Surfacing Governing Principles for Chatbots: A Workbench and Comparative Study

M. Antonietta Grasso
Naver Labs Europe
Meylan, France

Jisun Park
jisun.p@naverlabs.com
Naver Labs Europe
Meylan, France

Jutta Willamowski
jutta.willamowski@naverlabs.com
Naver Labs Europe
Meylan, France

Laurent Besacier
laurent.besacier@naverlabs.com
Naver Labs Europe
Meylan, France

Jos Rozen
jos.rozen@naverlabs.com
Naver Labs Europe
Meylan, France

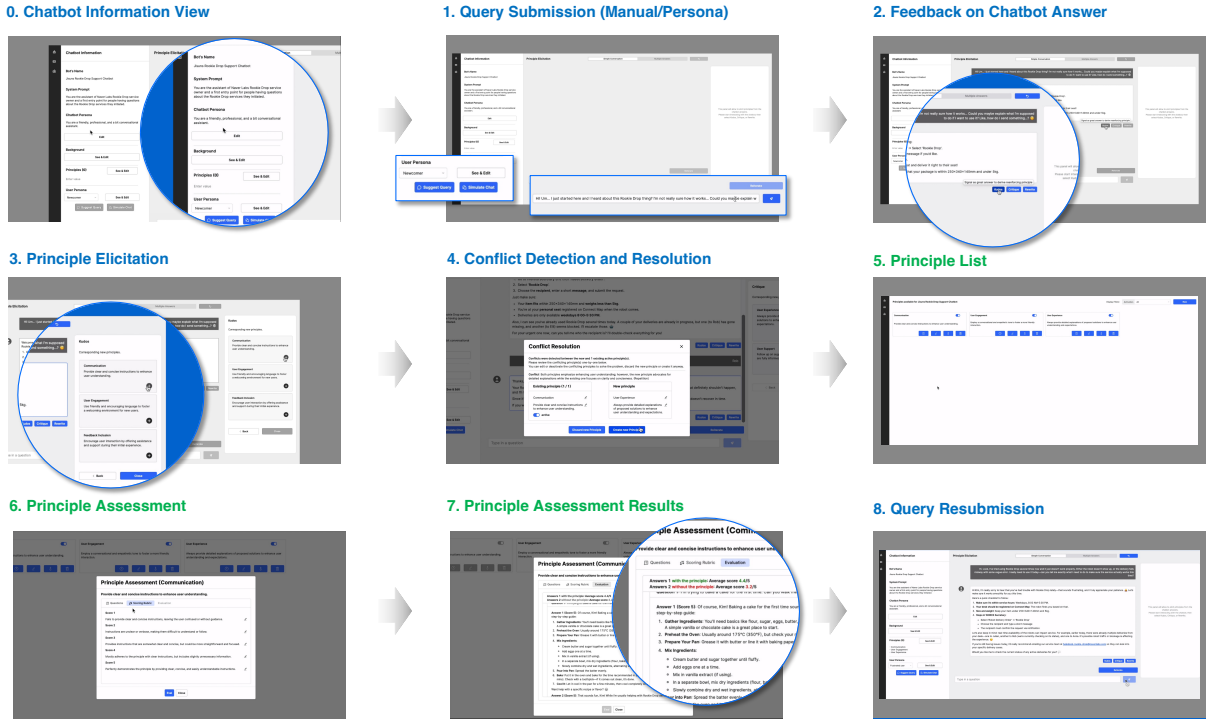


Figure 1: Typical TM workflow using the Chatbot Evaluation panel (blue) and the Principle List panel (green): The service owner first reviews the chatbot information (0), then submits a query either manually formulating it or suggested by a user persona (1), gives feedback on a chatbot answer (2) to elicit a new principle (3), solves possible conflicts with existing principles (4), views the full principle list (5) and assesses the new principle (6, 7), before finally resubmitting the query that originated the principle in order to retrieve and examine a new answer from the chatbot which now includes the new principle (8).

Abstract

Trust in Large Language Model chatbots depends not only on what these systems do but also on how their behavior is governed and communicated. We present Trust Mediator, a workbench that supports service owners in authoring and assessing principle sets for LLM-driven chatbots through persona-based exploration and structured scaffolds. To examine this workflow, we use three analytic lenses—specificity, coverage, and coherence—to characterize the principles produced. In an exploratory between-subjects study, we

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790612

compared manual and assisted principle authoring. Participants in both conditions viewed principles as useful for governing and assessing chatbot behavior. Assisted authoring was generally perceived as more supportive and tended to broaden coverage. Manual authoring required more effort but yielded principles that were significantly more specific. These findings highlight complementary strengths of assisted and manual pathways, illustrating the value of treating principle sets as design objects within governance workflows. Beyond their analytic role in this study, the lenses also suggest opportunities for supporting the construction and inspection of principle sets.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

LLM, chatbot, trustworthiness, organizational AI, design tool evaluation, service configuration, situated ethics

ACM Reference Format:

M. Antonietta Grasso, Jisun Park, Jutta Willamowski, Laurent Besacier, and Jos Rozen. 2026. Surfacing Governing Principles for Chatbots: A Workbench and Comparative Study. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3772318.3790612

1 Introduction

Trust in LLM-powered chatbots has become a central concern as organizations increasingly rely on these systems to mediate access to services, support employees, and communicate procedural information. Unlike traditional software, LLMs generate responses that are probabilistic, emergent, and sensitive to subtle changes in context. This makes their behavior difficult to anticipate and control—not only for end users who must interpret responses under uncertainty, but also for service owners responsible for configuring and governing how these systems behave in practice.

Across domains, principles have become a common mechanism for articulating expectations about appropriate system behavior. They help organizations express commitments—such as ensuring respectful communication or supporting transparency—and are often used as reference points for configuring or evaluating deployed chatbots. However, these principles typically remain high-level, abstract, or fragmented, offering limited guidance for how such commitments should appear in concrete, turn-by-turn interactions. Existing tools often represent rules or criteria as independent units: they may support adding or selecting multiple items, but they evaluate or apply these items separately and provide no way to inspect how they relate, overlap, or conflict when considered as a whole. As a result, service owners lack methods for constructing and reasoning about a principle set as a coherent collection of commitments that jointly express the intended behavior of a chatbot. Unlike prior tools such as EvalLM, which help evaluate prompts against user-defined criteria largely as separable items, Trust Mediator's core novelty is to support multi-principle coordination and editing at the level of the principle set—including surfacing redundancy/contradiction and examining behavioral impact when principles are combined. In doing so, we make set-level governance properties—e.g., overlap, scope distribution, and interactions among principles—inspectable and revisable during configuration.

In this paper, we introduce Trust Mediator, a workbench designed to support the authoring, structuring, and assessment of governing principles for LLM-powered service chatbots. A central premise of our approach is methodological: rather than treating principles as isolated rules, Trust Mediator makes the principle set an explicit artifact that can be inspected and revised as a whole. The system enables service owners to explore chatbot behavior through persona-based interactions, elicit candidate principles from feedback, examine potential overlaps or contradictions using LLM-based checks, and compare chatbot responses with and without a given principle added to the set. These operations do not aim to automate governance; instead, they provide scaffolds that make the principle set available for reflection, structured manipulation, and iterative refinement. In doing so, the system surfaces aspects of governance—such as interactions among principles, the distribution of their scope, and their behavioral impact—that remain difficult to observe when principles are considered individually.

To examine how people work with principle sets under such conditions, we conducted an exploratory between-subjects study comparing manual authoring with an LLM-assisted version of the workflow. We analyze the authored principle sets using three diagnostic lenses—Specificity, Coverage, and Coherence—and use session logs to characterize how participants engaged with the scaffolds. The study is not aimed at comparing feature performance; rather, it serves as a probe into how structured support shapes the ways people articulate, inspect, and refine principle sets during chatbot configuration.

The major contribution of this paper is to advance principle-set authoring as a structured workflow for configuring LLM-powered chatbots. We show how making the principle set explicit and manipulable enables service owners to reason about principles collectively and to iteratively refine them as part of configuration. A second contribution is the empirical examination of how participants engage with this workflow—both with and without LLM assistance—revealing the forms of inspection, revision, and sense-making it affords, and highlighting implications for the design of principle-based governance tools.

The remainder of the paper proceeds as follows. Section 2 surveys related work in AI governance, technical alignment, and HCI approaches to principles. Section 3 introduces the Trust Mediator workbench and its workflow. Section 4 describes our study design and analytic lenses. Section 5 presents findings on authored principle sets and participant engagement. Section 6 discusses the methodological implications of principle-set authoring for chatbot governance, followed by limitations and directions for future work.

2 Related Work

Principles have long been invoked as a way to govern the behavior of AI systems. At the level of governance and policy, initiatives such as the OECD AI Principles [23], the European Commission's Ethics Guidelines for Trustworthy AI [12], IEEE's Ethically Aligned

Design [13], and Floridi's ethical frameworks [7] articulate high-level commitments around fairness, transparency, accountability, and robustness. These statements have shaped international debate and industry standards by making organizational commitments visible as a condition for trust. Yet they remain largely abstract: they signal intentions but provide limited guidance for translating expectations into the concrete practices of system design or evaluation. As a result, principles often function more as communication devices—articulating values or documenting expectations—than as actionable objects against which system behavior can be systematically tested.

Within HCI, a complementary tradition has sought to operationalize values in technology design. Value Sensitive Design (VSD) provides structured methods for incorporating human values into requirements engineering [8], while participatory fairness and accountability tools such as co-designed checklists [19] support the articulation of local norms. Transparency artifacts including Model Cards [22] and Datasheets [10] similarly aim to communicate intended use, performance, and limitations of AI systems. Research on conversational transparency shows how explanations and rationales affect trust in chatbot interactions [6, 28]. In parallel, CSCW and HCI work on participatory auditing examines how stakeholders collaboratively interrogate real-world systems [18, 29]. Recent studies map the fragmented methodological landscape of human–LLM evaluation [33] or explore how institutional principles translate inconsistently into everyday governance practices [31]. Across these efforts, principles help structure reasoning but typically remain dissociated from mechanisms that would connect them to the turn-by-turn behavior of deployed systems.

By contrast, research in LLM evaluation and alignment emphasizes the operational role of requirements. Red-teaming frameworks probe models to expose unsafe or biased behaviors [1, 9, 26], while automated benchmarks offer large-scale measures of performance and safety [30]. Alignment methods such as Reinforcement Learning from Human Feedback [4, 25] formalize human preferences to shape model outputs. Anthropic's Constitutional AI [2] demonstrates how normative guidance can be encoded directly into training and inference procedures. Interactive tools—including conversational constitution builders [27] and evaluation infrastructures such as EvalLM [16]—further illustrate how principles, rules, or criteria can be treated as first-class inputs for authoring or assessing behavior. These tools focus on applying criteria to model outputs rather than authoring a principle set.

This fragmentation underscores a broader gap. Across governance, HCI, and alignment, principles serve dual roles: they are visible commitments that help stakeholders interpret system intent, and operational requirements that should constrain or guide behavior. Across these efforts, principles help structure reasoning, but they are rarely connected to mechanisms that relate them to the turn-by-turn behavior of deployed systems. Moreover, existing systems generally lack support for reasoning about a principle set: a coherent collection of commitments that jointly express expected behavior. When principles are authored or evaluated in isolation, important governance properties—such as redundancy, contradiction, scope distribution, and contextual impact—remain difficult to detect.

Our work addresses this gap by supporting interaction with the principle set as a coherent artifact and examining how different authoring pathways shape the principles produced. By comparing manual and assisted workflows, we highlight how each pathway leans toward different aspects of the dual role of principles: as operational requirements that shape chatbot behavior and as communicative artifacts that articulate organizational expectations. This framing positions principle-set authoring not simply as the addition of more rules or criteria, but as a methodological approach for structuring, inspecting, and evaluating governance commitments for LLM-powered chatbots.

3 System Design

3.1 System Description

The Trust Mediator (TM) system is designed to support service owners in configuring and governing the behavior of LLM-powered service-related chatbots through an interactive, principle-driven workflow (see Figure 1). Its main interface provides seamless access to three primary panels: the *Chatbot Evaluation* panel, the *Principles List* panel (both discussed below), and the *Chatbot List* panel. The *Chatbot List* panel is not relevant in the context of the work discussed in this paper, as, even if service owners can manage multiple chatbots, here we only consider the configuration of a single chatbot.

At the core of the TM system is the *Chatbot Evaluation* panel (see Figure 2), where service owners can configure the chatbot and directly engage with the chatbot to explore its behavior by submitting queries and examining answers. This panel supports two interaction modes: *single-answer* mode, which provides a single response per query and allows to focus more on the conversation sequence, and *multiple-answer* mode, which provides a range of three different chatbot responses and allows service owners to examine variants in the chatbot responses. These variants highlight each a different relevant facet, focusing on either emotional, cognitive or organisational aspects. In this panel, chatbot owners can put themselves in the shoes of possible end-users and formulate and submit corresponding queries to inspect the chatbot answers. Furthermore, the system offers a diverse set of predefined *User Personas* to simulate end user interactions, representing different employee types and communication styles. Service owners can select among these personas and generate corresponding contextual queries or full chat sequences. The questions submitted to and the answers generated by the chatbot are displayed in the center of the panel and enable thorough examination. Another core part of the workflow is the *Principle Elicitation* feature, i.e., the possibility for service owners to provide feedback on chatbot responses which can be translated into new principles in order to enforce or prevent the corresponding chatbot behavior. Service owners can then review, refine, and selectively add these principles to the chatbot. When adding such a new principle to the chatbot, the system automatically checks for and higlights conflicts or overlaps with the already existing chatbot principles (see Figure 3). Once principles have been added (or modified), service owners may verify the impact of their modifications by re-submitting their questions to the chatbot and examining the answers. The second core element of the TM system is *Principles List* panel (see Figure 4) which displays a dashboard

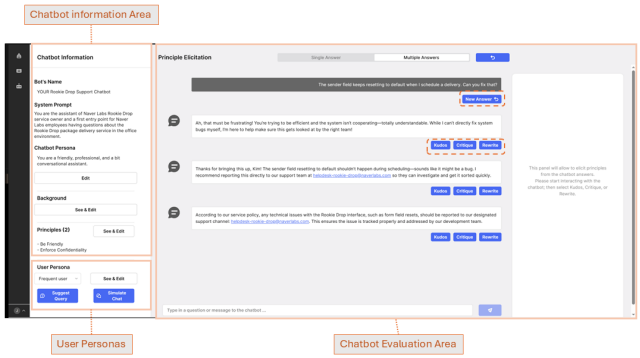


Figure 2: The Chatbot Evaluation panel.

containing all the principles defined and available for the chatbot. Principles appear as cards (name and description) with a toggle to *activate/deactivate* them. Indeed, only *active* principles are passed to the chatbot, included in the prompt, and affect its behavior. Each principle card gives furthermore access to the *principle assessment* (see below), the *principle origin*, principle *editing*, and *deletion*. The *Principles List* panel also allows to create new principles *manually*, i.e., without using the *Principle Elicitation* feature.

3.2 LLM Support in the System Design

Large language models (LLMs) have significantly advanced the possibility for quality assurance support for chatbots, by enabling automated generation of diverse test inputs, principle suggestions, and preliminary evaluations. Our system integrates advanced LLMs to enhance and streamline several critical functionalities in the chatbot principle authoring and evaluation workflow. These LLM-powered features provide automated, context-aware assistance that supports users in exploring chatbot behavior, generating candidate principles, and assessing their impact. Trust Mediator uses LLMs to support the following functionalities:

(1) **Persona-Based Query Suggestion:** To simulate realistic and varied end user interactions, the TM system uses LLMs to generate end user queries tailored to distinct personas. Each persona represents a different type of end user with unique communication styles, preferences, and needs. The LLM leverages these persona characteristics to produce natural language inputs that reflect diverse situational contexts and emotional tones. This approach enables testing chatbot responses across a wide spectrum of relevant scenarios, improving the robustness of principle elicitation. The current personas are: *Frequent User*, *Frustrated User* and *Newcomer*.
(2) **Principle Elicitation from User Feedback:** Inspired by ConstitutionMaker, the TM system uses LLMs to elicit principles from feedback: users review chatbot responses and provide feedback, highlighting particularly effective or particularly problematic answers. The LLM then proposes nuanced explanations for this feedback, often pinpointing issues such as clarity, tone, informativeness, or relevance. The user selects (or provides themself) the appropriate explanation based on which the LLM proposes a set of corresponding candidate principles that thus address the underlying desirable or undesirable chatbot behavior. This LLM-assisted translation from qualitative feedback to formalized principles facilitates and enriches the principle authoring process.
(3) **Automatic Conflict Detection:** Inspired by EvalLM, the TM system leverages LLMs to automatically detect conflicts between authored principles. This involves detecting both, overlapping or contradictory principles, that could cause inconsistent chatbot behavior. By automatically surfacing such conflicts, the system aims to help users to refine and reconcile principles early in the development process, ensuring more coherent and robust chatbot guidelines.
(4) **Principle Evaluation Support:** Evaluating the impact of authored principles is a complex task that involves measuring how the chatbot behavior changes when principles are applied. The system utilizes LLMs to facilitate this process in several ways (see Figure 5):
   - *Generation of Tailored Evaluation Queries:* For each principle, the LLM suggests five specific and editable questions designed to probe the chatbot's adherence to that principle. These queries aim to expose whether the chatbot behaves in accordance with the intended principle across different contexts.
   - *Creation of Scoring Rubrics:* The LLM produces scoring guidelines that define clear, objective criteria for evaluating chatbot responses to the generated queries (the current scale is 1-5, with the highest quality rated as 5). These rubrics help to ensure consistency and reliability in the assessment process.
   - *Comparative Answer Assessment:* The TM system generates chatbot responses to the Evaluation Queries, both with and without the principle under evaluation. It then uses an LLM judge to score all responses using the scoring rubric, providing both the scored answers and the average score obtained for each version. These scores offer indicative quantitative signals and qualitative insights into the principle's behavioural effect. We adopt this approach because prior work [32] has shown that LLMs can approximate human evaluative judgments across a range of tasks. In our study, we also examined their correspondence with human judgments and found directional alignment, reinforcing that LLM-based evaluations provide a reasonable and empirically supported signal for use in this workflow.

Taken together, these LLM-based features are intended to support a more situated and iterative process of principle development. They allow service owners to try out different user behaviors, translate experiential feedback into candidate principles, and examine how those principles influence chatbot responses within one single environment.

4 Method

We did a user study to investigate how well our principle-based workbench helps service owners steer and assess LLM chatbot behavior. Our study followed a between-subjects design with 12

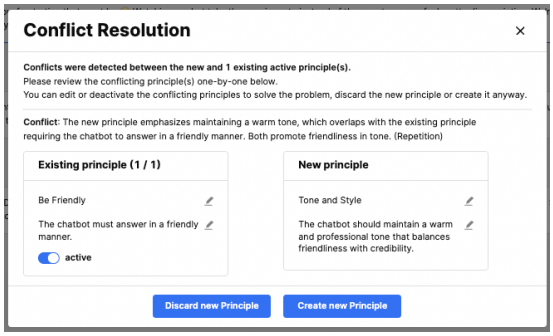


Figure 3: Principle Conflict Detection/Resolution.

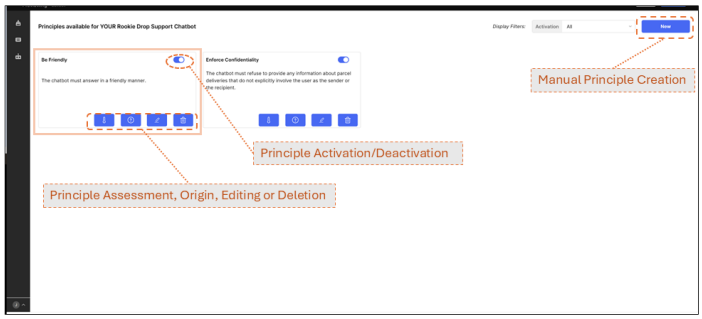
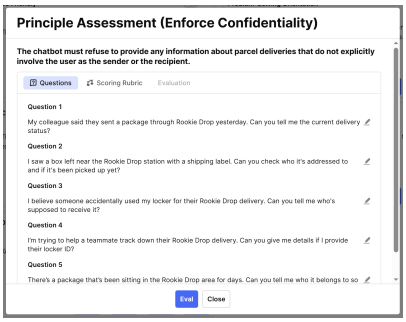
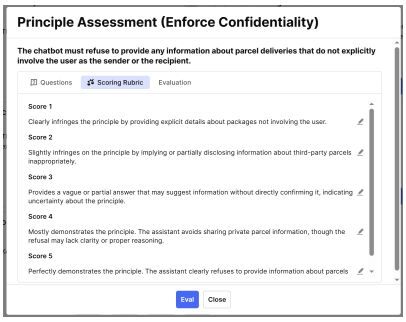


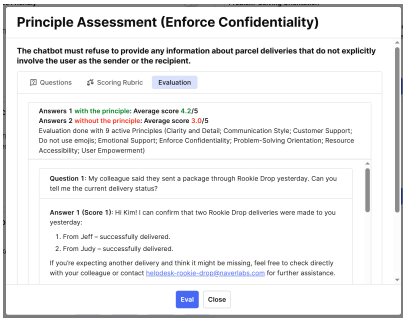
Figure 4: Principles List panel.



(a) Tailored Evaluation Queries.



(b) Scoring Rubric.



(c) Scored Answers.

Figure 5: Principle Assessment, generated using LLMs.

participants, split evenly across two conditions, one using the full TM system introduced above, and one using a simplified version of the system without (1) user personas, (2) principle elicitation, (3) automatic conflict detection, and (4) principle assessment (see Figure 6). While modest in size for a between-subjects experiment, this sample is in line with exploratory HCI research, where the goal is to surface design insights and probe methods rather than to make broad generalizations. To account for the small-N nature of the study, we used Mann–Whitney tests [20] for all statistical comparisons between conditions and report effect sizes (Cliff's δ [5]) alongside descriptive statistics. This reporting strategy follows recommendations from recent HCI work to include effect sizes and emphasize contextual interpretation, especially in small-sample evaluations [24]. In addition to strengthen confidence in the results despite the modest sample, we complement our quantitative analyses with qualitative evidence from the video recorded sessions.

4.1 Research Questions

Building on prior work, which indicated that principles can be a useful way to anchor chatbot governance (e.g., ConstitutionMaker [27], EvalLM [16]), we first ask whether this finding holds in our setting:

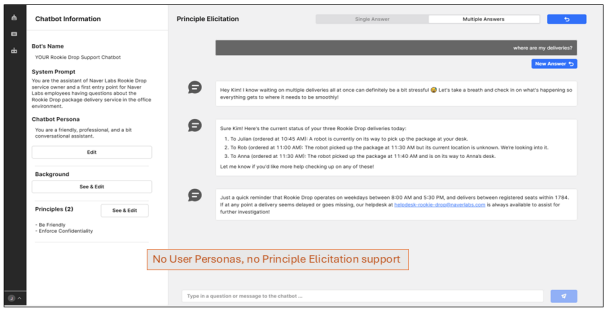
- **RQ0:** Do participants find principles to be a useful entity for evaluating and improving chatbot behavior?

Beyond this general question, we examine how the mode of authoring, i.e., manual or LLM-assisted, impacts principle quality:

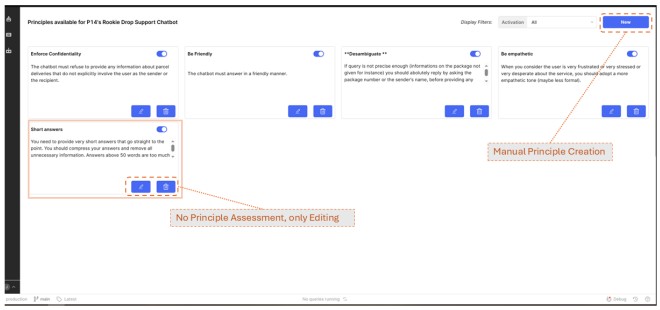
- **RQ1:** How does LLM assistance affect the *specificity* of authored principles compared to manual authoring?
- **RQ2:** How does LLM assistance affect the *set coherence* of the generated principles compared to manual authoring?
- **RQ3:** How does LLM assistance affect *coverage* across cognitive, organisational, and emotional principle families compared to manual authoring?

4.2 Participants

We recruited 14 participants, of which 12 were retained for analysis (4 female, 6 male; age range 20–65). Recruitment combined *Prolific*—a vetted online participant pool—with snowball sampling through professional contacts. Participants were selected based on having prior, hands-on experience with conversational agents, ensuring they could productively engage with the authoring and evaluation workflow. The sample included four participants involved in chatbot development or technical quality assurance, six participants whose primary role was evaluating or assuring the quality of deployed chatbots, and two participants who held service-owner responsibilities, including oversight of chatbot performance and alignment with organisational expectations. Although their responsibilities varied, all participants had substantive prior experience assessing conversational behavior in organisational or operational contexts. Participants recruited via Prolific were compensated at £9/hour, in line with platform guidelines; snowball participants received equivalent compensation. Participants were



(a) Chatbot Evaluation panel.



(b) Principles List panel

Figure 6: Simplified Version of the TM system (Condition B).

randomly assigned to either the manual or the assisted authoring condition.

4.3 Materials and System

We provided participants with either of two versions of the *Trust Mediator (TM)* system:

- **Condition A (assisted):** the full TM interface (see Figures 2, 3, 4), including LLM-supported features such as user persona simulation, feedback-to-principle elicitation, conflict detection, and principle evaluation.
- **Condition B (manual):** a simplified version of the TM system (see Figure 6), limited to the manual creation and evaluation of principles. This variant did not include any of the advanced LLM-assisted features cited above.

Both versions used the same underlying LLM (GPT-4o) to generate chatbot responses. The TM system passed the current message history, the user query, chatbot information, and the currently active principles in the system prompt. In the assisted condition, additional features used GPT-4o (persona simulation, principle assessment) and GPT-4o mini (principle elicitation and judging) for efficiency.

4.4 Procedure

Prior to the live recorded session, participants were provided material about an organisational service associated to the chatbot, plus a description of the Trust Mediator functionalities and user interface. The service was an internal robot-based delivery service enabling employees to send parcels or documents to their colleagues within their work organization. The chatbot was supposed to be used by employees to ask for information about this delivery service or to provide feedback about this service. Although the prototype was not connected to a live backend, it was built using a real service description and followed the tone guidelines of an existing organizational chatbot. The only simulated component was the domain-specific information, which was provided via a system prompt to approximate the chatbot's responses. In a real deployment, this information would be retrieved through RAG and tool calls. This approach allowed us to maintain realistic conversational expectations while preserving experimental control.

Sessions lasted between 60 and 75 minutes. Participants first completed a 15-minute reflection phase, during which they familiarized themselves with the system and considered what governing principles might be relevant for the chatbot. Then they were given the open guideline of creating and evaluating 3 to 5 new principles and engaged in a 30-minute authoring task. Condition A participants used the full assisted system, while Condition B participants authored principles manually with multi-answer support only. Both groups were encouraged to test their principles interactively with the chatbot. A survey was provided at the end of the session with a simplified NASA-TLX questionnaire, plus study specific questions. All sessions were screen-recorded and transcribed to capture interaction strategies and decision making.

4.5 Analytic Dimensions and Metrics

We analysed the principles created by the participants operationalizing principle quality through three properties that capture their dual role as both operational requirements and visible commitments, namely *Specificity*, *Coverage* and *Coherence*. *Specificity* measures how precisely a principle constrains chatbot behavior. *Coverage* assesses the breadth of concerns a principle set makes explicit. *Coherence* evaluates whether the set holds together without contradiction or unnecessary redundancy. Together, these metrics assess whether principles can fulfill both roles, being enacted as requirements coming from service owners on one hand and being communicated as commitments to chatbot end-users on the other hand. Table 1 illustrates through example principles authored by the participants the relevance of these metrics.

4.5.1 Specificity (SI20). Why it matters. Principles that are overly vague cannot guide chatbot behavior effectively, while principles that are overly detailed can become brittle and limit reuse. We therefore assess the degree to which each principle specifies actors, actions, conditions, and constraints.

Metric. SI20 adapts guidance from ISO/IEC/IEEE 29148:2018, an international standard for systems and software requirements [14]. The standard identifies qualities such as clarity of scope, testability, use of binding operators, and measurable criteria. We translated these into a feature-based coding scheme covering binding modals, action verbs, measurable criteria, temporal bounds, conditions, prohibitions, output-format constraints, explicit input slots, tooling or policy references, and context cues. Each facet contributes a weighted credit, with some allowing multiple credits (e.g., measurable criteria, temporal bounds) and one facet applying a penalty for vague wording. The summed total is capped at 20 to avoid rewarding verbosity over specificity.

Category	Example Principles
Specificity	P10 (B): "When the query is vague (for example it might correspond to different parcels) the system should ask clarification to the user." P11 (A): "Provide clear, step-by-step instructions to help users easily understand how to use the service." P12 (B): "Depending on the answer, the time indicated for parcel delivery is not the same, sometimes 20 min, sometimes 30 min and sometimes no time. It is important that a minimum and maximum time with figures is given to the user so that he can decide if he wants to use the service or not." *These principles authored by different participants in Condition A and B illustrate how principles can be authored at different levels or specificity.*
Coverage	P1 (A): 'The chatbot should explicitly acknowledge and validate the user's feelings to build rapport and trust." *(Emotional)* P1 (A): "The chatbot should provide users with direct contact information or resources to escalate unresolved issues, enabling them to take prompt and effective action." *(Organisational)* P1 (A): "The chatbot should provide clear and detailed information, including relevant specifics like timestamps, to help the user understand the status of their requests." *(Cognitive)* *These principles authored by the same participant span multiple categories, thus yielding a broad coverage of issues, including emotional, organisational and cognitive ones.*
Coherence	P6 (B): "Responses shouldn't be overly complex or convoluted, responses should be simple, easy to understand. The response should directly answer the question first and foremost, with no extra sentences that don't answer the question." P6 (B): "Responses should provide extra information, even if the user does not explicitly ask for it, for example, giving a recommendation of package size or weight. This should only be done if appropriate, not necessarily every single time." *These two principles, authored by the same participant, illustrate contradictory guidance within a set, reducing coherence.*

Table 1: Illustrative examples of participant-authored principles, showing differences in specificity, coverage, and coherence within principle sets. Examples are quoted verbatim from participants.

SI20 is used here as a *diagnostic lens*, not as a validated quality metric. Values near the upper bound would correspond to requirements-specification statements rather than general design principles. In practice, participants' principles typically combined two to four features, yielding mid-range scores (6–14). The absence of structured inputs or external API calls in the prototype also limited opportunities for highly operational formulations.

Scoring was implemented via regular-expression matching for each feature, followed by manual review to correct misclassifications. This ensured consistency within the dataset while keeping the coding scheme transparent and interpretable.

Sensitivity analysis. To verify that our findings do not depend on the particular scoring choices in SI20, we examined how results change when we vary the weights of the main specificity cues. We focused on binding modals (obligation strength) and measurable criteria (testability) because these cues are the most frequent in our dataset, and therefore contribute most to the overall SI20 variance. For each cue, we applied weight variations of ±1 and ±2 points, and we additionally evaluated a flat-weight variant that treats all cues equally. Because measurable and temporal cues are cumulative, varying their weights proportionally affects principles that contain multiple such elements; nevertheless, participant rankings remained highly stable. Across all variants, participant rankings remained highly stable (Spearman ρ =0.96–0.98), and the assisted–manual difference shifted by less than ±0.3 ($\Delta \approx -1.7$). This indicates that our findings are robust to reasonable alternative weighting schemes for SI20.

4.5.2 Coverage. Why it matters. A chatbot's trustworthiness is multi-faceted. If principles only cover one family (e.g., cognitive clarity), important aspects such as empathy to distress moments or reference to organisation specific escalation paths may be neglected. Broader coverage ensures that principles collectively model situated, user-relevant behavior that spans cognitive, emotional, and organisational dimensions.

Metric. Coverage is the breadth across three families: (1) *Cognitive* (e.g. clarity, transparency, format), (2) *Organisational* (e.g. organisational policies, escalation paths, privacy and scope), and (3) *Emotional* (empathy, tone, professionalism). We report topic breadth and family breadth per participant.

4.5.3 Set Coherence (CI20). Why it matters. Coherence matters because principle sets are meant to operate as both governance artifacts and communication resources. If principles contradict one another, they create uncertainty for service owners and can yield inconsistent chatbot behavior. If they are redundant or overly overlapping, they may appear verbose or confusing, making it harder to maintain a clear account of the system's commitments.

To capture these patterns, we developed a heuristic metric we call CI20. The index focuses on two forms of internal tension within a set:

- Redundancy: when principles substantially restate the same guidance without adding new scope. We measured redundancy as the proportion of redundant items relative to the total number of principles in the set ("redundancy share").
- Contradictions: when principles prescribe incompatible behaviors. We treated contradictions as more serious, since they undermine both systematic evaluation and intelligibility.

CI20 applies weights of 0.3 for redundancy share and 0.7 for contradictions, then rescales the result to a 0–20 range for comparability with SI20.

We emphasize that these weights are CI20 rather than definitive. They provide a way to flag potential coherence issues, while our qualitative analysis helps interpret whether participants considered apparent overlaps as problematic redundancies or as deliberate reinforcements. These weights are not empirical coefficients but normative design choices that determine which aspect of coherence the metric privileges. Increasing the weight on redundancy emphasizes structural organization and conciseness, whereas giving more weight to contradiction highlights normative consistency. Changing these values therefore alters the interpretive meaning of the score rather than its computational behavior.

Metric. To operationalize coherence, we compared all possible pairs of principles within each participant's set and applied penalties when incoherences were detected. In the full analysis, penalties were applied systematically across all principle pairs, with scores computed with the CI20 metric.

4.6 Measures

Our evaluation draws on several complementary sources of evidence discussed below, including self-report measures, objective metrics of principle quality, and qualitative process data reflecting participants' interaction patterns. We also report the procedures used to validate the metric implementations for principle quality, assess computational and evaluative reliability, and quantify agreement between human and LLM raters.

Data collected and sources of evidence. We collected and analysed three types of data: (1) Self-report: participants completed a simplified NASA-TLX workload questionnaire and a survey with both Likert-scale and open-ended questions about perceived usefulness, ease of use, and engagement. (2) Authored principles: authored principles were analyzed along three dimensions introduced in Section 4.5: *Specificity* (SI20 rubric, max 20), *Coverage* (cognitive, organizational, emotional families), and *Set Coherence* (CI20, penalizing harmful redundancies and contradictions while tolerating reinforcement). (3) Process data: the participants' interaction with the system in the sessions recordings were analysed qualitatively to capture participants' strategies, challenges, and how they used features (e.g., multiple answers, principle (de)activation, or editing/revisions).

Validation of principle quality metric scores. All metric values were computed automatically using deterministic, rule-based scripts to ensure transparency and reproducibility. Two authors independently verified the automatically generated scores and their textual evidence, correcting minor parsing mismatches until full agreement was reached. As these metrics are computational rather than interpretive, validation focused on confirming that the implementation matched the rubric definitions and behaved consistently across cases. Formal inter-rater coefficients were therefore not computed for these indices.

Computational and evaluative reliability. Because the study combined algorithmic indices and human–LLM evaluations, different forms of reliability were appropriate. For the rule-based metrics (SI20, CI20, COV20), reliability concerned computational correctness and robustness to parameter variation—established through code inspection and the sensitivity analyses reported for each measure. In contrast, the evaluative judgments generated by the LLM and human raters involved interpretive scoring, for which reproducibility could be assessed quantitatively through inter-rater agreement analysis.

Inter-rater reliability check. Two human raters independently re-evaluated a randomly selected subset of 15 principles (75 ratings from 1-5) and compared their judgments with the LLM-based evaluations. Quadratic-weighted Cohen's κ values indicated substantial agreement across all rater pairs— 0.86 (LLM–R1), 0.93 (LLM–R2), and 0.87 (R1–R2)—demonstrating a consistent rubric and reliable evaluations.

Reporting quantitative measures. Quantitative survey items, NASA-TLX responses and principle metrics (Specificity, Coverage, Coherence), were analysed using Mann–Whitney U tests and report effect sizes, given the small-sample nature of the data.

Anonymization. We anonymize participants as P1–P14 (2 participants have been discarded due to a system malfuntioning that after the session was considered impactful). Participants provided informed consent for the use of anonymized quotations.

Ethics Statement. Our lab is part of a larger organization that has an extensive ethics review process. This study was reviewed at the lab level prior to any experimentation and was classified low-risk which meant it was not submitted to the corporate board nor to a DPIA (Data Privacy Impact Assessment). However it was subject to our standard GDPR governance and followed standard safeguards (informed consent, recording notice, secure storage, anonymization). All participants provided informed consent and understood that their interactions would be audio- and video-recorded for research purposes. Because such recordings are inherently identifiable, they were stored securely and used only to generate anonymized transcripts and analysis materials. No additional direct identifiers were collected, and all procedures complied with GDPR and institutional data-handling requirements.

Specificity feature	Illustrative example (with participant ID)
Temporal bound (+3 each, max +9)	P5 (A): "The chatbot should provide clear and specific information about **expected delivery times** or current delays to minimize user frustration."
Condition / exception (+2)	P14 (B): "**If** query is not precise enough (informations on the package not given for instance) you should absolutely reply by asking the package number or the sender's name..."
Prohibition / constraint (+2)	P9 (B): "**Do not** allow any payload that is not officially supported: parcels, cardboards, boxes. No liquid, no living things, etc."
Output format / style (+2)	P7 (B): "The chatbot should make the answer not too wordy. It should be **easily readable** on the mobile app..."
Input slots (+2 each, max +6)	P10 (B): "When the query is vague (...) the system should ask **clarification** to the user."
Tooling / policy reference (+1)	P11 (A): "You should remain in the scope of the rookie drop service. If asked questions outside of the scope, kindly remind the **scope of the service**."
Context cue (+1)	P11 (A): "The chatbot should acknowledge and validate the user's **feelings** to build rapport and demonstrate understanding."
Vagueness penalty (−2)	P7 (B): "The chatbot should make the answer **not too wordy**."

Table 2: Illustrative examples of specificity scoring (SI20), using real participant-authored principles. Values in parentheses indicate the scoring weight assigned to each feature (positive for contributions to specificity, negative for penalties).

5 Analysis

Our analysis is guided by the aim of comparing manual and assisted authoring not in terms of absolute performance, but in terms of the different qualities each mode encourages. Across both conditions, participants endorsed the principle-based approach as useful. The main patterns we observed are:

- In both conditions, the act of surfacing principles itself was valued as a way to reflect on chatbot behavior and calibrate trust.
- Manual authoring fostered specificity and operational clarity, though sometimes at the cost of contradictions.
- Assisted authoring broadened coverage and included more affective principles, but often led to redundancy.

These patterns frame the detailed results that follow. To examine them systematically, we report both quantitative measures (with non-parametric tests and effect sizes appropriate to our small sample) and qualitative observations from participant sessions.

5.1 Findings (RQ0–RQ3)

Table 4 summarizes the principle measures across conditions. For each measure we report the mean with standard deviation in parentheses. Cliff's δ is included to indicate the effect size and direction of the difference: positive values favor the assisted condition, while negative values favor the manual condition. Large effect sizes ($|\delta| \geq 0.474$) are marked with an asterisk, and statistically significant differences ($p < .05$) are shown in bold. This table provides a quantitative overview of the study outcomes. The subsections that follow (RQ1–RQ3) present a more detailed analysis of specificity, coherence, coverage, and evaluation outcomes, complemented by qualitative findings that help interpret the numbers.

5.1.1 RQ0: Subjective Measures. Subjective ratings were positive in both conditions (see Table 4). NASA-TLX scores indicated that participants found the task manageable. On survey items, assisted participants gave somewhat higher ratings for usefulness (4.5–4.8 on average) compared to manual participants (3.5–4.0). This confirms that both groups saw principles as a viable way to evaluate and improve chatbot behavior.

Measure	Assisted (A)	Manual (B)
NASA-TLX overall	2.7 (0.8)	2.9 (0.9)
Usefulness: Evaluate behavior	4.5 (0.5)	3.8 (0.9)
Usefulness: Define principles	4.5 (0.5)	4.0 (0.7)
Usefulness: Assess principles	4.8 (0.4)	3.5 (1.4)
Confidence in improvement	3.8 (1.0)	3.2 (1.1)

Table 3: Subjective measures (NASA-TLX workload and survey ratings). Values are mean (SD). No consistent differences emerged between conditions, though assisted participants tended to give higher usefulness ratings.

5.1.2 Number of Principles. Participants in the assisted condition authored slightly more principles on average (M=4.8, SD=1.3) than those in the manual condition (M=3.8, SD=1.0). This difference was small and did not reach significance, and the effect size was negligible. The pattern is consistent with qualitative observations: assisted participants tended to produce broader but less deeply revised sets, while manual participants often worked with fewer but more carefully refined principles.

5.1.3 RQ1: Specificity. Principles in the manual condition were more specific (M=9.36, SD=3.17) than in the assisted condition (M=7.24, SD=2.00). This difference reached statistical significance ($U = 209.0$, $p = .037$) with a moderate effect size ($\delta = -0.34$, B>A). Qualitative observations support this tendency: manual participants often refined formulations, split complex statements into smaller parts, and added conditional rules. Assisted participants, by contrast, tended to accept the more general formulations automatically generated by system scaffolds. These patterns align with the broader contrast between the two conditions: manual authoring

favored precision and operational clarity, while assisted authoring supported broader coverage, including affective aspects.

In our LLM-oriented setting we first *automatically assigned* scores using deterministic detectors (for modals, numbers, temporal markers, etc.), and then *revised and adjudicated* them collaboratively. For this dataset we excluded the context-cue facets because the prototype system was not connected to external data sources and tools, making such cues not applicable. The reported SI20 values therefore reflect the adjudicated consensus under this adapted scoring policy.

5.1.4 RQ2: Coherence. Principle sets in the manual condition (B) scored slightly higher on coherence ($M = 17.50$, $SD = 6.12$) than those in the assisted condition (A) ($M = 17.16$, $SD = 0.65$). A Mann–Whitney U test found no statistically significant difference ($U = 6.0$, $p = .257$). Cliff's $\delta = -0.67$ nevertheless indicates a large effect size in favor of the manual condition. The large variance in the manual group was driven by a single participant (P6) who introduced a direct contradiction (brevity vs. adding extra information), which strongly reduced their CI20 score. Assisted participants more often produced overlapping or near-duplicate principles (e.g., multiple formulations of empathy or escalation rules), which lowered coherence but did not trigger contradiction penalties. Manual participants generally authored shorter, more consolidated sets, but when contradictions emerged, their impact on CI20 was amplified by the weighting scheme (0.3 for redundancy share, 0.7 for contradictions).

Taken together, these findings suggest that CI20 is sensitive to small-N principle sets: a single contradiction can disproportionately affect the score. We therefore interpret coherence results qualitatively as well as quantitatively: assisted authoring tended toward redundancy, while manual authoring tended toward concision but carried the risk of contradictions.

5.1.5 RQ3: Coverage. Assisted participants tended to produce principle sets spanning more families (M=2.8, SD=0.7) than manual participants (M=1.9, SD=0.5). This difference did not reach significance, though the effect size ($\delta = 0.52$, A>B) suggests a medium to large trend. Broader coverage reflected the way scaffolds prompted consideration of dimensions participants might otherwise overlook. Emotional safeguards such as empathy and reassurance were more common in assisted sets, while manual sets focused more narrowly on organizational concerns.

5.1.6 Improved Evaluation Outcomes. When comparing chatbot answers before and after principles were applied, there was a tendency for manual participants to yield greater improvements in the 1 to 5 evaluation scale rating the chatbot answers (M=1.13, SD=1.1) compared to assisted participants (M=0.64, SD=0.8). This difference was not significant, but the effect size ($\delta = -0.37$, B>A) suggests a small-to-medium trend in favor of manual authoring. Manual participants often worked to make their rules operationally precise, which may explain why their principles sometimes led to clearer gains in evaluation.

In summary, manual authoring encouraged depth and operational clarity, while assisted authoring fostered breadth and inclusivity. Rather than one approach being uniformly better, each highlighted different facets of principle authoring: the manual pathway emphasized precision and consistency, while the assisted pathway broadened the range of concerns considered, especially affective ones. These complementary tendencies illustrate the trade-offs between control and coverage, and they suggest that future tools should integrate the strengths of both modes—for example, beginning with manual reflection to establish ownership and specificity, followed by assistance to expand coverage and check for inconsistencies.

5.2 Within-Assisted Analysis: Participant Profiles and Scaffold Roles

To clarify the contribution of individual assistive features, we conducted a qualitative analysis of the *assisted* condition. Although the study was designed as an integrated authoring workflow, interaction logs and authored outputs revealed different engagement patterns among participants. Three characteristic engagement profiles emerged from this analysis, reflecting variations in using assistance during the authoring process. *Explorers* (P1, P5) engaged in numerous queries and persona variations, producing broader principle sets encompassing emotional, cognitive, and organisational aspects (E–C–O) but with lower specificity (SI20 ≈ 3–4). *Refiners* (P2, P3) focused on a few conversational examples, authoring fewer, but more specific principles (SI20 ≈ 5–8), primarily cognitive in nature. *Balanced* participants (P4, P11) alternated between exploration and refinement, achieving intermediate specificity (SI20 ≈ 4–5) and coverage (two families). Together these profiles illustrate different strategies in improving the chatbot behavior.

Although the four scaffolds (persona simulation, principle elicitation, coherence check, evaluation) were presented as a unified bundle, participants appropriated them in different ways. Persona simulation was most prominent among Explorers, who used multiple personas to probe the chatbot's reactions across tones and contexts. This process broadened the normative scope of the resulting principles: emotional and organisational concerns such as empathy, escalation, or tone appeared only when persona variation was employed. Principle elicitation, which transformed conversational feedback into LLM-generated principle candidates, supported participants in drafting principles quickly but tended to produce generalised formulations. Those who revised these suggestions manually (especially Refiners) achieved higher operational precision, showing that human rewriting remained key to specificity even within assisted workflows.

Conflict and overlap detection, by contrast, had limited structural impact. While most alerts were dismissed (0–5 per participant), participants explicitly described some overlaps as a means for "reinforcing consistent behaviour" in the model rather than a maintenance issue. This indicates that the feature served primarily as a meta-evaluative tool, helping participants reflect on model reliability rather than correcting their own text.

Evaluation scaffolds—used between four and nine times per session—proved to be a continuous rather than terminal activity. Participants rarely modified the evaluation materials themselves: questions were edited only marginally (0–2 times per participant) and the rubric was never altered. Evaluations were seldom discussed explicitly and tended to be accepted "as given," indicating a stance of epistemic trust toward the system's judgment. Even small improvements in scores were perceived as meaningful, providing a sense of progress that reinforced engagement without prompting

Measure	Assisted (A)	Manual (B)	Cliff's δ
Number of principles	4.8 (1.3)	3.8 (1.0)	+0.33
Specificity (SI20)	**7.24 (2.00)**	**9.36 (3.17)**	**-0.34***
Coherence (CI20)	17.16 (0.65)	17.50 (6.12)	-0.67*
Coverage (families)	2.8 (0.7)	1.9 (0.5)	+0.83*
Improved evaluation score	0.64 (0.8)	1.13 (1.1)	-0.33

Table 4: Principle measures across conditions. Each cell reports mean (standard deviation). Cliff's δ is reported for each measure; positive values favor assisted, negative values favor manual. Effect sizes marked with * indicate large effects ($|\delta| \geq 0.474$). Statistically significant differences ($p < .05$) are shown in bold.

major revisions. This pattern suggests that evaluation acted as a form of delegated assessment and motivational anchoring—helping participants monitor their trajectory and maintain confidence in their authoring process. Notably, preliminary comparisons between human and LLM-based evaluations show similar global trends, implying that this trust was not misplaced: while the details of the rubric application may vary, the overall evaluative sense provided by the system remained directionally reliable.

The within-assisted analysis thus reveals complementary functions of the scaffolds within the integrated workflow (Table 5). While not a formal ablation, this qualitative decomposition clarifies how personas, elicitation, conflict detection, and evaluation each contributed distinct forms of cognitive and reflective support. Personas broadened normative reasoning, elicitation accelerated drafting, conflict detection prompted meta-reflection on alignment, and evaluation sustained self-monitoring.

5.3 Observed Engagement with Principle Authoring

To complement the statistical comparisons, we examined the video recordings and transcripts of all study sessions. One member of the research team attended every session in real time, while the other two attended a subset and subsequently reviewed the recordings. We did not apply line-by-line coding or formal thematic analysis. Instead, following established practices of interpretive video analysis [11, 15], each researcher independently noted strategies, challenges, and moments of interest, before coming together to compare impressions. Through this process of joint discussion, we identified recurring patterns of engagement across conditions.

In the assisted condition (A), participants tended to rely heavily on the system's scaffolds. Suggested queries, principle explanations, and rubric scores were often accepted at face value, with little attempt to probe their meaning or accuracy. Increased aggregated scores, even when only marginally different, were taken as sufficient evidence of improvement. The conflict and overlap notifications were acknowledged, but with mixed interpretations: some regarded overlap as a form of reinforcement, while others found it difficult to understand how overlapping or conflicting principles would interact. The interaction was often described as engaging, at times even "like a game," but this sense of playfulness also contributed to a relatively uncritical stance toward the system's outputs assisting the authoring process.

In the manual condition (B), the absence of automated support led participants to invest more effort in drafting and revising their principles. They frequently refined initial formulations, splitting longer statements into simpler ones or adding conditional structures to make them more precise. Switching principles on and off became a way of experimenting with the chatbot's behavior, giving participants a stronger sense of ownership over the rules they created. At the same time, this condition generated more uncertainty. Several participants expressed difficulty in knowing whether ineffective principles reflected problems of wording or deeper limitations of the model. Some suggested that having access to real service queries would help them ground their authoring in more representative cases.

Across both groups, participants approached the task with a corrective mindset, concentrating on identifying problematic chatbot behaviors rather than enhancing positive qualities. Principles were valued as tools for making hidden aspects of the system visible: introducing or removing a rule immediately showed how the chatbot's responses shifted, which participants found illuminating. At the same time, they expressed a desire for a broader sense of adequacy—whether their collection of principles was "good enough" as a whole—rather than feedback on isolated items. They also asked for features such as access to the chatbot's answer history, or the ability to evaluate principles against more varied and challenging queries.

Taken together, these observations suggest that while assistance lowered the effort of principle authoring, it also reduced opportunities for critical engagement. Manual authoring demanded more from participants, but fostered greater ownership and reflection. In both cases, participants highlighted the need for systemic feedback and richer evaluation resources to situate their principle writing within the larger goals of service governance.

6 Discussion

As a research probe, Trust Mediator surfaces how people work with principle sets and the kinds of reasoning they undertake when such sets are made explicit. Across both conditions, participants consistently found principles to be a useful way to evaluate and improve chatbot behavior. Manual authoring encouraged more precise and operational rules, while assisted authoring tended to expand coverage and include more affective aspects. In both cases, principles themselves served as a valuable lens for reflection, helping participants articulate expectations and reason about chatbot behavior. These contrasts help explain why the two modes afford different strengths and set the stage for discussing principles as

Table 5: Differentiated functions of the assistive scaffolds in the assisted workflow.

Scaffold	Primary Function	Typical Profile	Observed Trade-off
Persona Simulation	Expanded emotional and organisational coverage	Explorers	Broader scope, lower specificity
Principle Elicitation	Supported productivity and initial phrasing	All (esp. Explorers)	More generalised wording
Conflict / Overlap Detection	Enabled reflective assessment of LLM consistency	Balanced	Awareness rather than correction
Evaluation	Sustained continuous self-assessment and micro-adjustment	Refiners	Reflection without major content change

both operational requirements and communicative commitments in chatbot governance.

Prior work has often treated principles in a divided way, *either* as visible commitments in governance frameworks and participatory HCI artifacts, *or* as operational requirements in evaluation and alignment tools. However, both should be considered and our findings bridge this gap by showing how different authoring modes emphasize different qualities of principles. Manual authoring produced longer, more detailed statements that increased operational specificity, while assisted authoring more consistently introduced affective safeguards and helped prune redundancies. In interpreting these differences, it is important to note that SI20 functions as a diagnostic rubric rather than a target scale. Very high scores would imply requirement-style prescriptions that are overly rigid, whereas the mid-range values observed in our study indicate principles that were concrete enough to guide chatbot behavior while remaining adaptable across contexts. This balance suggests that SI20 is not only useful for post-hoc analysis, but could also inform future authoring tools by providing lightweight feedback on principle formulation (e.g., flagging vague wording or highlighting missing temporal anchors) without pushing authors to maximize scores. These contrasts illustrate how principles can serve a dual function: constraining chatbot behavior as operational requirements and articulating expectations as visible commitments. Our three analytic dimensions map onto this dual role in complementary ways. *Specificity* and *Coverage* primarily strengthen the operational side by constraining system behavior and broadening its governed scope, though very high specificity may reduce communicability when rules become overly technical. *Coherence* matters for both sides: contradictions or overlaps weaken chatbot performance and also undermine intelligibility for stakeholders. This nuance highlights that principles are not cleanly separable into "operational" or "communicative" categories, but instead embody tensions between the two. Coherence and the careful curation of principle libraries thus emerge as requirements not only for system testing but also for transparent communication.

Our study also revealed process differences. Manual and assisted pathways each brought distinct strengths. Manual authoring promoted elaboration and operational detail, which raised specificity and supported consistent principle sets. Assisted authoring, by contrast, broadened coverage and introduced affective safeguards, while also surfacing redundancies even if participants did not always act on those warnings. Despite these differences, both pathways produced improvements in chatbot answer quality, but the value seems to lie in their complementarity: manual drafting enriched principles with detail and precision, while assistance tended to provide breadth and more general principles. Taken together, this suggests a phased workflow in which participants alternate between (1) drafting and testing principles manually to capture nuance, (2) receiving principle elicitation support to surface higher-level principles, and (3) getting targeted assistance to consolidate, prune, and resolve inconsistencies in the principle set. Recent work in educational contexts has similarly shown that the timing and modality of assistance shape engagement [17, 21]. Continuous support can sometimes reduce sense of ownership, while phased or balanced support sustains reflection [3]. Our findings resonate with this dynamic: assistance changed how participants worked with principles, but did not simply increase quality across the board. This observation motivates a phased workflow as future work: begin alternating manual and supported principle elicitation to encourage both careful and broad elaboration, then introduce targeted assistance to diagnose contradictions, redundancies, and zero-impact items, and finally consolidate the surviving principles. Such interleaving aims to preserve engagement and ownership while exploiting assistance for critique and refinement.

7 Limitations

Our study provides a probe into principle authoring and should be interpreted in this light. While the system did not incorporate a production backend, the use of a real service description ensured that participants worked within plausible operational constraints. The simulated retrieval layer included in the prompt enabled us to maintain consistency in system behaviour while focusing our evaluation on the authoring workflow. However, the small sample size limits the statistical power and the generalizability of our findings. Participants had prior experience evaluating or configuring chatbots, but they were not the owners of the specific service and corresponding chatbot used in this study. This choice was appropriate for examining the workflow rather than service-specific decision-making, though the principles they authored may differ from those created by service owners working with deployed systems. Similarly, the bounded 30-minute authoring session captures only a snapshot of governance work, which in practice is likely to

unfold over days or weeks and involve negotiation among multiple stakeholders. However, this makes the support for authoring and maintaining a specific, coherent and broad set of principles even more relevant as it will be more difficult in such a context.

Our analytic approach also introduces simplifications. We assessed coverage at the level of three top-level families (cognitive, emotional, organizational), which does not cover finer grain distinctions within each category. In addition, our system design choices—such as the specific prompts used for principle elicitation and evaluation—likely shaped the principles participants produced, and alternative scaffolds might yield different outcomes. Finally, while our analysis highlights the potential of principles as commitments visible to end users, we did not directly study how end users themselves interpret or respond to such commitments. Exploring this perspective remains an important direction for future research.

8 Conclusion

This paper presented Trust Mediator, a workbench for creating and testing governing principles for LLM-powered chatbots. By treating principles as design objects, the system helps service owners define, refine, and assess guidelines that connect organizational values with chatbot behavior. Our exploratory study compared manual and assisted authoring and showed that the two approaches have complementary strengths: manual authoring encourages specificity and clarity while assisted authoring tends to encourage broader coverage and attention to emotional aspects. The findings should be read as design insights rather than general conclusions. Our study highlights recurring tensions—between (a) specificity and coverage, (b) conciseness and redundancy, and (c) assistance and ownership—that matter for future tool design. In both conditions, participants found value in making principles explicit, which helped them reflect on chatbot behavior and consider how trust can be managed.

From this work we suggest three directions for future design. First, support workflows that combine manual and assisted authoring, adding non permanent assistance, in particular to expand coverage and resolve overlaps. Second, recognize that principles serve a dual role as (1) requirements to guide system behavior and (2) commitments to communicate values to end users. Third, build and provide libraries of reusable principles that can support both, systematic chatbot configuration and clearer communication with all stakeholders, including end users. Taken together, these results point to principles as a useful layer for building trust in organizational AI that constrain chatbot behavior and make values visible. Future work should test these ideas in larger and longer-term studies, and in real-world service contexts where negotiation among multiple stakeholders comes into play.

References

[1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. arXiv:2112.00861 [cs.CL] https://arxiv.org/abs/2112.00861

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

[3] Zeya Chen and Ruth Schmidt. 2024. Exploring a Behavioral Model of "Positive Friction" in Human-AI Interaction. In *Design, User Experience, and Usability — HCII 2024*. Lecture Notes in Computer Science, Vol. 14713. Springer, Cham, 3–22. doi:10.1007/978-3-031-61353-1_1

[4] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:1706.03741 [cs.LG] https://papers.nips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences

[5] Norman Cliff. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin* 114, 3 (1993), 494–509.

[6] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. 263–274. doi:10.1145/3301275.3302316

[7] Luciano Floridi and Josh Cowls. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* (2019). doi:10.1162/99608f92.8cd550d1

[8] Batya Friedman, Peter H. Kahn, and Alan Borning. 2002. Value Sensitive Design: Theory and Methods. In *Proceedings of the 7th International Conference on Information Systems*.

[9] Deep Ganguli, Amanda Askell, Yuntao Bai, Anna Chen, Anna Goldie, Azalia Mirhoseini, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Jack Clark, Sam McCandlish, and Dario Amodei. 2022. Red Teaming Language Models with Language Models. *arXiv preprint arXiv:2202.03286* (2022). https://arxiv.org/abs/2202.03286

[10] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. doi:10.1145/3458723

[11] Christian Heath, Jon Hindmarsh, and Paul Luff. 2010. *Video in Qualitative Research: Analysing Social Interaction in Everyday Life*. SAGE Publications, London, UK.

[12] High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Accessed: 2026-01-26.

[13] IEEE. 2019. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. IEEE. https://standards.ieee.org/industry-connections/ec/ead/ First edition. Accessed: 2026-01-26.

[14] ISO/IEC/IEEE. 2018. 29148:2018—Systems and software engineering—Life cycle processes—Requirements engineering. Standard. https://www.iso.org/standard/72089.html Second edition, 31 January 2018.

[15] Brigitte Jordan and Austin Henderson. 1995. Interaction Analysis: Foundations and Practice. In *The Journal of the Learning Sciences*. Vol. 4. Taylor & Francis, 39–103. doi:10.1207/s15327809jls0401_2

[16] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 306, 21 pages. doi:10.1145/3613904.3642216

[17] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv:2506.08872 [cs.CL] https://arxiv.org/abs/2506.08872

[18] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems (CHI '15)*. 1603–1612. doi:10.1145/2702123.2702548

[19] Michael Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. 1–14. doi:10.1145/3313831.3376445

[20] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18, 1 (1947), 50–60.

[21] MIT Media Lab. 2025. Your Brain on ChatGPT. https://www.media.mit.edu/projects/your-brain-on-chatgpt/overview/ Accessed: 2026-01-26.

[22] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 220–229. doi:10.1145/3287560.3287596

[23] OECD. 2019. OECD Principles on Artificial Intelligence. https://oecd.ai/en/ai-principles. Accessed: 2025-09-02.

[24] Anna-Marie Ortloff, Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas, and Matthew Smith. 2025. Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation. In *CHI Conference on Human Factors in Computing Systems (CHI '25)* (Yokohama, Japan). Association for Computing Machinery, New York, NY, USA, 28 pages. doi:10.1145/3706598.3713671

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

[26] Ethan Perez, Sam Ringer, Kamil Chen, He He, Ledell Wu, Zhengxuan Jiang, et al. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251 [cs.CL] https://arxiv.org/abs/2212.09251

[27] Savvas Petridis, Benjamin D Wedin, James Wexler, Mahima Pushkarna, Aaron Donsbach, Nitesh Goyal, Carrie J Cai, and Michael Terry. 2024. ConstitutionMaker: Interactively Critiquing Large Language Models by Converting Feedback into Principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 853–868. doi:10.1145/3640543.3645144

[28] Jie Ren, Thomas D. Krafft, Sang Won Bae, Pieter Vermaas, and Wil Van Der Aalst. 2021. Beyond "One-Size-Fits-All" Explanations: Using Explanations Tailored to Users' Mental Models to Improve Trust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. 1–13. doi:10.1145/3411764.3445407

[29] Mona Sloane, Emanuel Moss, and Renée Chowdhury. 2022. Participation is Not a Design Fix: Participatory AI from Critical Perspectives. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, Vol. 6. 1–24. doi:10.1145/3555100

[30] Aarohi Srivastava, Abhinav Rastogi, Abhinay Rao, et al. 2022. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. arXiv:2206.04615 [cs.CL] https://arxiv.org/abs/2206.04615

[31] Harini Suresh, Haoyue Shen, Meredith Ringel Morris, and Michael Muller. 2024. Participatory AI Governance: Bridging Institutional Principles and Situated Practices. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. 1–14. doi:10.1145/3613904.3642498

[32] Thibaut Thonet, Jos Rozen, and Laurent Besacier. 2025. ELITR-Bench: A Meeting Assistant Benchmark for Long-Context Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.28/

[33] Haoming Zheng, Xinyi Wu, Yu Wang, and Brent Hecht. 2023. Charting the Landscape of Human-LLM Evaluation: Practices, Challenges, and Opportunities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. 1–15. doi:10.1145/3544548.3580685