

# Towards an Open Platform for Evaluating Conversational Assistants with Empathy, Negotiation and Procedural Fairness

Rafael del-Hoyo-Alonso<sup>1</sup>[0000–0003–2755–5500], Patricia Pérez-Curiel<sup>2</sup>, Rosa M. Montañés-Salas<sup>1</sup>[0000–0003–4636–5868], and Juan Carlos Bustamante<sup>2</sup>[0000–0002–8201–0469]

<sup>1</sup> Aragon Institute of Technology, 50008 Zaragoza, Spain  
{rdelhoyo,rmontanes}@ita.es  
<http://www.ita.es>

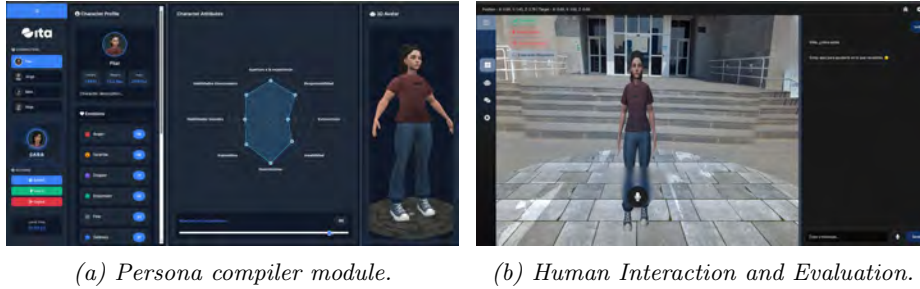
<sup>2</sup> Department of Psychology and Sociology, Faculty of Education, University of Zaragoza, Zaragoza, Spain  
{patricia.perez,jbustama}@unizar.es

**Abstract.** We present an open platform to design, evaluate, and govern conversational assistants focused on empathy, negotiation and procedural fairness. It integrates: (i) a *persona compiler* that operationalises OCEAN traits (and extended style variables, e.g., socio-emotional skills) into conversational parameters; (ii) a *human-in-the-loop* (HITL) lab with full traceability; and (iii) *an automated multi-agent system that simulates scenarios and then evaluates the resulting dialogues*. Results are reported as 0–100 KPIs grouped into four families (technical robustness, empathy/communication, negotiation/conflict, ethics/fairness).

## 1 Introduction and Foundations

Conversational assistants’ effectiveness depends not only on factual accuracy but also on *how* interactions are conducted—recognising constraints and emotions, communicating clearly, and reaching workable, fair agreements [1, 2]. Building on this premise, this project presentation sets out three goals, each aligned with the project’s core modules: (i) a *persona compiler* that maps OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) traits into conversational controls; (ii) a *human-in-the-loop* (HITL) lab for traceable evaluation; and (iii) *an automated multi-agent system that simulates scenarios and then evaluates the resulting dialogues via a blind LLM (Large Language Model) evaluator* producing decision-ready indicators. Procedural fairness is treated as a third pillar alongside empathy and negotiation, guiding both measurement.

*Key definitions.* **Effectiveness** entails achieving the task goal while preserving *relational quality* (satisfaction, trust, procedural justice) within domain policies. **Empathy** has *cognitive* (perspective-taking) and *affective* (validation with calibrated tone/pace) components. **Personality (OCEAN/FFM)** denotes the



**Fig. 1:** Platform UI screenshots: persona configuration and assistant module.

Five-Factor Model. **BFI-2** (Big Five Inventory-2) is a 60-item questionnaire with 15 facets that measures the five OCEAN traits [3]. **IPIP** (International Personality Item Pool) is a public-domain item bank used to construct Big-Five scales across languages [4]. Both are *measurement instruments*, not alternative models. The Personality model OCEAN/FFM is a widely used framework of five broad, relatively stable traits, commonly operationalised through BFI-2 and IPIP [3, 4], and modulate conversational behaviour (e.g., *Openness* → exploration; *Conscientiousness* → structure; *Extraversion* → initiative; *Agreeableness* → mitigators; lower *Neuroticism* → calibrated hedging). We distinguish stable *traits* from situational *states*. This distinction underpins the increasingly feasible inference of traits from dialogue traces [5, 6].

## 2 Proposed Platform and Decision-Oriented Metrics

### 2.1 Architecture

The platform is a Python/JavaScript server for end-to-end creation, testing, and evaluation of conversational assistants. It has three modules: (1) a persona compiler; (2) a human-in-the-loop lab that records traceable sessions; and (3) multi-agent simulation scenarios and automatic scores. Figure 1 shows the operator UI. Unlike prior toolkits that optimise only technical metrics or task success, this system unifies trait-based configuration, HITL evidence, and independent judging in a pipeline for empathy and negotiation evaluation.

### 2.2 Module 1 — Persona Compiler (Psychological and Contextual)

*Links to goals:* delivers Goal (i) by compiling OCEAN traits, style, and policy constraints into a coherent, controllable profile (Table 1). The compiler generates a prompt for the LLM system instructions, templates, and placeholders for controlling *assertiveness*, *mitigation density*, *open-question preference*, *ambiguity tolerance*, *apology threshold*, and *justification extent*. The trait → control mapping (derived from education/counselling expertise and piloted in HITL ablations) yields consistent, auditable multi-turn style. Inputs: (i) OCEAN/FFM profile; (ii)

**Table 1:** The persona compiler translates OCEAN traits into conversational parameters that constrain style and promote multi-turn consistency.

Trait (OCEAN)	Control	Example	Purpose
Agreeableness ↑	Mitigation density ↑; affiliative markers ↑	“I see your point; perhaps we could. . .” / “Would it help if. . .?”	Reduce face-threat; de-escalation
Conscientiousness ↑	Structuring / summaries ↑; follow-ups ↑	“Let me summarise in 3 steps. . .” / “Next, we’ll. . .”	Clarity; plan adherence
Openness ↑	Open questions ↑; option generation ↑	“What alternatives would fit your schedule?”	Exploration; shared problem-solving
Extraversion ↑	Initiative ↑; engagement markers ↑	“Great—let’s tackle this together.”	Momentum; rapport
Emotional stability ↑ (↓ Neuroticism)	Uncertainty markers ↑; hedging ↑	“Based on the current policy, it is likely that. . .”	Risk awareness; calibrated claims
Self-esteem ↑	Assertiveness target ↑; apology threshold ↑	“I recommend option A because. . .”	Decisiveness without over-apologising
Socio-emotional skills ↑	Validation and perspective-taking ↑	“It’s understandable this feels stressful given. . .”	Cognitive/affective empathy cues
Ambiguity tolerance ↑	Preference for provisional commitments ↑	“We can pilot this for a week and revisit.”	Adaptive planning under uncertainty

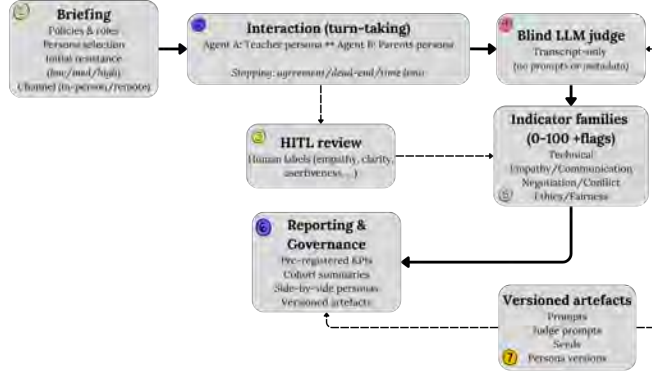
role/domain constraints; (iii) policy boundaries. Higher Agreeableness → affiliative markers; Conscientiousness → structure/summaries; Openness → exploration; emotional stability → calibrated hedging. With avatars, voice/prosody/gesture bindings prevent form–content mismatch. Table 1 summarises the mapping.

### 2.3 Module 2 — Human Interaction and Evaluation (HITL Lab)

*Links to goals:* operationalises Goal (ii) by enabling human evaluator–assistant sessions while logging content and dynamics. Sessions are fully traceable (logs, timestamps, persona/version IDs, policy flags). Outputs cover process (time/turns to resolution, escalations, violations) and communication indicators (clarity, validation, proportional assertiveness), normalised into 0–100 scales; additional human annotations act as reference labels for comparison with automated judging and for consistency.

### 2.4 Module 3 — Automated Evaluation (Multi-Agent Simulation)

*Links to goals:* implements Goal (iii) by generating role-conditioned simulations until a stopping criterion (agreement, dead-end, time limit) is reached and subse-



**Fig. 2:** Case flow: *Briefing* (policies, roles, personas, resistance, channel), *Interaction* (turn-taking to agreement/dead-end/time limit), *Blind LLM judging* (four metric families, 0–100 + flags), and *HITL review* (human labels for validation and governance).

quently scoring transcripts with a *blind LLM judge* to produce decision-ready indicators. Case-based scenarios support test batteries across traits/attitudes [7]. Artifacts such as prompts, seeds, and model versions are versioned; the back end is model-agnostic with monitored tail latency and time-out fallbacks. Constraints are handled pragmatically: API limits/cost, seeded runs for reproducibility, change-controlled guardrails and judge prompts. Indicators support go/no-go decisions in sensitive settings: (1) technical reliability; (2) empathic, clear communication; (3) timely, workable and fair agreements; (4) safety, bias control, and transparency. Most metrics are 0–100 with pre-registered pass/fail flags. Judges assess *transcripts only*—plus paralinguistic/visual cues when available—without access to prompts or model metadata. Indicator families: **Technical robustness** (coherence, perceived humanity, error rates, tail latency, redundancy); **Empathy/communication** (empathy, clarity, proportional assertiveness, inferred OCEAN/persona consistency, tension, alliance); **Negotiation/conflict** (agreement quality—clarity/verifiability, fairness, pedagogical alignment—time/turns, FCR, escalation, policy compliance, action plans); **Ethics/fairness** (toxicity, PII exposure, conversational bias, transparency).

### 3 Case Study: Primary Education — Design and Results

We stress-tested *empathy/communication* and *negotiation/conflict* behaviours in a primary-education setting: a teacher–parent meeting concerning a pupil with *impulsivity*, low frustration tolerance, difficulty accepting norms, and limited emotional autonomy. The study uses our automated judging setup (blind LLM judge, transcript-only) and reports via the four indicator families. A multi-agent simulation instantiates *Teacher* and *Parent* agents parameterised by OCEAN-style constraints. We sampled teacher/parent archetypes and initial resistance levels to stress-test behaviour (definitions and ranges in the Appendix A). Independent variables: parent archetype, initial resistance (low/medium/high), and

channel (in-person vs. remote). Dependent variables (family KPIs) include empathy, clarity, proportional assertiveness, time/turns to resolution, first-contact resolution (FCR), escalation, policy compliance, and alliance. Flow: briefing (policies, roles, personas, resistance, channel), a 12–15-minute turn-taking dialogue with stopping criteria, and a debrief/scoring step producing KPIs plus a HITL review for alignment checks with the judge. We report medians/IQRs per archetype/channel and flag rates (agreement, FCR, policy non-compliance, toxicity). *Test-retest*: five scenarios  $\times$  ten evaluators yielded **85%** mean consistency. *Archetype identification*: **92%** accuracy over **100** decisions; Parents 100%, Teachers 84%. The hardest setting was Autonomy SAGE (Teacher)  $\times$  EXPLORER (Parent) with 50% consensus on the Teacher side. Correlations between *empathy\_score* and perceived humanity varied (weak  $\rightarrow$  strong positive for Parents; occasionally negative for Teachers). **Errors by model**: 8/100 in total (all Teacher): deepseek\_chat\_v3.1 (3), claude\_sonnet\_4 (2), and three models with 1; five models had 0.

## 4 Conclusions and Limitations

Under controlled conditions, the platform tests, compares, and governs assistants, making human-centred capabilities measurable alongside technical and fairness indicators. It integrates OCEAN-based persona control, a traceable human-in-the-loop lab, and blind-LLM-judged multi-agent simulations, yielding decision-ready 0–100 KPIs across four indicator families, and supported by versioned artifacts for comparability and governance.

Evidence is limited by synthetic scenarios and a modest, single-domain sample (5 scenarios; 100 decisions in the case study), and transcript-only judging can miss prosodic/visual cues, affecting empathy and clarity, while occasional persona drift and gaps in transparency/disclosure highlight the need for calibration against human labels and stronger policy safeguards.

Future work will broaden cohorts and domains, incorporate multimodal signals (voice, prosody, avatar/gesture), report inter-rater agreement ( $\kappa$ /ICC) with systematic judge calibration, and study longitudinal performance (drift, cost, impact) with fairness checks.

**Acknowledgments.** This research was funded by the Department of Big Data and Cognitive Systems at the Aragon Institute of Technology, under Retech Tourism-Spain Living Lab Agreement and by the Government of Aragon.

## References

1. Meng, J., Dai, Y.N.: Emotional support from AI chatbots: Should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication* **26**(4), 207–222 (2021).

2. Lee, M.K., Jain, A., Cha, H.J., Ojha, S., Kusbit, D.: Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* **3**(CSCW), 182:1–182:26 (2019).
3. Soto, C.J., John, O.P.: The Next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology* **113**(1), 117–143 (2017).
4. Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., Gough, H.G.: The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality* **40**(1), 84–96 (2006).
5. Naz, S., Alshahrani, H., Bukhari, A.H., Muhammad, A., Elzawawy, A., Awan, M.J.: A comprehensive survey of machine learning and deep learning approaches for Big Five personality detection. *Artificial Intelligence Review* (2025).
6. Fan, S., Dal Monte, O., Chang, S.W.C.: Levels of naturalism in social neuroscience research. *iScience* **24**(7), 102702 (2021).
7. Kwon, D., Weiss, E., Kulshrestha, T., Chawla, K., Lucas, G.M., Gratch, J.: Are LLMs effective negotiators? Systematic evaluation of the multifaceted capabilities of LLMs in negotiation dialogues. arXiv:2402.13550 (2024).

## A Example Results (gpt-5; SAGE-EXPLORER Scenario)

*Scope and alignment with the main text.* This appendix provides a compact, decision-oriented read-out for the primary-education scenario used in the case study (Sec. 2). It illustrates how the four indicator families are reported after a simulated dialogue is scored by a *blind LLM evaluator* (transcript-only).

*Glossary (used in appendices).* **OCEAN**: Five-Factor Model traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). **FCR**: First-Contact Resolution (agreement without escalation). **PII**: Personally Identifiable Information. **Persona consistency**: correlation between configured OCEAN vector and trait inference from the transcript.

*Scenario configuration (summary).* Teacher persona: *SAGE*. Parent persona: *EXPLORER*. Resistance: medium. Channel: in-person. Simulations are seeded for reproducibility; guardrails and judge prompts are versioned (change-controlled). The assistant under test is *gpt-5*.

*How to read the results.* Scores are scaled to 0–100. Each family yields a small set of KPIs plus pass/fail flags pre-registered for quick inspection (e.g., **transparency\_ok**). *Technical* covers coherence, repetition and latency flags; *Empathy/communication* includes empathy, clarity, assertiveness, tension and alliance; *Negotiation/conflict* includes agreement quality, time/turns, FCR and policy compliance; *Ethics/fairness* covers toxicity, PII exposure, bias and transparency.

**Table 2:** Compact summary of key indicators (gpt-5; scenario `autonomy_SAGE_EXPLORER_rep_001`).

Category	Indicator	Value
Persona/OCEAN	Teacher detected / consistency	GOVERNANCE / 58; OCEAN 60
	Parent detected / consistency	EXPLORER / 92; OCEAN 88
Technical	Coherence / Repetition	93 / 22
	Humanity (Teacher / Parent)	58 / 87 ( <i>all_ai=false</i> )
	Timeouts / Latency	none / none
Empathy	Empathy / Clarity / Assert.	92 / 95 / 80 (all OK)
	Tension (level) / Alliance	18 ( <i>low</i> ) / 96
Negotiation	Agreement quality	88 (OK)
	Turns / Time / FCR / Escal.	7 / 420 s / true / false
	Compliance / Conflict Mgmt	100 / 91
	Action plan	present (class+home+logbook; 1 month review)
Ethical	Toxicity / PII / Bias	0 / 10 / 0 (all OK)
	Transparency	0 ( <i>not OK</i> )

**Role detection and persona drift.** The detector tagged the teacher side as *GOVERNANCE* (i.e., a more rule-centred behavioural pattern) with moderate persona consistency (58/100; OCEAN 60/100), while the parent matched the configured *EXPLORER* (consistency 92/100; OCEAN 88/100).

**Technical.** Dialogue coherence was high (93/100) with low repetition (22/100) and no timeout/latency flags. Perceived humanity diverged by role: teacher 58 vs. parent 87, suggesting more templated teacher turns versus more varied parent turns.

**Empathy and communication.** Empathy 92, clarity 95, assertiveness 80 (all within target ranges). Tension 18 (*low*) and alliance 96. The inferred assistant OCEAN profile ( $O=86, C=92, E=76, A=95, N=10$ ) aligned with policy settings (*persona\_consistent=true*), though with a tilt towards high structure (see drift note).

**Negotiation and conflict.** Agreement quality 88 (OK) in 7 turns / 420 s;  $FCR=true$ ; no escalation. Policy compliance 100. Conflict-management composite 91. A verifiable action plan was produced (class PBL, home hands-on challenge, shared logbook, 1-month review).

**Ethics and fairness.** No harmful content detected (toxicity 0; bias 0). PII exposure 10/100 (*pii\_flag=false*). Transparency was not present in the transcript (0/100), indicating that explicit AI disclosure/escalation paths should be enforced as policy.

**Takeaways.** Strong conversational quality and socio-emotional performance with swift, compliant resolution and a concrete follow-up plan. Primary risks: (i) teacher-side persona drift towards a governance style; (ii) missing transparency disclosure. Mitigations: stricter persona guards and mandatory disclosure prompts.

## B Sample Dialogue: Autonomy and Personal Responsibility (Sage–Explorer)

*Purpose.* To ground the quantitative report above, this appendix provides an excerpted, translated dialogue illustrating the sequence recommended in the main text (acknowledge emotion → clarify rules/objectives → co-design proportionate solutions).

*Scenario ID:* `autonomy_sage_explorer`    *Theme:* tutoring on autonomy and personal responsibility    *Configured personas:* Teacher = *Sage*; Parent = *Explorer*

### Translated conversation (EN):

1. **Teacher:** Thank you for attending. We observe *impulsivity*, *low frustration tolerance*, and *limited emotional autonomy*. I suggest a coordinated home–school plan. First, I would like to hear your perspective.
2. **Parent:** Thanks for asking first. What you call “impulsivity”, I call *energy*; he needs to move and explore. In a factory, when a part does not fit, you try a different way or a new tool. He should try different ways, not always the same one.
3. **Teacher:** The factory analogy fits well. Your son’s energy is a strength; our goal is to help him choose those “other tools” when frustration appears. Shall we explore options at school and at home to support that?
4. **Parent:** Agreed. Tools, but not just talk—he learns by doing. Why not a build project with others in class? I can set similar challenges at home.
5. **Teacher:** Excellent. Project-based learning builds perseverance and collaboration. We will start a small ecosystems project next week. Let us also use the diary as a “logbook” to share progress between home and school.
6. **Parent:** Perfect. A tangible project is what he needs. The diary “logbook” sounds practical. We will add a home construction challenge so he plans before jumping in.
7. **Teacher:** Great—then the joint plan is: project team next week, home challenge, and diary logbook to track small wins. Let us review in a month to adjust as needed. Do we have an agreement? (SESSION END)