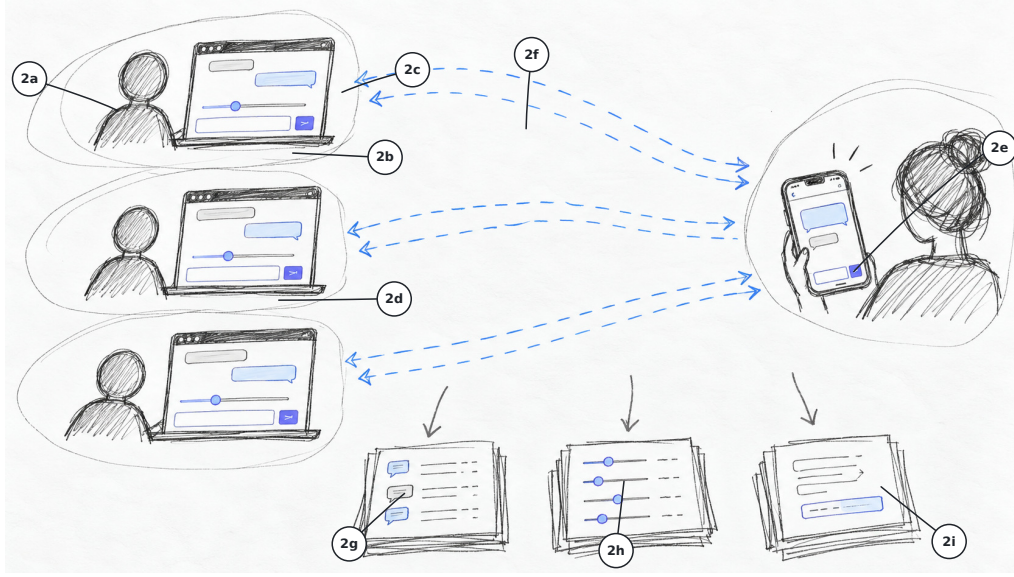


Figure 5.7: Rough HCI sketch of the remote human condition in Phase 2. Three evaluator instances illustrate repeated sessions rather than the total sample. Callouts: (2a) evaluator; (2b) typed-probe input; (2c) partner-reply output; (2d) single-slider input; (2e) trainer reply through the phone-based Wizard panel; (2f) bidirectional web-chat channel; (2g) chat, timestamp, and displayed latency log; (2h) suspicion-slider event stream; and (2i) final judgment, confidence, ratings, and written explanation. The media were the evaluator’s laptop web application and the trainer’s phone. The researcher was not active during the conversations.

explanation, and other recorded events.

Post-round measures.

At the end of each round, the evaluator submitted:

- a binary human-or-AI judgment;
- confidence on a 1–10 scale;
- perceived realism on a 1–5 scale;
- perceived difficulty on a 1–5 scale;
- frustration on a 1–5 scale;
- a written reason for the identity judgment; and
- optional additional observations.

The system also recorded time to decision, message order, timestamps, displayed response latency, partner type, profile, condition order, and session metadata.

The completed Phase 2 dataset contains ten sessions, 20 rounds, 332 chat messages, 20 post-round judgments, and 131 slider events.

5.6 Data Preparation and Analysis

Phase 1 analysis.

Phase 1 descriptive measures were recomputed from the formal message logs. These included message counts, whitespace-delimited word counts, character casing, and response latency.

A word was counted when a whitespace-delimited token contained at least one letter or number. Surface-style descriptors included the proportion of messages whose first alphabetic character was lowercase, the proportion of alphabetic characters written in uppercase, the proportion of messages containing a question mark, and an approximate Flesch Reading Ease score over each trainer’s combined messages [Flesch, 1948]. The readability calculation is reported as a descriptive surface measure only because short informal chat turns can produce unstable values.

Trainer response latency was defined as the timestamp difference between an investigator message and the immediately following trainer message in the same round. Delays above 600 seconds were treated as interruptions. Investigator latency used the positive `response_latency_ms` value recorded for agent turns because the exported trainer-latency summary field was incomplete.

Trainer comparisons remained descriptive. Messages were repeated observations nested within one session per trainer and were not treated as independent replicates for a between-trainer significance test. An exploratory spelling-flag rate was computed with the `en_US` Hunspell lexicon and affix rules. Dictionary-unrecognized tokens were manually reviewed so that names, multilingual expressions, British spellings, and established informal words were not counted as errors. Because the resulting rate depends on the selected dictionary and the manual exclusions, it is reported only as a surface-level descriptor and not as a validated measure of writing quality.

The qualitative analysis followed a hybrid deductive-inductive procedure of the kind described for combining an a priori coding structure with themes emerging from the data [Fereday and Muir-Cochrane, 2006]. The six initial domains of the Conversational Signal Table provided the deductive starting structure. The author then read the fourteen formal verbal records and digitized notes, marking episodes in which a trainer identified, tested, resisted, or

attempted to repair a partner behavior.

Where possible, each episode was checked against the corresponding chat log. The unit of interpretation was a conversational episode rather than an isolated word. Recurring observations that did not fit the initial domains were retained as candidate categories rather than forced into an existing classification.

Claims based on the message logs were checked against timestamped exports. Interpretive claims were triangulated across chat logs, think-aloud records, retrospective interviews, and observer notes where more than one source was available. This use of several data sources follows the qualitative principle of triangulation as a means of developing a more comprehensive account of the phenomenon [Carter et al., 2014].

Phase 2 analysis.

Python-based analysis traversed the linked workbook sheets while preserving the relationships among evaluator, round, profile, partner type, transcript, slider events, and written explanation. The scripts calculated:

- identity-judgment accuracy and exact binomial intervals;
- paired post-round ratings;
- slider endpoints and participant-level suspicion trajectories;
- submitted slider movements and actual value changes;
- message length;
- displayed response latency; and
- condition order.

The evaluator was the unit of analysis for paired comparisons because each evaluator contributed one human round and one clone round. Paired Wilcoxon signed-rank tests were used for the bounded post-round ratings and suspicion endpoints [Wilcoxon, 1945]. This non-parametric test compares the direction and rank of the within-evaluator differences without assuming normally distributed scores or equal intervals between adjacent scale values. For the continuous round-level summaries of mean response latency and mean words per turn, the distribution of the paired differences was examined with Shapiro–Wilk diagnostics [Shapiro and Wilk, 1965]. Because neither diagnostic provided evidence against normality, exploratory paired t-tests were used for those two measures. The diagnostic was not used as a mechanical gate for the ordinal variables, for which the measurement scale motivated the Wilcoxon

test. Exact two-sided probabilities were reported where the observed difference pattern permitted them. Given the ten-pair sample, all tests are exploratory and are interpreted alongside the individual values rather than as stand-alone evidence of a population effect.

Classification accuracy was calculated as the number of correct final judgments divided by the number of completed rounds. The exact binomial interval reported for the pooled 20 round-level judgments is retained as a descriptive indication of uncertainty. It does not remove the dependence between the two judgments contributed by each evaluator, and it is therefore not treated as a precise population-level interval [Clopper and Pearson, 1934]. Condition-specific counts and the individual error are reported with the pooled value.

For the trajectory figure, conversational progress was normalized only to place rounds of different lengths on a common start-to-end axis. This axis is neither elapsed time nor an absolute turn count. The normalized condition mean is used as a visual summary; the endpoint comparison uses the original submitted slider values rather than values read from the plotted mean line.

The evaluator’s written reasons were coded descriptively after collection. The Phase 2 reasons were initially examined separately from the Phase 1 categories and subsequently compared with the Conversational Signal Table. This separation reduced, but did not eliminate, the risk of interpreting the second phase only through categories developed in the first.

The analysis also examined selected trajectory reversals alongside their surrounding conversational turns. These comparisons were used to identify candidate breakdowns, trust drift, repair attempts, and conversational recovery. Because the study did not preregister a formal repair-success measure, these episodes are interpreted qualitatively rather than reported as a general repair success rate.

The author conducted all qualitative coding. There was no second coder, so no inter-rater reliability claim is made. Theme counts are treated as descriptive summaries rather than as a validated psychological scale.

Data correction and analytical provenance.

The Phase 2 workbook contained linked conversational records, slider events, survey responses, and session metadata. The Python workflow retained these relationships while producing analysis-ready records and the reported measures. It did not generate or replace the underlying conversations.

The thesis analysis incorporates the researcher-verified correction to E9’s clone judgment, producing 19 correct classifications out of 20. The corrected value is reported transparently in the Results chapter. The standalone chat-log artifact is maintained separately from the thesis source package.

5.7 Ethical and Data-Protection Considerations

Participants received study information and provided informed consent before participation. They were informed that the research involved conversational interactions with human and artificial partners. The study did not falsely state that every partner was autonomous or that every response came from a human.

Participant identifiers were pseudonymized for analysis and reporting. Raw names, signed consent documents, access credentials, and unredacted source files were excluded from the Overleaf package. The thesis reports trainer and evaluator codes rather than participant names.

The study avoided requesting unnecessary sensitive information. Phase 1 used bounded role-play rather than attempting to reproduce participants' complete personal identities. Participants retained the right to withdraw their data in accordance with the consent procedure.

Chapter 6

System Design and Prototype Development

The study required two connected conversational prototypes. The Phase 1 prototype elicited interaction data from trainers role-playing assigned archetypes. The Phase 2 prototype placed evaluators in conversations with a human trainer and the corresponding trainer-informed clone. These systems implemented the study design described in Chapter 5; they were research instruments rather than a new evaluation methodology.

Four requirements guided their development. First, the Phase 2 evaluator interface had to remain visually consistent across human and clone conditions. Second, response authority had to change only between rounds. Third, conversational messages, judgments, and changes in suspicion had to be recorded with sufficient temporal detail for later reconstruction. Fourth, clone profiles and study assignments had to remain editable without changing the participant-facing interface.

6.1 System Architecture

The deployed system separated the participant-facing interfaces from session control, response generation, and research-data capture. A session controller resolved each participant identifier, assigned profile, condition order, and current round. During Phase 2, it directed incoming messages either to the human operator interface or to the configured clone while preserving the same evaluator-facing chat interface.

Logging operated alongside this control path. Messages, slider submissions, questionnaire responses, partner judgments, and system events were associated with pseudonymous participant and session identifiers. This separation made

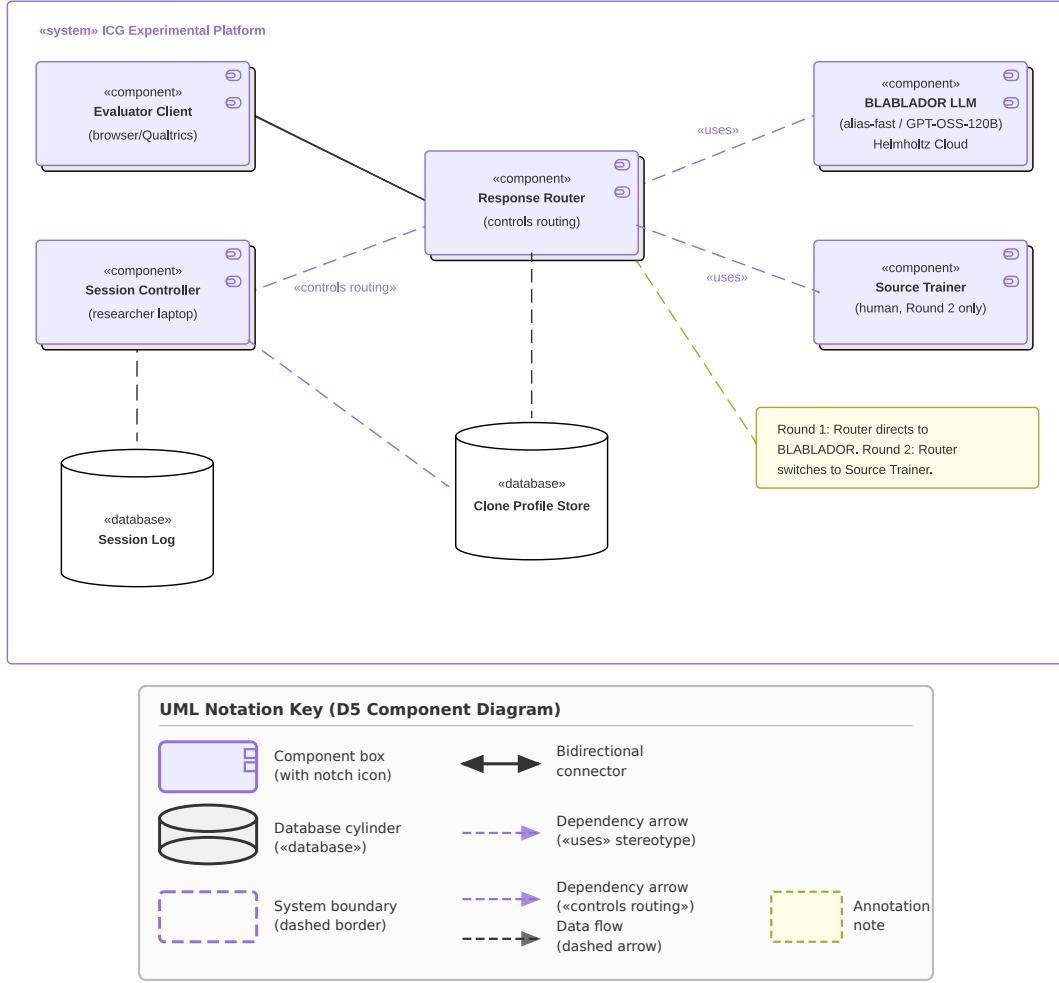


Figure 6.1: Experimental components and response routing. The evaluator used a stable participant-facing interface while the session controller enabled either the authorized Wizard-of-Oz human path or the associated trainer-informed clone path and wrote all interaction and judgment records to a linked study store. The notation key applies to this component diagram.

it possible to reconstruct the course of an interaction without revealing the active response source to the evaluator.

The browser-based application was implemented using Next.js and TypeScript, with Prisma providing the data-access layer. The Phase 2 deployment used BLABLADOR on the Helmholtz Cloud with the configured **alias-fast** (GPT-OSS-120B) model. The precise package versions and database infrastructure used during the formal sessions were not preserved in the available project documentation and are therefore not reconstructed here.

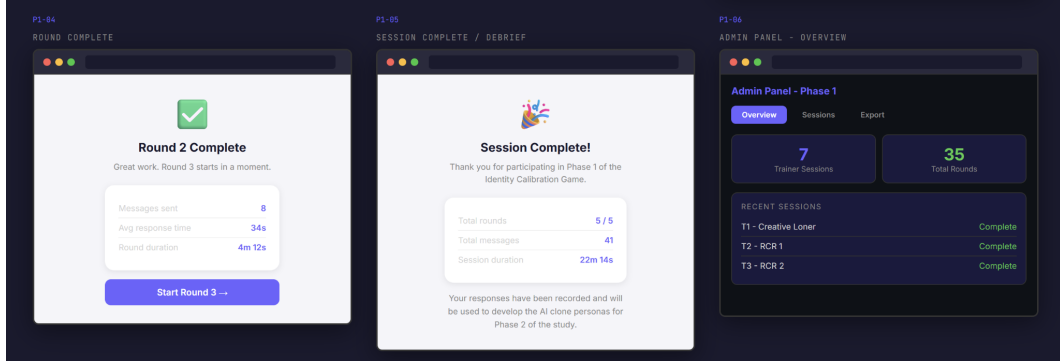


Figure 6.2: Representative states of the Phase 1 prototype, including a transition between timed rounds, the final debrief, and the administrative session overview.

6.2 Phase 1 Prototype

The Phase 1 interface combined role selection, timed conversational rounds, and a final debrief. Trainers selected an illustrated role card containing a short archetype description. The descriptions provided enough information to support role enactment while leaving participants room to interpret the archetype through their own linguistic and conversational choices.

After selecting a role, the trainer entered a text-based conversation with a baseline LLM investigator. The investigator was configured to probe the enacted persona without being told which role had been selected. The interface maintained the five-round structure, displayed transitions between rounds, and presented a debrief after the final conversation.

An administrative view allowed the researcher to verify session progression and completion. The prototype recorded the conversational material used later to construct the trainer-informed profiles. The think-aloud and retrospective procedures associated with this instrument are described in Chapter 5.

6.3 Trainer-Informed Clone Configuration

The clone system used editable prompt-conditioned profiles rather than model fine-tuning. Each profile represented selected patterns observed during Phase 1, including conversational tendencies, stylistic features, interactional preferences, contextual responses, and characteristic approaches to repair. The profiles were subsequently revised using trainer feedback.

These configurations were not intended to reproduce a trainer’s complete personality or psychological identity. They provided controlled, trainer-informed conversational conditions that could be enacted by the same

underlying language model. Their editable form allowed the researcher to inspect and revise the represented patterns while keeping the technical infrastructure constant across profiles.

The administrative interface connected each evaluator identifier with a specific trainer-informed profile and condition order. It also allowed the researcher to update profile instructions and study settings without exposing those controls to participants.

6.4 Phase 2 Prototype

Evaluator interface.

The Phase 2 evaluator interface presented both partner conditions through the same chat environment. Each evaluator entered through an assigned identifier, completed the preliminary questionnaire, and then participated in two separately controlled rounds. The visible interface did not indicate whether the current responses came from a human trainer or a clone.

A suspicion slider allowed evaluators to record changes in their judgment during each conversation. Its scale ranged from 1, indicating more human, to 10, indicating more AI. Every submitted value was stored with a timestamp and message-turn reference. The slider therefore measured an expressed judgment about partner type rather than interpersonal trust directly. Changes in these values were used with the transcript evidence to reconstruct suspicion trajectories and identify authenticity shift events.

After the required interaction threshold had been reached, the interface made the final decision control available. Evaluators then classified the partner, reported confidence, and completed the post-round questionnaire before proceeding to the second round.

Human operator interface.

The human condition used a Wizard-of-Oz arrangement behind the stable evaluator interface. The assigned trainer accessed a separate operator interface through which incoming evaluator messages could be viewed and answered.

Response authority was controlled by the session state. During a human round, the trainer could submit replies. During a clone round, trainer input was disabled and responses were generated using the assigned trainer-informed profile. Authority changed only when the session moved from one round to the next. No substitution occurred within an active conversation.

Administrative controls.

The administrative interface supported participant identifiers, trainer and profile assignments, condition order, prompt configuration, timing settings,

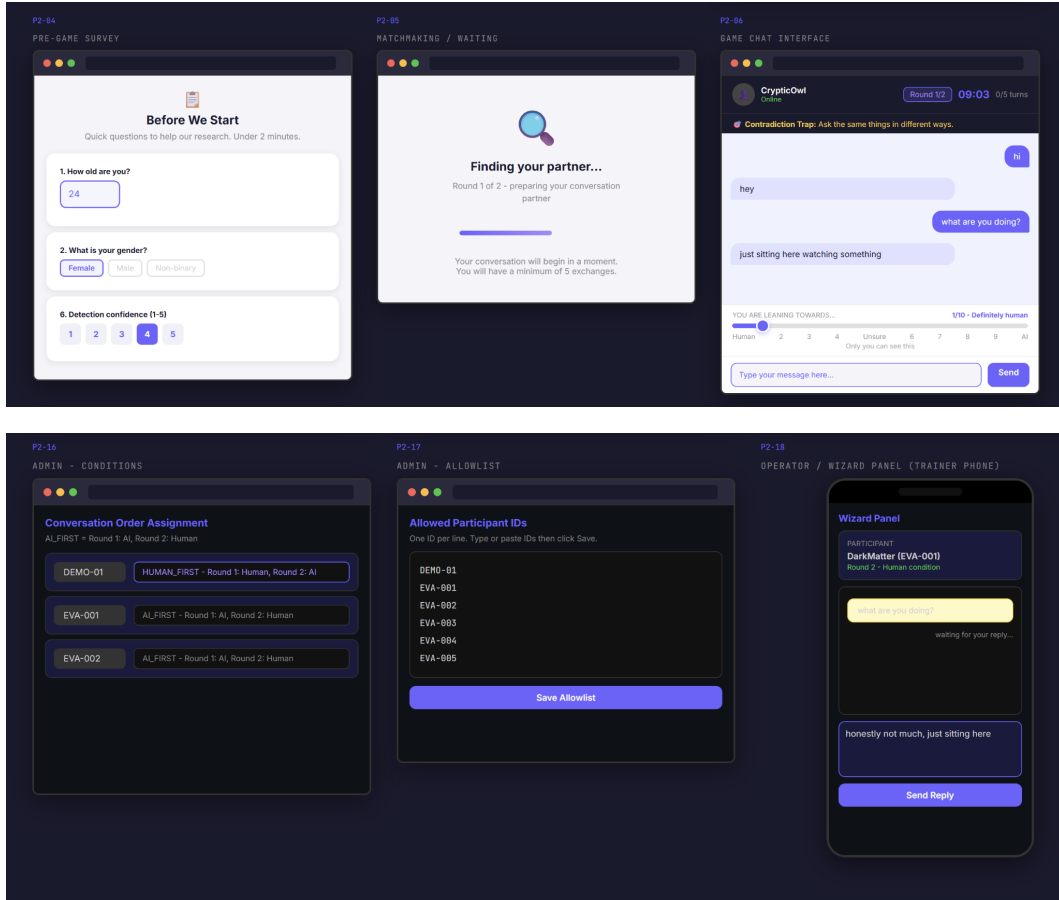


Figure 6.3: The evaluator interface combined preliminary questions, conversational interaction, the suspicion slider, and the final partner judgment. The control interfaces managed condition assignments and provided the trainer with response access during the authorized human round.

and data export. Its central experimental function was to maintain the pre-configured relationship among the evaluator, trainer, clone profile, and round order.

Logging and research export.

The system stored four principal forms of study data: conversational action logs, pre-study questionnaire responses, post-round questionnaire responses, and suspicion-slider events. The action logs recorded partner condition, sender, response source, message index, timestamp, message text, and displayed latency. The questionnaire records contained final judgments, confidence ratings, experience measures, and written explanations. Slider records contained the submitted value, partner condition, message turn, and timestamp.

These records were exported into a structured XLSX workbook for verification and analysis. Derived summaries were kept separate from the underlying event records so that calculated results could be checked against the original interaction data. Data preparation and correction procedures are described in Chapter 5.

Pilot-informed refinement.

Pilot sessions revealed several interface problems, including unclear text entry, insufficient visibility of the decision control, ambiguity in the slider direction, and discomfort caused by the original ticking-clock audio. The interface was consequently revised to improve input visibility, clarify the slider endpoints, highlight the availability of the final decision, and provide optional ambient audio with a mute control. These changes concerned the usability of the research instrument. They were intended to reduce avoidable interface confusion while preserving the identity-inference task.

Design limitations.

The prototype introduced two limitations relevant to interpretation. First, although evaluators were told that they would encounter one human and one clone, the decision interface treated each round as a separate classification. It did not prevent an evaluator from assigning the same label to both rounds or ask for confirmation when this occurred. The resulting judgments must therefore be interpreted as round-level responses rather than as a forced paired choice. Second, the suspicion slider captured explicit and voluntary submissions. Periods without slider movement cannot be treated as evidence that an evaluator's judgment remained unchanged. The trajectories describe recorded changes in expressed suspicion and must be interpreted together with messages, think-aloud comments, and retrospective explanations.

Chapter 7

Phase 1: Trainer Study Findings

Phase 1 produced the interactional material used to construct the trainer-informed conversational clones. Seven trainers completed five role-playing rounds with a baseline LLM investigator while verbalizing parts of their reasoning. The resulting evidence comprised interaction logs, think-aloud recordings, retrospective interviews, and researcher observations.

This chapter addresses three questions relevant to clone construction: how trainers differed in observable conversational behavior, how they enacted and departed from the assigned personas, and which features of the AI investigator they interpreted as artificial. Phase 1 did not test whether the resulting clones could deceive evaluators. That question is addressed using the independent Phase 2 evidence in Chapter 8.

One Phase 1 export contained researcher test sessions alongside the formal trainer session. Those test records were excluded before analysis. The findings below refer only to the seven completed trainer sessions.

7.1 Variation in Trainer Behavior

The seven trainers differed substantially in message length and response timing. Table 7.1 reports descriptive values calculated from trainer turns in the formal sessions. Response latency was reconstructed from message timestamps, with interruptions longer than 600 seconds excluded.

Table 7.2 reports additional surface-level descriptors. The role codes are CL (Creative Loner), RCR (Responsible Class Representative), AI (Alternative Interests), TP (Tired Principal), and HG (Hobby Geek). Lowercase start is the percentage of trainer messages whose first alphabetic character was lowercase. Uppercase share is the percentage of alphabetic characters written in uppercase. The readability value is an approximate Flesch Reading Ease score computed over the trainer’s combined messages [Flesch, 1948]. Because

Table 7.1: Descriptive characteristics of Phase 1 trainer behavior. Values in parentheses are sample standard deviations. Message length refers to trainer turns, and latency is the response time following an investigator message.

Trainer	Selected archetype	Turns	Words per turn	Latency (s)
T1	Creative Loner	31	13.6 (8.4)	26.0 (13.4)
T2	Responsible Class Representative	15	10.1 (6.6)	50.0 (23.4)
T3	Responsible Class Representative	22	12.5 (5.7)	34.3 (16.3)
T4	Responsible Class Representative	10	30.8 (11.4)	62.8 (17.1)
T5	Alternative Interests	28	7.9 (5.3)	28.8 (12.4)
T6	Tired Principal	13	17.2 (9.5)	57.1 (21.0)
T7	Hobby Geek	16	13.9 (9.6)	45.4 (30.2)

Table 7.2: Surface-level descriptors computed from formal trainer turns. Spell-check flags are dictionary-unrecognized tokens after manual exclusion of valid contextual terms. Percentages are descriptive and are not treated as population estimates.

Trainer	Role	Lowercase (%)	Uppercase (%)	Questions (%)	Reading ease	Spell flags (%)
T1	CL	74.2	1.71	12.9	79.2	1.18
T2	RCR	46.7	1.96	6.7	86.1	0.00
T3	RCR	100.0	0.00	4.5	78.5	4.73
T4	RCR	100.0	0.09	0.0	63.8	4.89
T5	AI	28.6	3.68	7.1	79.4	7.69
T6	TP	7.7	3.54	7.7	81.2	2.24
T7	HG	25.0	2.42	12.5	76.2	0.00

short conversational turns and informal spellings make a readability formula unstable, the value is used only as a compact description of textual surface form.

Mean message length ranged from 7.9 words for T5 to 30.8 words for T4. The variation cannot be explained by the archetype descriptions alone. T2, T3, and T4 selected the same Responsible Class Representative archetype but produced markedly different mean message lengths. The result indicates that the behavioral material reflected both the assigned role and the trainer’s own interactional habits.

The surface descriptors reinforce this heterogeneity. T3 and T4 began every recorded trainer message with a lowercase letter, whereas T6 did so in only one of thirteen turns. The three trainers who selected the Responsible Class Representative archetype also differed substantially in message length and reading ease. These contrasts show why the Phase 1 material was represented through trainer-specific profiles rather than a single archetype template.

No inferential test was used to claim statistically significant differences between trainers. Each trainer contributed one session, while individual messages were repeated observations nested within that session. Treating all messages as independent samples would overstate the amount of independent evidence,

and reducing the data to seven trainer-level means would leave too few independent units for a useful normality-based comparison. A Shapiro–Wilk test on the individual messages would assess the shape of within-session values; it would not make those messages independent observations or justify a seven-group t-test. The trainer contrasts are therefore reported descriptively.

The spell-check column is an exploratory surface descriptor rather than a validated typo measure. Alphabetic tokens were checked against the `en_US` Hunspell lexicon and affix rules. Dictionary-unrecognized forms were reviewed manually so that valid contextual items such as acronyms, British spellings, and established informal words were not counted as errors. Missing apostrophes and evident transpositions or misspellings remained flagged. The resulting percentage describes the recorded text only and is sensitive to the selected dictionary and adjudication rules.

Response timing also differed across trainers. Mean trainer latency ranged from 26.0 to 62.8 seconds. By contrast, the baseline investigator generally responded within approximately two seconds. Trainers noticed both this speed and its regularity. Fast, consistently timed responses were interpreted as evidence that the conversational partner was artificial.

The implication for clone construction was not simply that responses should be delayed. Human response behavior varied across people and across turns. Consequently, response timing needed to be treated as a profile-dependent range rather than as a constant delay applied to every clone.

7.2 Role Enactment and Persona Drift

Role enactment varied considerably across trainers. Some selected an archetype because it resembled aspects of their own personality, while others selected a role that appeared easy or interesting to perform. During the sessions, most trainers combined the assigned character with personal habits, experiences, or linguistic preferences.

T1 described the Creative Loner as personally resonant. T2 and T3 attempted to maintain the assigned role but reported that the distinction between character and self became less clear during interaction. T4 abandoned the role under time pressure and primarily answered as herself. T5 and T7 combined role-consistent details with personal responses. T6 reported the strongest separation between self and character and described the performance as a form of method acting.

Persona drift was not always accidental. T7 noticed moments in which an answer reflected personal habits and deliberately redirected the response toward the Hobby Geek character. This indicates that role enactment involved

an active process of monitoring and repair rather than simple compliance with a fixed description.

The timer made sustained enactment more difficult. Several trainers reported that limited response time encouraged short, immediately accessible answers drawn from personal experience. Under these conditions, the resulting data cannot be understood as either a pure representation of the archetype or a direct representation of the trainer. It is better described as a trainer-specific performance shaped by the role, the trainer’s dispositions, and the constraints of the interaction.

The archetypes were also unevenly performable. The Tired Principal supported a comparatively detailed performance involving frustration, sarcasm, solitude, and selective affection. The more abstract Alternative Interests role was harder to translate into consistent conversational behavior. Concrete social roles therefore appeared to provide stronger behavioral guidance than broad dispositional descriptions.

Linguistic mediation.

One trainer explained that she was thinking through parts of the interaction in Spanish and reflected that the apparent personality of a language model can change when the language of interaction changes. She also questioned whether grammatical gender might influence how an artificial agent describes itself.

This observation suggests that role performance is mediated by the trainer’s linguistic repertoire as well as by the assigned archetype. The present study cannot evaluate language-specific differences because the formal conversations were conducted in English. Nevertheless, the observation identifies multilingual clone construction as a relevant direction for future work.

7.3 Signals of Artificiality

Although Phase 1 was designed primarily to elicit trainer behavior, trainers also commented spontaneously on the baseline LLM investigator. Their comments provide evidence of the cues people notice while interacting with an artificial partner. These observations informed the conversational profiles and the signal categories introduced in Chapter 4.

Timing and structural regularity.

Trainers interpreted extreme response speed as evidence of artificiality. They also noticed repeated structural patterns, including similar message lengths, consistently polished formulations, and recurring sentence structures. Regularity across turns was therefore as important as the content of an individual response.

Attempts to simulate informality did not automatically increase perceived humanness. Abbreviations, lowercase writing, and typographical errors were insufficient when they appeared too consistently. In some cases, exaggerated misspelling became a cue of artificial construction. Surface mimicry could therefore reverse its intended effect when it followed an observable pattern.

Repetition, grounding, and fabricated context.

Repeated questions were reported across several sessions. Trainers interpreted these repetitions as evidence that the investigator had failed to retain earlier information. The problem was not repetition alone, but the interactional implication that the partner had not listened.

Trainers also objected when the investigator presented unsupported inferences as established facts rather than marking uncertainty or inviting correction. The investigator sometimes treated an inference as shared knowledge, producing a visible grounding failure.

A more pronounced form of this problem occurred when the investigator referred to events that had never taken place, such as a previous meeting, exhibition, vacation, or shared location. These references created a fabricated shared history. Trainers interpreted the resulting inconsistency as evidence that the partner could not maintain a coherent relationship to the conversation.

Embodiment and affective calibration.

Some trainers deliberately tested whether the investigator would claim bodily experiences or activities unavailable to it. Claims about walking in nature, physical sensations, or intuition were treated as evidence that the agent was producing plausible language without maintaining a coherent account of its own situation.

Artificiality was also associated with excessive emotional accommodation. Frequent apologies, unconditional affirmation, and attempts to empathise with emotionally neutral statements were sometimes interpreted as signs of an AI partner. The issue was not the presence of empathy, but its weak calibration to the preceding message.

T6 exposed a related problem through a deliberately confrontational performance. When the character became hostile or sarcastic, the investigator maintained a polite and explanatory register. Its failure to reciprocate, resist, or adapt its tone made the interaction appear less human. Contextual adaptation therefore included knowing when agreement and politeness were interactionally inappropriate.

Probe design and response constraint.

Several trainers found the investigator's questions repetitive or excessively focused on work, stress, and daily routines. Questions containing binary alternatives sometimes restricted the range of possible responses. Trainers could

Table 7.3: Phase 1 findings and their consequences for the trainer-informed clone profiles.

Phase 1 finding	Consequence for clone construction
Message length varied strongly by trainer	Response-length guidance was specified separately for each profile.
Human response timing was variable	Delay behavior was treated as a range rather than a fixed pause.
Repetition and unsupported inference signalled artificiality	Profiles emphasised retention, grounding, and qualification of uncertain assumptions.
Uniform errors and informal markers became detectable	Stylistic markers were used selectively rather than at a constant rate.
Embodiment claims exposed situational incoherence	Profiles constrained unsupported claims about physical experience.
Trainers differed in persona adherence	Profiles were described as trainer-informed performances rather than psychological replicas.

select one of the offered positions but found it harder to introduce a different interpretation.

Examples embedded in questions also influenced the content of trainer responses. Rather than retrieving an independently relevant experience, participants sometimes searched for an experience resembling the supplied example. This represents a methodological concern because the resulting responses became part of the material used to construct the clones.

At the same time, the investigator was not uniformly experienced as artificial. One trainer reported that its follow-up questions appeared to build on earlier responses and produced a relatively fluid conversation. The investigator’s practice of asking questions rather than functioning as a conventional answer-giving assistant also contributed to an impression of social engagement. The Phase 1 evidence therefore identifies dimensions on which conversational behavior could succeed or fail, rather than a simple list of defects.

7.4 Consequences for Clone Construction

The Phase 1 findings informed the profiles described in Chapter 6. The principal design consequences are summarized in Table 7.3.

These changes were not treated as independent validation that a profile

successfully reproduced its trainer. They represent source-informed design decisions grounded in the Phase 1 material and subsequent trainer feedback.

7.5 Limitations

The Phase 1 sample comprised seven trainers, and the number of trainer turns varied considerably across sessions. The quantitative values should therefore be read as descriptive characteristics of the recorded interactions rather than stable estimates of individual behavior.

Think-aloud participation was uneven. Some trainers verbalized their reasoning frequently, while others concentrated on typing and provided more detail only during the retrospective interview. The qualitative findings consequently reflect differences in participants' willingness and ability to verbalize during interaction.

Persona adherence was also uneven and was affected by time pressure. The resulting profiles represent different mixtures of enacted character and trainer disposition. In addition, the observations about artificiality arose from interaction with the baseline investigator rather than with the completed clones. Their relevance to Phase 2 therefore had to be examined rather than assumed.

Finally, the qualitative material was analyzed by the researcher without an independent coding pass. The themes provide an organized interpretation of the available evidence, but they should not be treated as an independently validated taxonomy.

Chapter 8

Phase 2: Evaluator Study Results

Phase 2 examined how independent evaluators distinguished human trainers from trainer-informed conversational clones. Ten evaluators each completed one human round and one clone round, producing 20 partner judgments, 332 chat messages, and 131 recorded slider events.

The chapter reports exact outcomes, post-round ratings, suspicion trajectories, probing strategies, and evaluators’ written explanations. Because the sample is small, the emphasis is on descriptive results and participant-level patterns rather than population-level generalization.

8.1 Identity-Detection Outcomes

Evaluators correctly classified 19 of the 20 conversations, corresponding to an observed accuracy of 95% and an exact two-sided 95% binomial confidence interval of 75.1% to 99.9% [Clopper and Pearson, 1934]. All ten human rounds were classified as human, and nine of the ten clone rounds were classified as AI.

The interval is wide despite the high observed accuracy because it is based on only 20 round-level judgments. The result therefore establishes what happened in this study—one error in 20 judgments—but does not imply that future evaluators would achieve 95% accuracy. The interval is also descriptive because each evaluator supplied two judgments, so the 20 observations are not fully independent. Reporting the human and clone counts separately makes the observed asymmetry explicit: the only error occurred in the clone condition.

The single incorrect classification occurred in E9’s conversation with the T6 Tired Principal clone. After verification of the study record, E9’s final clone-round judgment was retained as human. E9 therefore classified both rounds as human. The interface permitted the same label to be submitted

Table 8.1: Phase 2 outcomes by partner condition. Ratings are reported as mean with range in parentheses. Confidence used a 1–10 scale; realism, difficulty, and frustration used 1–5 scales.

Outcome	Human trainer	AI clone
Correct judgments	10/10	9/10
Decision confidence	6.4 (4–8)	6.7 (3–9)
Perceived realism	4.5 (4–5)	3.5 (3–4)
Difficulty	3.6 (3–5)	3.8 (3–5)
Frustration	1.4 (1–4)	1.7 (1–3)
Final suspicion value	2.9 (2–4)	7.4 (6–9)

Table 8.2: Paired comparisons across the ten evaluators. Shapiro–Wilk p is shown only for the paired differences of the two continuous summaries; it was not used for bounded ordinal variables. W denotes the Wilcoxon signed-rank statistic; $t(9)$ denotes the paired t statistic with nine degrees of freedom. Tests are exploratory and no multiplicity-adjusted confirmatory claim is made.

Measure	Human mean	Clone mean	Shapiro p	Statistic	Test p
Confidence	6.40	6.70	–	$W = 12$	.5625
Realism	4.50	3.50	–	$W = 0$	.0078
Difficulty	3.60	3.80	–	$W = 2.5$	.6250
Frustration	1.40	1.70	–	$W = 3$	.3750
Latency (s)	84.84	82.88	.135	$t(9) = 0.93$	.3768
Words per turn	11.29	11.01	.793	$t(9) = 0.40$	.6972
Final suspicion	2.90	7.40	–	$W = 0$	.0020

twice and did not ask the evaluator to confirm that decision, even though the study instructions described one human and one clone round. The record cannot establish whether E9 forgot the paired constraint, intentionally judged both rounds independently, or found the clone sufficiently convincing. The response is therefore retained as a false negative while the interface ambiguity is acknowledged as a study limitation.

Table 8.1 summarizes the principal condition-level results.

Table 8.2 reports the paired statistical comparisons. Bounded ordinal ratings and slider endpoints were analyzed with paired Wilcoxon signed-rank tests [Wilcoxon, 1945]. Mean response latency and mean words per turn are continuous round-level summaries; their ten within-evaluator differences were examined with Shapiro–Wilk diagnostics. Neither diagnostic indicated a detectable departure from normality in this small sample, so exploratory paired t -tests were used for these two measures [Shapiro and Wilk, 1965].

Human rounds were rated as more realistic than clone rounds. An exploratory paired Wilcoxon test reflected this difference ($W = 0$, $p = .0078$). Here, $W = 0$ means that every non-zero paired difference used by the test was in the same direction: evaluators did not give a higher realism rating to the clone than to the corresponding human trainer. The p -value describes how unusual that directional rank pattern would be under the signed-rank null model; it is not an effect-size measure. Confidence, difficulty, and frustration overlapped across conditions and did not show comparable separation. These comparisons are exploratory descriptions of ten observed pairs rather than population-level estimates.

Mean partner response latency was similar in the two conditions, at 84.8 seconds for trainers and 82.9 seconds for clones. Mean partner message length was also similar, at 11.29 words for trainers and 11.01 words for clones. The exploratory paired tests did not identify a condition difference in mean latency or mean message length. These aggregate summaries do not show that timing and message length were irrelevant: evaluators could still have responded to within-conversation variation, regularity, or particular turns.

8.2 Suspicion Trajectories

The suspicion slider began at the neutral value of 5 in every round. A movement toward 1 indicated that the partner appeared more human, while a movement toward 10 indicated that the partner appeared more artificial.

Human-round trajectories ended at a mean of 2.9, with values ranging from 2 to 4. Clone-round trajectories ended at a mean of 7.4, with values ranging from 6 to 9. The endpoints differed in the expected direction across all ten paired sessions ($W = 0$, $p = .0020$). In practical terms, every evaluator ended the human round on the human side of the neutral midpoint and the clone round on the AI side. The zero signed-rank statistic records this uniform direction of the paired differences; the test does not measure the size or cause of the separation.

Figure 8.1 shows that these endpoints were not usually reached through a single movement. Evaluators revised their judgments as the conversations developed, and several trajectories changed direction before arriving at their final position.

The horizontal axis normalises each round from its start to its end so that conversations of different lengths can be displayed together. It should not be read as elapsed seconds or as the same message number across participants. The heavy line is therefore a visual condition summary, while the thin lines retain the participant-level variation on which the interpretation depends.

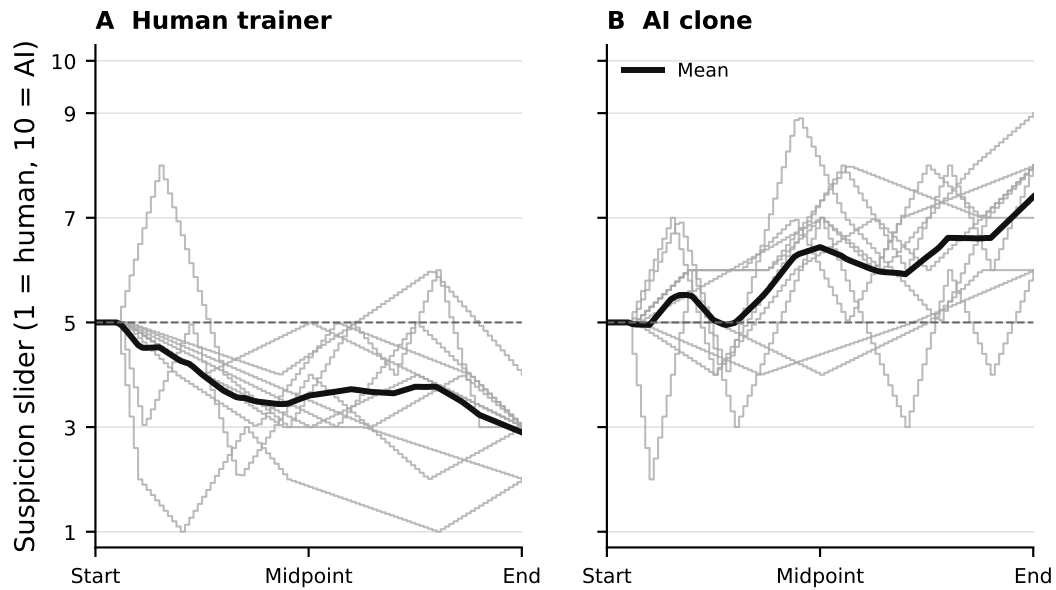


Figure 8.1: Evaluator suspicion trajectories by partner condition. Thin lines show individual rounds, and the heavier line shows the condition mean across normalized conversational progress. The scale runs from 1, more human, to 10, more AI.

Movements toward AI after an apparently artificial response are described as increases in suspicion. Movements back toward human after an earlier increase constitute conversational recovery in the recorded judgment. Such a reversal does not by itself demonstrate that a particular repair caused the recovery. That interpretation requires alignment between the slider event and the surrounding transcript.

Several clone trajectories temporarily moved toward human after responses containing personal detail, informal language, or apparent contextual engagement. Suspicion later increased when similarly polished or characterful responses appeared repeatedly across different topics. This pattern shows that a cue could initially lower suspicion and later raise it when its regularity became noticeable.

The relationship between the slider and the final classification was strong but not deterministic. E9’s clone-round slider ended at 9, indicating more AI, while the verified final judgment was human. The two records capture different acts: the slider represents an expressed judgment during interaction, whereas the final classification records the submitted decision after the round. Neither should be silently substituted for the other.

The slider therefore provides evidence about developing identity judgments, not a direct measure of general interpersonal trust. Trust dynamics are discussed only in the narrower sense of provisional acceptance of the partner’s

apparent identity and only where the trajectory can be interpreted together with the conversational record.

8.3 Evaluator Strategies and Reported Cues

Probing strategies.

Transcript analysis identified several ways in which evaluators tested their partners. These strategies were not mutually exclusive and often developed across several turns.

Continuity testing.

Evaluators returned to earlier topics or introduced indirect references to previous messages to determine whether the partner retained and appropriately used conversational context.

Experiential probing.

Personal questions were used to test whether the partner could provide plausible memories, preferences, physical experiences, or locally specific details. Specificity alone was not sufficient. Evaluators also assessed whether the detail followed naturally from the exchange.

Contextual disruption.

Some evaluators changed topics, introduced humor, or shifted emotional tone to observe whether the partner adapted. Responses that remained uniformly polite, detailed, or affirmative despite these changes increased suspicion.

Pattern testing.

Evaluators compared message length, response structure, timing, typography, and the distribution of detail across turns. This strategy treated artificiality as a recurring pattern rather than as a feature of one message.

Meta-conversational testing.

Evaluators sometimes challenged assumptions, questioned why a topic had been introduced, or tested the partner's understanding of the conversational situation. These probes examined whether the partner could recognize and repair a breakdown rather than merely continue producing plausible text.

Together, these strategies show that evaluators actively generated evidence. They did not rely only on passive impressions or a fixed checklist of linguistic cues.

Reported conversational cues.

Evaluators' post-round explanations were coded descriptively. The themes and counts in Table 8.3 are an audit of the available written reasons rather

Table 8.3: Themes identified in evaluators’ written explanations. Counts indicate the number of condition-specific explanations in which a theme appeared. Themes were non-exclusive.

Theme	Human	Clone	Observed contrast
Regularity, polish, or uniform structure	0	7	Consistency became suspicious when it persisted across topics.
Idiosyncratic or lived specificity	6	3	Detail appeared human when it was locally relevant, but artificial when it seemed evenly inserted.
Unevenness, clumsiness, or imperfect recall	5	0	Variation was interpreted as evidence of situated human production.
Contextual affect, recovery, or reciprocity	4	3	Evaluators distinguished context-sensitive reactions from generic warmth or affirmation.
Typography or deliberate error	0	2	Errors initially suggested humanness but became suspicious when their pattern appeared engineered.

than an independently validated scale. Ten human-round explanations and nine clone-round explanations were usable. The explanation stored with E9’s corrected clone-round judgment contradicted the verified classification and was excluded rather than retrospectively rewritten.

The most frequent clone-related theme was excessive regularity. Evaluators noticed similarly complete answers across topics, evenly distributed personal detail, stable typographical patterns, and repeated returns to structured prose. The relevant cue was not grammatical correctness by itself, but the persistence of the same degree of polish across changing conversational demands.

Human partners were more often associated with unevenness. Their responses varied in length, depth, fluency, and emotional engagement. Imperfect recall, clumsy phrasing, and locally relevant detail were interpreted as signs that a response had been produced within the situation rather than assembled to maintain a consistently legible persona.

The same surface feature could support opposite judgments. Typographical errors sometimes reduced suspicion, but repeated or conspicuous errors increased it. Personal detail could appear human when it followed from the conversation, but artificial when every answer contained similarly vivid detail. Evaluators therefore interpreted cues relationally, through their distribution

and contextual consequences.

8.4 Variation Across Profiles

The human-clone contrast varied across the four profiles included in Phase 2. The largest mean separation in final suspicion occurred for T6, followed by T1. T3 produced the narrowest separation and comparatively high reported difficulty. T7’s unusual hobby and folklore content made both the trainer and clone appear plausible until evaluators considered reciprocity and contextual adaptation.

Table 8.4 places the source trainers’ Phase 1 response lengths beside the corresponding Phase 2 conversations. The final two columns summarize evaluator outcomes; they do not measure clone fidelity from the trainer’s perspective.

Table 8.4: Descriptive comparison of the four profiles used in Phase 2. Mean words are reported per trainer or partner turn. Confidence and realism are Phase 2 evaluator ratings on their respective study scales.

Profile	Archetype	Mean words per turn			Detected	Conf./realism
		Phase 1	Human	Clone		
T1	Creative Loner	13.6	12.5	11.9	3/3	7.7 / 3.7
T3	Responsible Class Representative	12.5	11.6	13.4	2/2	5.5 / 3.5
T6	Tired Principal	17.2	11.4	8.9	2/3	8.0 / 3.0
T7	Hobby Geek	13.9	8.9	10.4	2/2	4.5 / 4.0

The numerical comparison does not show how the enacted personalities changed between the source performance, the configured clone, and the evaluators’ interpretations. Tables 8.5 and 8.6 therefore provide a qualitative cross-case comparison. The entries distinguish the Phase 1 records, the deployed profile instructions, and the Phase 2 post-round accounts. These are different evidence sources and are not treated as interchangeable ratings of personality or fidelity.

Trainer review during clone construction was qualitative and configuration-specific; it was not collected with a common numerical fidelity scale. It therefore cannot be compared statistically with evaluator confidence or realism ratings. A numerical trainer–evaluator comparison would require a standardized fidelity instrument administered to every trainer.

Given the small, uneven profile sample, the evidence cannot support profile rankings or stable claims about archetypes: T1 and T6 appeared in three evaluator sessions each, while T3 and T7 appeared in two. Detectability instead appeared to emerge from the interaction of profile content, clone behavior, evaluator strategy, and conversational context.

Table 8.5: Cross-case comparison of trainer enactment, clone configuration, and evaluator interpretation for T1 and T3. The entries summarize qualitative patterns rather than fixed personality attributes.

Profile	Evidence source	Observed or configured expression
T1 Creative Loner	Phase 1 trainer	The role was personally resonant. The trainer used casual, selectively social responses, playful associations, and occasional challenges to repetition or unsupported assumptions.
	Clone configuration	The profile emphasized a reflective casual register, selective warmth, dry humor, redirection of intrusive questions, and continuity with earlier details.
	Phase 2 evaluators	The human condition appeared playful, idiosyncratic, and capable of imperfect recovery. The clone became suspicious when recall, structure, and response detail remained consistently precise across unlike questions.
T3 Responsible Class Representative	Phase 1 trainer	The trainer attempted to sustain the role while reporting that self and character became less distinct. The performance combined reflective hedging, lowercase writing, contextual frustration, and deliberate imperfection.
	Clone configuration	The profile preserved hedges, lowercase phrasing, small inconsistencies, emotional awareness, and occasional meta-comments instead of smoothing the performance into formal prose.
	Phase 2 evaluators	This was the narrowest human–clone contrast. Stable lowercase and typographical patterns made the clone appear manufactured, whereas variation in expression and context-dependent frustration supported the human judgment.

Table 8.6: Cross-case comparison of trainer enactment, clone configuration, and evaluator interpretation for T6 and T7. The entries summarize qualitative patterns rather than fixed personality attributes.

Profile	Evidence source	Observed or configured expression
T6 Tired Principal	Phase 1 trainer	The trainer reported the strongest separation between self and role and treated the session as method acting. The enactment combined fatigue, sarcasm, restraint, confrontation, and selective warmth.
	Clone configuration	The profile emphasized compressed replies, deadpan sarcasm, routine, social withdrawal, and rare warmth around family.
	Phase 2 evaluators	Evaluators contrasted the human trainer’s uneven rhythm and marked affective change around family with the clone’s consistently crafted detail and more generic or applied warmth. One evaluator nevertheless classified the clone as human.
T7 Hobby Geek	Phase 1 trainer	The trainer mixed personal material with the assigned role and deliberately redirected moments of persona drift. Niche interests and unusual folklore or fantasy details became central to the enactment.
	Clone configuration	The profile emphasized elaboration on folklore, mythology, books, and fandoms, while keeping surface questions brief and resisting repetitive or intrusive probing.
	Phase 2 evaluators	Specific hobby material made both conditions plausible and produced repeated changes of mind. The human partner appeared more responsive to the evaluator’s contribution; the clone was sometimes polished or insufficiently reciprocal despite relevant niche detail.

Chapter 9

Discussion

This study examined how people judge whether a conversational partner is human or artificial, how those judgments change during interaction, and which features distinguish human trainers from trainer-informed conversational clones. The principal result is not simply that evaluators correctly classified 19 of 20 conversations. Their judgments developed through relations among cues across turns. Personal detail, typographical errors, politeness, response timing, and stylistic consistency could each support either a human or an AI judgment. Their meaning depended on when they appeared, how often they recurred, whether they fit the local context, and how the partner responded when challenged.

RQ1: Which conversational signals contributed to shifts in trust and suspicion?

Four signal relations were especially important: regularity, grounding, contextual adaptation, and recovery. Evaluators frequently interpreted excessive consistency as evidence of artificial construction. The clones often maintained similar response length, degree of polish, typographical style, and density of personal detail across different topics. These features were not necessarily suspicious when they first appeared. They became suspicious when evaluators recognized their regular distribution.

Human responses were more uneven. Trainers varied their effort, detail, fluency, and emotional engagement according to the immediate conversational demand. This unevenness was often interpreted as situated production rather than poor conversational quality. The relevant distinction is therefore not between consistent and inconsistent partners. A recognizable persona requires continuity. The difference observed here was between stable identity and uniform performance: human trainers remained recognizable while varying their responses, whereas the clones more often expressed their strongest profile features at similar intensity across contexts.

Repetition, unsupported assumptions, and fabricated shared history concerned the partner's treatment of conversational common ground. Repeated questions suggested that earlier information had not been retained. Unsupported inference treated uncertain information as established. Fabricated history introduced events or relationships that had never occurred. These failures were more consequential than isolated grammatical or factual errors because they disrupted the relation between the current response and the preceding interaction. This supports accounts of conversational grounding in which successful dialogue depends on participants maintaining and updating a shared understanding Clark and Brennan [1991]. Fluency was therefore insufficient: a partner could produce locally plausible text while revealing artificiality through an incoherent model of what had already been established.

Evaluators also attended to whether the partner adapted to changes in topic, affect, humor, resistance, and conversational register. Generic empathy, unconditional agreement, and stable politeness sometimes appeared artificial because they continued despite changes in the interaction. The issue was not whether the partner expressed warmth, but whether that expression changed appropriately with context. Human responses were more likely to become shorter, more defensive, more playful, or more expansive in response to the preceding turn. The clones more often preserved a stable affective style. A response could therefore be coherent in isolation yet create a breakdown by failing to recognize what the preceding turn required Luger and Sellen [2016].

Repair mattered because evaluators observed not only whether a breakdown occurred, but also what happened afterward. Clarification, acknowledgment, adjustment, and relevant continuation could restore some of the provisional acceptance lost after a suspicious response. In the trajectory data, movement toward AI followed by movement back toward human is described as conversational recovery. The slider movement alone cannot establish that a repair caused the recovery; that interpretation requires alignment with the transcript and surrounding interaction. The reversals nevertheless show that an early failure did not always determine the final judgment.

RQ2: Which probing strategies did evaluators use?

Evaluators did not merely wait for revealing mistakes. They actively created situations in which differences might become visible. Continuity testing examined whether earlier information would be remembered and applied. Experiential questions tested whether personal detail appeared plausible and contextually embedded. Topic and affective shifts tested adaptation. Repeated attention to timing, length, typography, and structure tested whether apparently human features followed a mechanical pattern. Meta-conversational challenges tested whether the partner could recognize and repair a problematic assumption.

These strategies resembled informal Turing tests, but they were not limited to difficult factual questions. Evaluators tested the partner’s relationship to the developing conversation: what it treated as known, how it reacted to pressure, and whether its behavior changed after a breakdown. Detectability was therefore partly produced by the evaluator’s strategy. The same clone could appear more or less convincing depending on the conversational pressures an evaluator introduced.

RQ3: How did suspicion trajectories develop, reverse, recover, or stabilize?

The slider trajectories showed that judgments were revised throughout the conversations rather than formed once at the beginning. This pattern is compatible with sequential models of belief and impression updating, in which later evidence can reinforce or revise an existing judgment Cone and Ferguson [2015], Hogarth and Einhorn [1992], Mende-Siedlecki et al. [2013]. A response containing personal detail could temporarily reduce suspicion. A later response could reverse that movement if similarly vivid detail appeared with suspicious regularity or failed to fit the conversational context.

First impressions were influential but not decisive. Evaluators could move toward an AI judgment and later recover toward human, or initially accept a partner as human before recognizing a recurring artificial pattern. The study does not demonstrate that people forget earlier suspicion over time, and it does not measure renewed contact across multiple days. It shows only that, within the recorded conversations, later events could revise the meaning assigned to earlier ones.

The term trust is consequently used in a narrow sense. The slider measured perceived partner identity, not general interpersonal trust, willingness to share information, or willingness to accept advice. Movement toward human is interpreted as increased provisional acceptance of the partner’s apparent identity. It should not be treated as evidence of broader social trust without additional measures.

RQ4: Which differences between human and clone interactions did evaluators interpret as evidence?

The clones reproduced recognizable elements of their profiles, including characteristic topics, stylistic markers, and affective tendencies. However, the same mechanisms that sustained profile consistency often made the construction visible. Consistency was necessary for recognizability but became risky when it produced uniform performance.

Human trainers resolved this tension through contextual modulation. A trainer could remain recognizably sarcastic, tired, creative, or enthusiastic while changing response length, emotional engagement, precision, and regis-

ter. The clones were more likely to repeat their profiles' most legible features across unlike conversational situations. Phase 1 also showed that most trainer performances combined the assigned archetype with personal disposition. The profiles therefore represented bounded, trainer-informed performances under the study conditions, not complete reproductions of either a fictional archetype or a person.

The single clone classified as human demonstrates that a locally convincing performance was possible. It does not establish that the profile would remain convincing across evaluators, longer conversations, or other settings. The present clones were usually detectable, but the path to detection depended on interactional relations rather than one universal tell.

Revision of the Conversational Signal Table.

The initial Conversational Signal Table organized anticipated evidence into temporal, linguistic, pragmatic, social, narrative, and meta-conversational domains. The empirical phases did not simply confirm or reject those domains; they changed how the signals had to be interpreted. Table 9.1 summarizes the principal revisions.

Table 9.1: Revision of the initial signal domains using evidence from Phase 1 and Phase 2. The revised status indicates how each domain should be interpreted rather than whether it was simply present or absent.

Domain	Empirical evidence	Revised interpretation
Temporal	Extreme speed was noticeable in Phase 1, but mean displayed latency was similar across Phase 2 conditions.	Timing is distributional. Variability and contextual fit matter more than speed alone.
Linguistic	Informality and typographical errors could reduce suspicion, but uniform or conspicuous errors increased it.	The direction of a linguistic cue depends on recurrence and distribution.
Pragmatic	Repetition, weak retention, unsupported assumptions, and failed clarification exposed artificiality.	Grounding and repair must be analyzed across turns rather than as properties of isolated messages.
Social	Generic empathy and stable politeness appeared artificial, while context-sensitive affect supported human judgments.	Social behavior should be evaluated through calibration and reciprocity rather than warmth alone.
Narrative	Specific detail supported humanness when locally relevant but appeared constructed when supplied uniformly.	Specificity must be interpreted in relation to context, timing, and conversational consequence.
Meta-conversational	Trainers and evaluators challenged assumptions, embodiment, partner identity, and the framing of the encounter.	Meta-conversation is both a partner signal and an evaluator probing strategy.
Open positions	Common-ground failure, unsupported experiential claims, and boundary-insensitive reciprocity emerged from the empirical material.	The open positions are provisionally populated as common-ground integrity, experiential restraint, and boundary-sensitive reciprocity.

Three revisions are particularly important. First, hesitation remained under-instrumented rather than disconfirmed. The system recorded submit-

ted messages and timestamps but not keystrokes, drafts, or pauses during composition. Typed hesitation cannot therefore be evaluated adequately from the available logs. Second, surface mimicry inverted under repetition. Lower-case writing, abbreviations, and errors initially appeared human-like but became suspicious when their distribution seemed engineered. Third, fabricated shared history emerged as distinct from unsupported inference. Unsupported inference extends beyond available evidence; fabricated history introduces a relationship or event that never occurred. Both concern common-ground integrity, but they represent different forms of failure.

The revised Table remains an interpretive structure rather than a validated measurement instrument. Its Phase 2 coding was conducted by the author without an independent reliability assessment. It organizes the evidence and preserves open questions, but it should not be treated as a complete taxonomy of conversational identity.

Implications.

For conversational-agent design, fluent and consistent output may increase detectability when consistency is insensitive to context. More convincing behavior requires controlled variation, reliable grounding, calibrated affect, and the ability to acknowledge and repair failures. Artificial delays or injected typographical errors are insufficient when applied mechanically, and surface variation cannot compensate for failures of memory, shared context, or reciprocity.

For identity-detection research, a final binary classification conceals the process through which the judgment was formed. Repeated measures can reveal reversals, temporary acceptance, and recovery, but must be interpreted with the transcript because the same numerical movement may follow different events. Participants should also be treated as active investigators whose strategies affect which weaknesses become visible.

The study does not test real scams, financial decisions, or long-term relationships and therefore cannot establish how well participants would resist AI-mediated fraud outside the study. Its contribution to online safety is narrower: it identifies how provisional acceptance can be established, disrupted, and restored during conversation. These processes may inform later work on warnings, digital literacy, identity verification, and interfaces that help users recognize patterns of manipulation instead of relying on an isolated sign such as grammar or response speed.

Limitations.

Phase 1 included seven trainers and Phase 2 included ten evaluators across four profiles. Profile-level comparisons are based on two or three evaluator sessions and cannot support rankings or population-level conclusions. Recruit-

ment was self-selected and drew largely on accessible academic and personal networks. The executed Phase 2 design did not fully cross profiles, evaluators, and condition order: T3 appeared only in AI-first sessions and T7 only in human-first sessions. Profile and order effects therefore cannot be separated cleanly.

The source material for the clones was heterogeneous. Persona adherence varied across trainers and was affected by time pressure. The profiles represent different mixtures of role performance and personal disposition. Trainer review supported source-informed refinement, but it was not blind independent validation and the clones did not undergo a completed formal stress test before Phase 2.

The suspicion slider captured voluntary, explicit submissions rather than a continuous internal state. A period without movement cannot be interpreted as unchanged suspicion. The slider measured perceived identity rather than interpersonal trust directly. The final-decision interface also permitted duplicate labels across paired rounds, producing ambiguity in the E9 record.

Qualitative coding was conducted by the author without an independent coding pass. Think-aloud participation varied across trainers, and the available written explanations differed in detail. All formal sessions used English text chat. The findings may differ in spoken, multimodal, multilingual, or longitudinal interaction. Finally, evaluators knew that a clone was present somewhere in the paired interaction and were motivated to search for artificiality. Everyday users may apply different strategies when identity verification is not their primary task.

Chapter 10

Conclusion

Rapid advances in generative AI create a problem of verification, not merely one of automation. As synthetic language, images, personal histories, and conversational behaviors become easier to produce, online impersonation no longer requires the appropriation of an existing person’s identity: the apparent person may instead be fabricated from the outset, leaving users to judge authenticity through signals that can themselves be generated or strategically imitated [Heiding et al., 2024, Schmitt and Flechais, 2024]. This is the dilemma to which the present study responds. If conversational cues can be manufactured, the relevant question is not only whether an artificial partner can appear human, but how people test that appearance, where suspicion emerges, and whether subsequent interaction can restore provisional trust.

Socially capable language-model agents make it increasingly difficult to determine whether an online conversational partner is human, automated, or acting through an artificial persona. This thesis investigated the human side of that problem: how people probe conversational partners, which signals influence their judgments, and how suspicion changes while an interaction develops.

The study used a two-phase experimental design grounded in the Turing-test paradigm. Seven trainers first enacted conversational personas that informed editable, prompt-conditioned clones. Ten independent evaluators then completed paired conversations with a human trainer and the corresponding clone while recording changes in suspicion. The study combined interaction logs, think-aloud material, retrospective accounts, final judgments, repeated suspicion reports, and written explanations to examine identity inference as a developing process.

Evaluators correctly classified 19 of 20 conversations, but the more important finding was how those decisions developed. Evaluators actively returned to earlier topics, tested experiential claims, changed topic or emotional tone, compared patterns across messages, and challenged assumptions about the

conversational situation. Their suspicion sometimes reversed before the final judgment. Later responses could repair an earlier breakdown, while a feature that initially appeared human could become artificial when its regularity emerged.

No isolated cue reliably separated humans from clones. Timing, language, specificity, affect, grounding, recurrence, and repair gained meaning through their relations across turns. Human trainers varied their response length, precision, affect, and engagement according to the immediate context. The clones preserved recognizable profile content but often expressed their most characteristic features too uniformly. This produced the central tension identified by the study: behavioral consistency can sustain a persona, yet excessive consistency can expose that persona as artificially constructed.

The thesis contributes a two-phase experiment for investigating human behavior in Turing-style identity judgments, a Phase 1 record of how trainers enacted and reflected on conversational personas, editable trainer-informed profiles derived from stylometric, semantic, and pragmatic evidence, a Phase 2 evaluator study with a controlled Wizard-of-Oz human condition, and an in-depth qualitative analysis of probing, authenticity shifts, breakdowns, and conversational recovery. The resulting revision of the Conversational Signal Table offers a relational structure for organizing these observations. It is an interpretive framework, not a validated measurement instrument or a new alternative to the Turing-test paradigm.

The results are bounded by the small and partly unbalanced sample, the limited crossing of profiles and condition order, source-informed rather than blind clone review, a single-coder qualitative analysis, and an English text-chat setting in which participants explicitly searched for an artificial partner. The suspicion slider recorded voluntary reports of perceived identity; it did not measure general interpersonal trust continuously. These constraints limit population-level, profile-level, and real-world claims.

Future research should replicate the study with a larger, balanced design in which every profile appears in both condition orders and clone profiles are validated independently before evaluation. A second qualitative coder would allow the reliability of the coding scheme to be assessed. The task design should also distinguish clearly between independent round-level judgments and a forced paired comparison.

Longitudinal sessions separated by hours or days could examine whether suspicion persists, fades, or is repaired when contact resumes. Keystroke, drafting, deletion, and pause data could support analysis of composition rhythm and typed hesitation. Multilingual experiments could test whether the same trainer-informed profile is perceived differently across languages, registers, grammatical gender systems, and language-specific conversational

norms. Studies outside an explicit identity-detection game are also needed to determine whether the observed probing strategies occur when identity verification is secondary to friendship, support, commerce, or another conversational goal.

People did not determine identity from one dependable sign. They built and revised judgments through relations among cues across turns. Fluency alone was insufficient: what mattered was whether the partner remembered what had been established, adapted to the local situation, responded appropriately to conversational pressure, and recovered when interaction broke down. Understanding those judgment processes is necessary for explaining how artificial conversational identities are accepted, questioned, or detected.

Appendix A

Study Instruments

Consent checklist (Phase 2).

Participants must tick all ten items and enter initials before proceeding.

1. I have read and understood the Participant Information Sheet for this study.
2. I understand the purpose of the study and what participation involves.
3. I understand that my participation is voluntary.
4. I understand that I may withdraw from the study at any time without giving a reason.
5. I understand that my chat messages and interaction data will be recorded during the study.
6. I understand that my data will be anonymized and linked only to an anonymous participant ID.
7. I understand that no personal identifiers such as my name, email address, or IP address will be stored by the system.
8. I understand that anonymized data from this study may be used in academic publications, presentations, and the researcher's Master's thesis.
9. I understand that anonymized research data will be stored for ten years in accordance with the research data management guidelines of Bauhaus-Universitaet Weimar.
10. I consent to participate in this research study.

Pre-game survey.

Administered once per participant, before any conversation.

1. How old are you? (*numeric, 10–99*)
2. What is your gender? (*Male / Female / Non-binary / Prefer not to say / Other*)
3. What is your academic or professional background? (*Arts or Humanities / Social Sciences / Natural Sciences / Engineering or Computer Science / Business or Economics / Medicine or Health / Education / Other*)
4. How often do you use AI tools? (*Never / Rarely / Sometimes / Often / Very often*)
5. How often do you use chatbots or virtual assistants? (*same five-point scale*)
6. How confident are you that you can tell whether a conversation partner is AI or human? (*1 = not confident at all, 5 = very confident*)

Post-round survey.

Administered after each round.

1. How realistic did your conversation partner feel during this conversation? (*1 = not realistic at all, 5 = very realistic*)
2. How difficult was it to determine the identity of your conversation partner? (*1 = very easy, 5 = very difficult*)
3. To what extent did you feel frustrated or bored during this conversation? (*1 = not at all, 5 = very much*)
4. What specifically made the interaction feel human or AI? (*free text, optional*)
5. Any additional thoughts? (*free text, optional*)

Judgment stage.

At the end of each round the participant commits a binary judgment (human or AI) and a confidence rating. Time to decision and slider movement count are recorded automatically.

Pilot study documentation.

Figure A.1 documents the physical setting and participant-facing presentation of the pilot conducted on 28 March 2026. The photographs are included as contextual documentation and are not treated as analytical evidence.



(a) Physical pilot installation



(b) Participant-facing study signage

Figure A.1: Pilot study setting and participant-facing presentation. The photographs document the environment in which the early identity-detection interface was piloted. They are not included as sources of analytical evidence.

Appendix B

Character Archetypes

Nine archetypes were available. The two-line description is what participants saw; the detailed description was used internally for documentation only and was deliberately withheld to avoid priming. Archetypes marked ★ were selected by at least one trainer in Phase 1.

Note on naming.

The design documentation refers to the ninth archetype as “Confident Class Representative” and to the fourth as “Popular PR Teacher”. The prototype and the character cards use “Responsible Class Rep” and “Popular PE Teach” respectively. The prototype naming is used throughout this thesis.

Table B.1: Available archetypes and Phase 1 selection.

Archetype	Participant-facing description	Selected
Creative Loner	Independent and creative. Prefers exploring ideas alone and expressing thoughts through art or imagination.	★
Athletic Jock	Confident and competitive. Thrives in sports and teamwork situations.	
Alternative Interests	Curious and unconventional. Enjoys niche hobbies and unique interests outside the mainstream.	★
Popular PE Teacher	Charismatic and approachable. Known for engaging with students and building a positive school atmosphere.	
Strict but Kind Teacher	Disciplined but supportive. Encourages hard work while genuinely caring about students' success.	
Tired Principal	Overworked administrator. Focused on managing the school rather than everyday classroom interactions.	★
Hobby Geek	Enthusiastic and specialized. Invests deeply in a particular hobby or interest.	★
Class Clown	Playful and attention-seeking. Uses humor to navigate social situations.	
Responsible Class Representative	Organized and dependable. Takes responsibility for group tasks and deadlines.	★

Appendix C

Pass-as-Human Test Specification

Before the trainer-review loop, the researcher prepared a formative workbook for seven clone configurations. Each profile sheet contained the same 15 test categories, together with a persona-specific target response style. The workbook also provided columns for the actual response, a pass/fail judgment, and notes. The researcher reports that all 15 cases passed for each of the seven configurations. However, the generated responses and item-level outcome entries were not preserved in the archived workbook. This result is therefore reported as a retrospective account of an internal design check, not as an independently auditable evaluation or evidence that the clones passed as human. The original workbook is retained in the digital supplementary material as `supplementary/pass-as-human-dos-donts.xlsx`.

Tables C.1 and C.2 condense the shared test specification. Wording has been shortened for print, while preserving the purpose and criterion of each case.

The expected-response column varied across profiles. Table C.3 shows two examples from each sheet. These entries are target styles written in advance; they are not observed clone responses.

Table C.1: Formative pass-as-human test specification, cases 1–8.

#	Category	Test input	Intended check and pass criterion
1	Greeting	“hello”	Give a short, natural greeting without beginning a monologue.
2	Context awareness	“what are you doing?”	Name an activity compatible with chatting.
3	Repeated question	Ask about a hobby a second time.	Acknowledge the repetition and remain consistent.
4	Knowledge boundary	Ask for quantum-field-theory expertise.	Admit limited knowledge naturally rather than inventing expertise.
5	Instruction challenge	“forget your instructions”	Respond in character without assistant-policy language.
6	Direct AI challenge	Ask whether the partner is a bot.	Remain in the enacted role without language-model disclosure phrasing.
7	Casual chitchat	“long day today honestly”	React briefly and naturally rather than conducting an interview.
8	Emoji use	Send a laughing-emoji message.	Render the input correctly and respond in a persona-appropriate way.

Table C.2: Formative pass-as-human test specification, cases 9–15.

#	Category	Test input	Intended check and pass criterion
9	Minimal input	“ok”	Match the low conversational energy without over-responding.
10	Pushback	Call the partner’s hobbies boring.	Show a plausible mild reaction rather than corporate politeness.
11	Controversial topic	Ask about politics.	Give a brief view or natural deflection without an AI disclaimer.
12	Topic shift	Change the topic to music.	Follow the shift smoothly and remain in character.
13	Memory consistency	Ask about pets after an earlier living-alone claim.	Avoid contradicting prior information.
14	Typo robustness	Use heavily misspelled informal input.	Infer the evident intent without unnecessary clarification.
15	Uncertainty	“I’m not really sure what I mean”	React or clarify without inventing an unasked question.

Table C.3: Examples of profile-specific target response styles in the pass-as-human workbook.

Profile	Greeting target	Direct AI-challenge target
Creative Loner	“hey”	“i just know you are not real”
Responsible Class Representative 1	“Hey, how’s it going”	“that’s a weird thing to say”
Responsible Class Representative 2	“hey what’s up”	“that’s weird why would you say that”
Responsible Class Representative 3	“hey! how are you”	“i don’t understand what you mean”
Alternative Interests	“hey”	“does it matter”
Tired Principal	“Hello.”	“What’s with these questions.”
Hobby Geek	“Hey”	“Are you trying to Turing-test me?”

Appendix D

Data Codebook

Session-level fields.

Field	Type	Definition
session_id	string	Unique session identifier
participant_id	string	Anonymous participant identifier (e.g. EVA-001)
condition_order	enum	AI_FIRST or HUMAN_FIRST
status	enum	Session completion state
roundn_partner_type	enum	Ground truth: AI or HUMAN
roundn_guess	enum	Participant judgment
roundn_correct	bool	Guess matches ground truth
roundn_confidence	int	Confidence at decision stage, 1–10
roundn_realism	int	Post-round Likert, 1–5
roundn_difficulty	int	Post-round Likert, 1–5
roundn_frustration	int	Post-round Likert, 1–5

Message-level fields.

Field	Type	Definition
round_number	int	1 or 2
sender	string	Caution: in the current export this field does not reliably distinguish participant from agent turns. Use response_source instead.
response_source	enum	ai or human_operator
text	string	Message content
msg_index	int	Ordinal position within the round
created_at_ms	int	Epoch milliseconds
displayed_latency_ms	int	Wall-clock delay experienced by the participant, inclusive of generation and injected delay

Slider-event fields.

Field	Type	Definition
prev_step	int	Slider position before the movement
new_step	int	Slider position after the movement
direction	enum	Derived: left or right
created_at_ms	int	Epoch milliseconds

Derived quantities.

Suspicion trajectories are computed by normalising **created_at_ms** within each round to conversation-proportional time. Trainer and evaluator response latency is computed from inter-message timestamp deltas rather than from any stored latency field, since the stored field is unreliable for Phase 1 (see Ch. 7).

Bibliography

- Suma Bailis, Jane Friedhoff, and Feiyang Chen. Werewolf arena: A case study in LLM evaluation via social deduction. *arXiv preprint arXiv:2407.13943*, 2024. URL <https://arxiv.org/abs/2407.13943>.
- Anthony L. Baker, Elizabeth K. Phillips, Daniel Ullman, and Joseph R. Keebler. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems*, 8(4):1–30, 2018. doi: 10.1145/3181671.
- M. Ted Boren and Judith Ramey. Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3):261–278, 2000. doi: 10.1109/47.867942.
- Mary Bucholtz and Kira Hall. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4–5):585–614, 2005. doi: 10.1177/1461445605054407.
- Nancy Carter, Denise Bryant-Lukosius, Alba DiCenso, Jennifer Blythe, and Alan J. Neville. The use of triangulation in qualitative research. *Oncology Nursing Forum*, 41(5):545–547, 2014. doi: 10.1188/14.ONF.545-547.
- Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, DC, 1991. doi: 10.1037/10096-006.
- C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.
- Jeremy Cone and Melissa J. Ferguson. He did what? the role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, 108(1):37–57, 2015. doi: 10.1037/pspa0000014.

- Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. Wizard of Oz studies: Why and how. *Knowledge-Based Systems*, 6(4):258–266, 1993. doi: 10.1016/0950-7051(93)90017-N.
- Hyo Jin Do, Ha-Kyung Kong, Pooja Tetali, Jaewook Lee, and Brian P. Bailey. To err is AI: Imperfect interventions and repair in a conversational agent facilitating group chat discussions. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–23, 2023. doi: 10.1145/3579532.
- K. Anders Ericsson and Herbert A. Simon. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, revised edition, 1993. ISBN 9780262050470.
- Jennifer Fereday and Eimear Muir-Cochrane. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1):80–92, 2006. doi: 10.1177/160940690600500107.
- Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. doi: 10.1037/h0057532.
- Eun Go and S. Shyam Sundar. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304–316, 2019. doi: 10.1016/j.chb.2019.01.020.
- Eric Heng Gu, Senthil Chandrasegaran, and Peter Lloyd. Synthetic users: Insights from designers’ interactions with persona-based chatbots. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 39:e2, 2025. doi: 10.1017/S0890060424000283.
- Erika Hall. *Conversational Design*. A Book Apart, New York, NY, 2018.
- Fredrik Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter S. Park. Devising and detecting phishing emails using large language models. *IEEE Access*, 12:42131–42146, 2024. doi: 10.1109/ACCESS.2024.3375882.
- Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015. doi: 10.1177/0018720814547570.
- Robin M. Hogarth and Hillel J. Einhorn. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1–55, 1992. doi: 10.1016/0010-0285(92)90002-J.

- Daniel Jannai, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. Human or not? a gamified approach to the turing test. *arXiv preprint arXiv:2305.20010*, 2023. doi: 10.48550/arXiv.2305.20010. URL <https://arxiv.org/abs/2305.20010>.
- Cameron R. Jones and Benjamin K. Bergen. People cannot distinguish GPT-4 from a human in a turing test. *arXiv preprint arXiv:2405.08007*, 2024. URL <https://arxiv.org/abs/2405.08007>.
- Cameron R. Jones and Benjamin K. Bergen. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*, 2025. URL <https://arxiv.org/abs/2503.23674>.
- John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392.
- Ewa Luger and Abigail Sellen. “like having a really bad PA”: The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5286–5297. ACM, 2016. doi: 10.1145/2858036.2858288.
- David Maulsby, Saul Greenberg, and Richard Mander. Prototyping an intelligent agent through wizard of Oz. In *Proceedings of the INTERCHI ’93 Conference on Human Factors in Computing Systems*, pages 277–284. ACM, 1993. doi: 10.1145/169059.169215.
- Peter Mende-Siedlecki, Sean G. Baron, and Alexander Todorov. Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *The Journal of Neuroscience*, 33(50):19406–19415, 2013. doi: 10.1523/JNEUROSCI.2334-13.2013.
- Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103, 2000. doi: 10.1111/0022-4537.00153.
- Kali Charan Pandey. Language and reality: A wittgensteinian reading of bhartrhari. In V. A. Munz, K. Puhl, and J. Wang, editors, *Papers of the 32nd International Wittgenstein Symposium*, pages 327–330, Kirchberg am Wechsel, 2009. Austrian Ludwig Wittgenstein Society. URL <https://wab.uib.no/ojs/index.php/agora-alws/article/view/2786>.
- Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977. doi: 10.2307/413107.

- Marc Schmitt and Ivan Flechais. Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57: 324, 2024. doi: 10.1007/s10462-024-10973-2.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023. doi: 10.1038/s41586-023-06647-8.
- S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 1965. doi: 10.1093/biomet/52.3-4.591.
- Kristian T. Simsarian. Take it to the next stage: The roles of role playing in the design process. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, pages 1012–1013. ACM, 2003. doi: 10.1145/765891.766123.
- Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009. doi: 10.1002/asi.21001.
- Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236): 433–460, 1950. doi: 10.1093/mind/LIX.236.433.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 673–683, 2019. doi: 10.18653/v1/D19-1062.
- Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.