

Neurodiversity-Informed AI Communication Support: Development and Evaluation of an Individualized Calibration Method

Steven Boyle | Northline Advisory | 19 October 2026

Northline Research | Research Paper | Version 1.0

Abstract

Communication-support systems can respond to individual preferences, but effective personalization depends on whether those preferences can be elicited, represented and applied in a particular task. This study developed a 28-item calibration instrument comprising 22 questions and six situational scenarios and evaluated its outputs using controlled synthetic comparisons. An exploratory 12-persona study and an independent 60-case comparison indicated improved judged suitability of individualized support recommendations over ordinary personalization. The corrected questionnaire-only 180-case study estimated a mean advantage of +0.335 points on a weighted 0–4 scale (95% confidence interval, +0.282 to +0.388). This statistically supported improvement did not satisfy the prespecified provisional +0.40 magnitude criterion, and one critical safety violation prevented satisfaction of the zero-violation gate. Later compression, source-mapping, domain-ablation and task-routing investigations examined whether questionnaire-derived accommodations improved ready-to-send messages. Neither the complete instruction pack nor prompt-only task routing established superiority in the tested tasks, and routing diagnostics identified omitted accommodations and unauthorized interventions. The findings support further investigation of structured elicitation while motivating—but not validating—governed interpretation and execution. All evaluations used synthetic cases and AI judging; no human communication benefits have been established.

Index Terms

Assistive artificial intelligence; communication accommodation; human–AI interaction; individualized calibration; neurodiversity; personalization; synthetic evaluation.

I. Introduction

Autistic and other neurodivergent individuals can encounter difficulties in communication environments that rely heavily on implicit meaning, nonverbal cues, rapid responses and unstated social expectations. Differences in how information is processed, interpreted and expressed may lead to misunderstandings, particularly when the intentions of others are uncertain or when communication takes place under stress. These difficulties are not necessarily attributable to one participant. Research concerning reciprocal misunderstanding has demonstrated the importance of considering how differences in communication expectations affect both parties to an interaction.

Large language models (LLMs) offer a potential means of supporting individuals in these circumstances. They can examine written exchanges, explain alternative interpretations, identify assumptions and help formulate responses. Existing systems also provide personalization through conversational context, saved preferences and user-defined instructions. An individual can therefore adapt some aspects of an assistant's behaviour without requiring a specialized communication-support system.

The effectiveness of this approach, however, depends in part on the individual's ability to identify and communicate the support they require. Someone may know that certain interactions are difficult without being able to specify the instructions that would make an AI assistant more helpful. Preferences may also differ between professional and personal circumstances, change under pressure or involve competing objectives. For example, an individual may want direct feedback while also wanting the assistant to preserve their original language when preparing a sensitive message. Ordinary conversational personalization can accommodate such preferences when they are expressed, but it does not necessarily provide a systematic process for identifying, examining and resolving them.

The practical origin of this investigation was an informal, individually developed set of AI instructions for organizing information and supporting executive attention. It asked an assistant to distinguish verified facts from inferences, prioritize consequential matters, make conflicts and uncertainties visible, and present information in a form that supported action. Its apparent usefulness in one workflow was not experimentally measured. More importantly, the instructions had accumulated through experience, lacked a consistent research-linked method for eliciting other people's preferences, and could not establish whether a similar approach would be useful beyond that individual case. The research question arose from the need to make this form of accommodation explicit, inspectable and testable.

This research investigates whether a structured method of eliciting communication preferences can address that

limitation. Drawing on established research concerning communication differences, accommodation, reciprocal understanding, cognitive demands and individual autonomy, a calibration instrument was developed comprising 22 questions and six situational scenarios. The instrument examines communication-support requirements from several perspectives, including information presentation, interpretation of ambiguity, decision support, contextual variation, response to pressure and preservation of personal communication style. The proposed full method would allow individuals to clarify responses and review the behavioural instructions derived from them before those instructions were applied to an AI assistant. This interactive review and authorization step was a design objective, not an activity completed by human participants in the reported experiments.

The principal research question is whether information obtained through this structured calibration method produces more appropriate individualized AI communication support than assistance generated without personal information or assistance informed by ordinary conversational personalization. To investigate this question, controlled synthetic experiments compared recommendations generated under three conditions: a basic LLM supplied only with the communication task, an LLM supplied with an open-ended description of individual preferences, and an LLM supplied with information derived from the structured calibration instrument.

The investigation evaluates the suitability of the resulting recommendations against specified individual preferences and communication circumstances. It does not assume that additional information necessarily improves support, nor does it treat conformity to conventional communication expectations as an inherently desirable outcome. The experiments first examine whether structured intake adds value under progressively stronger comparison controls. Later developmental studies investigate whether the questionnaire answers translate into relevant instructions and whether those instructions are reliably expressed in communication tasks. These are related but distinct questions: an instrument may capture useful preferences without a model applying all of them appropriately. They do not establish the effectiveness of the proposed interactive interview with human participants, which remains a separate stage of evaluation.

II. Literature and Conceptual Foundations

The development of an individualized communication-support method requires an understanding of both the difficulties that may arise during interpersonal communication and the conditions under which assistance is useful. Research concerning communication accommodation, reciprocal

understanding, cognitive demands and individual autonomy provides a basis for identifying these conditions. Although these areas of research approach communication from different perspectives, they collectively suggest that effective support depends on the characteristics of the individual, the circumstances of the interaction and the purpose of the communication.

A. Communication Accommodation and Individual Preferences

Communication accommodation theory examines how individuals modify their communication in response to others. Giles, Taylor and Bourhis [1] demonstrated that accommodation influences interpersonal interaction and the responses of communication partners. The relevance to AI communication support is that successful communication cannot be assessed solely by the clarity or conventional appropriateness of an individual message. The preferences and expectations of those involved, together with the intended outcome, must also be considered.

For neurodivergent individuals, communication environments may place demands on the interpretation of indirect language, implicit expectations, nonverbal information and conversational timing. However, the presence of a particular difficulty does not establish which form of assistance an individual wants. One person may prefer explicit explanations of implied meaning, while another may prefer several possible interpretations or a question seeking clarification. These differences make it inappropriate to derive a standard set of AI communication instructions from a diagnostic category alone.

An established precedent for eliciting individualized accommodations is the AASPIRE healthcare toolkit described by Nicolaidis and colleagues [2]. Its Autism Healthcare Accommodations Tool allows autistic adults to communicate their own accommodation requirements to providers. That healthcare-specific work supports the practice of making preferences explicit; it does not establish that CSA questionnaire items or AI outputs are valid.

The calibration method therefore examines individual preferences for directness, information presentation, response pacing and interpretation of ambiguity. These preferences are treated as information to be elicited rather than characteristics that can be assumed from an individual's neurotype.

B. Reciprocal Understanding and Communication Authenticity

Milton's account of the double empathy problem [3] challenges explanations that locate communication difficulty exclusively within autistic individuals. Misunderstanding may arise when people with different experiences and expectations interpret one another's behaviour through their own assumptions. This perspective is relevant to communication-support systems

because an assistant that evaluates only the neurodivergent person's behaviour may reinforce an incomplete account of the interaction.

Crompton and colleagues [4] provided empirical evidence relevant to this reciprocal account by comparing information transfer among autistic and non-autistic participants. Their findings reinforce the importance of examining communication as an interaction between people rather than inferring all difficulty from one person. They do not determine the appropriate AI intervention in a particular exchange.

An AI assistant may help an individual consider several interpretations of another person's message, identify assumptions that are not supported by the available information or examine how their own communication might be understood. Such assistance should not presume that one participant's interpretation is correct or that the objective is necessarily to conform to conventional social expectations.

Research by Hull and colleagues [5] on social camouflaging among autistic adults provides a related consideration. Participants described modifying or concealing aspects of their behaviour to meet social expectations, often with significant personal consequences. This raises an important distinction between helping someone communicate effectively and encouraging changes that conflict with their preferred way of expressing themselves.

Consequently, the calibration instrument examines both the assistance an individual wants and the boundaries they place on that assistance. It asks whether the AI should identify possible misunderstandings, challenge interpretations, suggest alternative wording or preserve an individual's original communication style. The appropriate response may vary by situation, and a potentially useful alternative should not automatically replace the individual's own words.

C. Cognitive Demands, Context and Decision Support

The Autistic SPACE framework of Doherty and colleagues [6] identifies sensory needs, predictability, acceptance, communication and empathy as considerations in autistic healthcare encounters. It provides context for investigating processing pace and situational accommodations, while its clinical setting limits any direct claim about CSA performance in ordinary conversations.

Communication support also involves the way information is presented and decisions are considered. Complex or ambiguous exchanges may require an individual to interpret several possible meanings, retain contextual information, evaluate consequences and formulate an appropriate response. The demands associated with these activities can vary with time pressure, fatigue, sensory conditions, emotional intensity and the importance of the decision.

These considerations suggest that communication preferences should not be treated as fixed instructions that apply identically in every circumstance. An individual who normally wants comprehensive explanations may prefer shorter responses when overloaded. Someone who generally welcomes direct recommendations may instead want alternatives and additional reasoning when a decision has substantial consequences.

The calibration instrument addresses these variations by examining information format, processing pace, uncertainty, recommendation strength and the circumstances that change an individual's support needs. Situational scenarios supplement the direct questions by examining how stated preferences apply when communication objectives or contextual demands conflict.

These dimensions are used as practical categories for developing the instrument. They are not presented as independently validated psychological factors, and the present research does not establish that any particular adaptation is universally beneficial.

D. Individual Autonomy and Human–AI Interaction

Self-determination theory identifies autonomy as an important component of human motivation and well-being [7]. In the context of AI communication support, autonomy concerns the individual's ability to determine the purpose of the assistance, control the extent of intervention and retain responsibility for their own communication and decisions.

Research on human–AI interaction provides additional guidance. Amershi and colleagues [8] identified principles concerning the communication of system capabilities, the management of uncertainty, user correction and control over AI behaviour. These principles are particularly relevant when a model interprets an individual's answers and converts them into instructions that may influence subsequent interactions.

An individual's description of a communication difficulty does not necessarily constitute permission for the assistant to intervene. Similarly, a preference expressed in one context may not authorize the same behaviour in another. A calibration process must therefore distinguish between the person's original response, the system's interpretation of that response and the behavioural instruction proposed for use by the AI.

The CSA method incorporates this distinction by retaining the source of each proposed accommodation and providing for individual review, modification or rejection before the instruction is approved. This approach is consistent with the broader emphasis on transparency, accountability and human oversight in the NIST AI Risk Management Framework [9]. These principles inform the design of the method but do not, by themselves, establish its effectiveness.

E. Implications for Instrument Development

[See Figure 1]

The literature identifies several considerations that must be addressed when developing individualized AI communication support. Communication preferences differ between individuals, misunderstandings may be reciprocal, contextual demands can change the assistance required, and effective support must preserve the individual's authority over their own communication.

Research on personalized language models also provides an important comparison. Salemi and colleagues introduced LaMP [10], a benchmark evaluating personalized language-model outputs across varied tasks and profile information. That work illustrates existing approaches to personalization; the question tested here is more specific: whether a structured communication-calibration record improves the appropriateness of support recommendations relative to independently elicited ordinary preferences.

These findings informed the development of a structured calibration instrument covering communication preferences, information processing and presentation, decision support, contextual variation, reciprocal interpretation and the boundaries of AI intervention. The instrument combines direct questions with situational scenarios to capture both stated preferences and the conditions under which those preferences apply.

The initial design also drew on the shortcomings of an informal personal communication-support skill. That precursor expressed a working set of preferences as direct operating instructions, without a systematic means of eliciting equivalent guidance for another person or establishing why particular instructions were appropriate. The research-informed instrument was developed to replace that informal accumulation with explicit questions, source-linked candidate behaviours and a reviewable path from stated preference to possible accommodation. The precursor explains the engineering motivation; it does not serve as evidence that the instrument improves outcomes.

The underlying research provides a rationale for examining these dimensions, but it does not prescribe the particular wording of the questions or demonstrate that the resulting instrument improves AI assistance. Those are matters of instrument design and empirical evaluation.

These considerations explain the design requirements for a communication-support instrument: it must elicit preferences without inferring them from diagnosis, distinguish difficulty from desired assistance, and record conditions under which support should change. Section III describes how these requirements were expressed as questions, scenarios and candidate model instructions.

III. Instrument Development and Behavioural Mapping

A. Development of the Calibration Instrument

The instrument was constructed to elicit the assistance a person wants from an AI system rather than infer support needs from a diagnostic label. This addressed the first limitation of the informal precursor: instructions that happened to work for one person could not simply be generalized to others. Its questions translate the design requirements established in Section II into choices about the assistant's communication, interpretation, decision support and intervention boundaries. The instrument was developed for calibration; it was not intended to measure a psychological trait or diagnose neurodivergence.

The design problem was to connect a person's own statements to assistance that an LLM could provide. It therefore asked both about ordinary communication preferences and about what the assistant should do when a message is ambiguous, a decision is difficult or the person wishes to preserve their original wording.

The resulting research instrument contains 22 questions and six situational scenarios. The questions combine structured response options with opportunities for individuals to explain their preferences in their own words. Structured options provide a consistent basis for comparison and interpretation, while narrative responses allow individuals to describe circumstances that may not be adequately represented by predefined categories.

The instrument is intended to form the basis of an interactive interview rather than operate solely as a questionnaire completed without assistance. During the proposed interview, an individual may clarify an answer, describe an example or explain how their preferences change under different circumstances. The structured questions maintain consistency across interviews while allowing the process to reflect the individual's actual experience.

The questions and scenarios were selected as engineering responses to the literature and to practical communication-support requirements. They were not validated as a psychometric scale, nor were they designed to classify or diagnose individuals.

B. Communication-Support Areas Covered by the Instrument

[See Table I]

The first group of questions examines communication preferences and the interpretation of information. These include the importance an individual places on precision, directness, contextual detail and being understood as intended. Questions

also examine how individuals interpret implied meaning, how their own communication may be misunderstood, which information formats they find useful and whether they prefer additional time or incremental explanations when processing complicated information.

These questions address a practical distinction between an individual's communication characteristics and the assistance they want from an AI system. For example, someone who finds indirect language difficult may want the assistant to identify possible interpretations, ask for more context or explain why the available information is insufficient. The appropriate response cannot be determined from the difficulty alone. It must be established through the individual's expressed preference.

A second group examines decision support and uncertainty. Questions address how alternatives should be presented, whether the assistant should make recommendations, what information is needed to evaluate a decision and how uncertainty should be communicated. This allows the calibration process to distinguish between an individual who wants a direct recommendation and someone who prefers a comparison of alternatives before making their own decision. The instrument also examines the circumstances under which these preferences may change.

A third group considers the effect of context and personal state on communication. Stress, fatigue, sensory conditions, time pressure and interpersonal conflict may affect the type or amount of assistance an individual finds useful. The questions ask individuals to identify relevant conditions and describe whether they want the assistant to adjust its behaviour. These responses are intended to support conditional instructions rather than permanent assumptions about the person's abilities or state.

The final group examines how the assistant should behave when interpreting interpersonal communication, challenging a conclusion or rewriting a message. These questions are particularly relevant to reciprocal understanding and autonomy. An individual may want the assistant to identify unsupported assumptions or explain how another person might interpret their words, while retaining control over whether any suggested changes are adopted. The instrument therefore asks about acceptable intervention, preservation of personal voice and behaviours the assistant should avoid.

The questionnaire was initially organized for presentation as communication (Q1–Q6), decision support (Q7–Q10), state and context (Q11–Q14), AI behaviour (Q15–Q22), and scenario calibration (S1–S6). For the later 48-case component study, these items were reorganized into six prespecified operational domains. Presentation and processing comprised Q1, Q2, Q5 and Q6; decisions and evidence comprised Q7–Q10, Q17, Q18 and S3; state and context comprised Q11–Q14 and S5; interpretation and uncertainty comprised Q4, Q15 and S1;

authenticity and rewriting comprised Q19–Q22, S2 and S4; and reciprocal perspective comprised Q3, Q16 and S6. Every item appeared in one domain. These were functional units for testing instruction contributions, not validated psychological factors. Their experimental results are presented later in the paper, after the comparative evaluation that motivated component testing.

C. Situational Calibration

Direct questions are useful for identifying preferences, but a stated preference may not provide enough information to determine how it should be applied. Someone who generally wants direct feedback, for example, may prefer a more cautious approach when discussing a personal relationship. Similarly, an individual who wants the assistant to challenge unsupported conclusions may not want every subjective interpretation questioned.

Six situational scenarios were developed to examine these distinctions. They address ambiguous interpersonal messages, frustrated outgoing communication, unsupported general conclusions, conflicts between effective wording and personal authenticity, information overload and reciprocal misunderstanding.

Each scenario presents a communication problem and asks what assistance the individual would want. The objective is not to identify a correct response to the situation. Instead, the scenarios explore how the individual's stated preferences apply when circumstances introduce uncertainty, competing objectives or potential interpersonal consequences.

For example, the authenticity scenario describes an accurate message that the AI believes could receive a more favourable response if expressed more warmly. The individual is asked whether the assistant should change the message, explain the trade-off, present alternatives or preserve the original wording. This distinguishes a preference for successful communication from permission to modify the individual's personal style.

The reciprocal-misunderstanding scenario examines an interaction in which two people may interpret one another's behaviour differently. It asks whether the assistant should explore both perspectives and how it should represent uncertainty about the other person's intentions. This provides a practical application of the double empathy concept without assuming that either participant is responsible for the misunderstanding.

The scenarios therefore supplement the questionnaire by identifying circumstances in which general preferences require qualification. They also provide information for developing instructions that depend on the context of the interaction rather than applying uniformly to every communication task.

D. Translating Responses into Behavioural Instructions

The intended output of the calibration process is a set of proposed instructions linked to the individual's answers. The design mapping records, for each item, its research basis, the support behaviour an answer may indicate, and the interpretation or intervention that must not follow without authorization. In the proposed human process, an answer remains distinct from the interpretation placed on it and from an instruction the person has actually approved.

The source mapping specifies different treatment for explicit choices and descriptions of experience. A selected preference such as "Ask one question at a time" (Q6) supports a provisional instruction for incremental questioning. By contrast, reporting that pressure changes communication (Q11) or listing sources of difficulty (Q12) does not authorize the model to shorten responses, defer action or challenge the individual. Question 13 asks what assistance is actually wanted in those circumstances. Free-text and scenario responses likewise require selective interpretation and confirmation rather than automatic conversion into rules.

Each candidate rule retains its questionnaire item and original response. For synthetic-case construction, the record could mark an item as used, context only, withheld pending confirmation, or in conflict. These dispositions preserve the difference between information that describes a person and information that permits a particular intervention. They are not evidence of consent from human participants; none participated in these experiments.

The design also sets an order for resolving competing instructions. Provider safety constraints and confirmed prohibitions or authenticity boundaries take precedence over optional support actions. A context-specific preference may replace a general default only when the relevant context is known or confirmed. Directly selected support actions take precedence over inferred needs; conflicting selections call for clarification rather than silent optimization. The scenarios may expose an exception, but do not themselves override a questionnaire answer. These are design rules for constructing candidate support instructions, not proof that a model always follows them.

The proposed interactive process therefore includes review of the resulting instructions. The individual would be able to accept, modify or reject each meaningful accommodation before it becomes part of an approved calibration profile. This is intended to prevent the system from treating an interpretation or an apparent communication difficulty as sufficient authority to change its behaviour.

This division of authority constitutes the interpretation-governance requirement of the proposed method: the source answer must remain inspectable, an inferred need must not

silently become permission, contextual qualifications must travel with the instruction, and the person must be able to confirm, modify, reject or withdraw an accommodation. These are design requirements for a future participant-facing system, not measured properties of the synthetic evaluation. They also differ from the subsequent question of whether a model correctly applies an authorized instruction when composing a particular message.

E. Examples of Answer-to-Strategy Mapping

The following illustrations use response options from the frozen questionnaire registry and behaviours from the research-to-behaviour mapping. They describe proposed actions subject to confirmation, not experimentally validated effects of particular questions.

Question 2 asks about preferred explicitness. “Say exactly what they mean” supports a more direct communication baseline; “I am comfortable inferring meaning from context” does not call for routine explanations of every implied meaning. Neither choice establishes the person’s ability or inability to understand indirect language. Question 6 distinguishes “Ask one question at a time” from “Give me the information and let me think.” The first may justify incremental questioning, while the second supports giving the person room for independent processing without unnecessary follow-up prompts.

Question 17 concerns the circumstances in which an assistant should challenge an interpretation. “Challenge more strongly when evidence contradicts me” supports an evidence-dependent challenge; “Ask questions before challenging” calls for a different approach. Question 21 concerns the trade-off between effectiveness and authenticity. “Preserve my original style” protects the person’s own expression, while “Show both versions and explain the trade-off” supports an optional alternative. Neither authorizes unrequested replacement of the original message. The mapping also requires Question 20’s stated authenticity boundaries to be respected.

Question 16 addresses reciprocal interpretation: “Only when I ask” restricts unsolicited discussion of how another person may interpret the individual’s words, whereas “Only when misunderstanding seems likely” sets a conditional trigger. The mapping prohibits the assistant from inventing the other person’s motives. The difference between these selections matters because a generally plausible accommodation may be intrusive or inappropriate for someone who has not requested it.

For example, two fictional people may request help interpreting the same ambiguous message. One may authorize explicit discussion of competing interpretations and a clarifying question before any challenge; the other may prefer only a brief statement of what is known, with alternative interpretations offered upon request. The same incoming message could

therefore warrant different assistance. This is an explanatory example of conditional calibration, not a participant observation or a measured experimental result.

The complete 22-question, six-scenario instrument and its response options should be provided as Supplementary Instrument A. A separate mapping crosswalk should reproduce the research construct, item, response-dependent candidate behaviour, clarification requirement and prohibited inference. The original registry and design mapping are distinct sources: the registry records exact response options, while the mapping records proposed behaviour and restrictions. These artifacts define the intervention under study; they do not establish that each item is necessary or beneficial.

F. Calibration Profile and Experimental Application

[See Figure 2]

The complete proposed method consists of four distinct activities: eliciting a person’s preferences through a guided interview; interpreting answers into proposed, conditional accommodations; allowing that person to inspect, change or reject those accommodations; and applying the approved instructions to later communication tasks. Only part of this sequence was represented in the synthetic experiments. No person completed a live interview or approved a generated profile as part of the reported studies.

Instead, the experiments constructed fictional cases with specified communication circumstances and questionnaire answers, then generated recommendations using differing amounts and types of individualized information. Consequently, the principal comparisons test whether structured questionnaire-derived information can improve the judged suitability of responses under controlled synthetic conditions. They do not test interview usability, individual approval, or sustained personal benefit. This defines the experimental intervention introduced next.

Subsequent source-linked compilation and domain-removal work also addressed a different concern: whether collecting more preferences necessarily produces better messages. A traceable answer-to-instruction mapping establishes what the model was permitted to do, but not whether every permitted instruction is needed for every task. The later component and task-routing studies were therefore developmental investigations of instruction selection and execution, rather than new confirmations of the original questionnaire-only comparison.

IV. Experimental Design and Methods

A. Experimental Questions and Common Comparison Framework

The experimental program developed through a sequence of connected questions. The first study asked whether a structured intake produced enough signal to warrant a larger test. The next addressed whether that signal survived when ordinary personalization was elicited independently. Larger studies tested the intervention under more demanding information, scoring and safety controls; a correction was then required when one execution exposed additional Profile Builder information in the structured condition. Finally, diagnostic and component studies examined whether output length, mapping density, irrelevant accommodations and prompt-only delivery could explain weaknesses revealed by the earlier comparisons. The experiments were not one pooled trial of a fixed product: each addressed a defined uncertainty created by the work that preceded it.

The use of language models as judges has methodological precedents as well as recognized limitations. Zheng and colleagues [11] examined model-based assessment of conversational outputs and identified possible position, verbosity and self-enhancement biases. Accordingly, the blinded repeated judgments in the CSA experiments measure performance under the specified synthetic evaluation procedure, not agreement with unobserved human preferences.

Three information conditions formed the principal comparison. The generic condition (A) received a communication task without a personal preference profile. The ordinary-personalization condition (B) received a description of the person obtained through a neutral, open-ended intake rather than the CSA questionnaire. The structured-calibration condition (C) received information derived from the 22 questions and six scenarios. The main comparison was C against B: it asked whether structured intake added value beyond ordinary personalization, not simply whether personalization was preferable to receiving no information.

Each synthetic case was a fictional individual represented by a defined set of communication preferences, circumstances and a task requiring assistance. The same case supplied the underlying communication problem for each condition. A separate generator produced a response under each permitted information boundary. The resulting responses were presented without condition labels to language-model judges, which scored them against the case requirements and the study rubric. Comparing C and B within the same case reduced variation attributable to differences between fictional people or tasks. The case, rather than the individual judgment, was therefore the unit of the paired comparison.

The frozen evaluation rubric assessed five dimensions on a 0–4 scale: research justification (25% of the composite), fidelity to

the individual's preferences (25%), contextual appropriateness (20%), autonomy and authenticity (20%), and conflict treatment (10%). The weighted score was $0.25(\text{RJ}) + 0.25(\text{PF}) + 0.20(\text{CA}) + 0.20(\text{AA}) + 0.10(\text{CT})$. Scores of 0, 2 and 4 had explicit behavioural anchors for each dimension; scores of 1 and 3 were intermediate judgments between the adjacent anchors. Two blinded judgments were averaged for each case and condition. This makes the rubric reproducible in its dimensions, weights and main anchors, but not fully behaviourally specified at every score level. Critical violations were assessed separately and could fail the safety requirement regardless of the weighted score.

To illustrate the comparison without reproducing an actual retained case, consider a fictional person seeking help with an ambiguous workplace message. In A, the model would receive the message and the request for help. In B, it would also receive the person's independently elicited, open-ended preferences. In C, it would receive the corresponding structured questionnaire-derived guidance within the permitted boundary. The study asks whether the resulting advice fits that same person's specified requirements more closely under C than under B. This illustration describes the design; it is not an additional experimental observation.

B. Feasibility Experiment and Independent Held-Out Comparison

The initial feasibility experiment used twelve constructed personas. Each persona received a recommendation under the generic, ordinary-personalization and structured-calibration conditions, yielding 36 recommendations and 72 judgments. Generation and judging were assigned separate model roles; the outputs were blinded before scoring. Within-persona differences were evaluated using exact sign-flip tests, bootstrap intervals and multiplicity correction. This was an exploratory proof-of-concept designed to determine whether a larger study was warranted, not to supply a final efficacy estimate. Its central limitation was that the ordinary-personalization briefs had been prepared with knowledge of the underlying personas. They were therefore not independent representations of what an ordinary open-ended intake would elicit.

The encouraging initial signal justified further testing, but its ordinary-personalization briefs were not independent. The second experiment addressed that independence weakness directly. Seventy-five new candidate cases were produced, 74 were eligible and 60 entered the comparison. For each case, the neutral open-ended answer used for ordinary personalization was elicited and frozen before the structured questionnaire condition was prepared, so B could not receive questionnaire answers or the instructions derived from them. Every case then yielded an A, B and C recommendation, producing 180 responses and two judgments per response. Separate pinned model roles performed case simulation, response generation and judging. The analysis used 20,000 bootstrap draws and

200,000 permutation draws, with Holm adjustment for the two principal comparisons.

An audit of the first 360 judgments found that the judging prompts included internal case identifiers. These judgments were excluded, and the original 180 recommendations were rejudged after removal of the identifiers. Only the replacement judgments entered the reported analysis. The correction restored the intended information boundary for judging without regenerating recommendations or changing the three comparison conditions.

C. Larger Confirmatory Studies and Information Boundaries

The larger confirmatory design fixed 180 matched synthetic cases and two primary contrasts: structured calibration against ordinary personalization (C minus B) and generic assistance (C minus A). The primary estimate was the case-paired mean difference. The frozen plan specified statistical tests with multiplicity adjustment and, separately, a provisional research-design threshold requiring the lower bound of the two-sided 95% confidence interval for C minus B to exceed +0.40 points on the 0–4 scale. The +0.40 criterion was researcher-selected before confirmatory results were observed; it was not derived from human-user validation, published minimal-important-difference research or an established clinical standard. The later power analysis used the criterion to plan sample size: assuming a true difference of +0.60 and a paired standard deviation of 0.80, approximately 169 cases would give 90% power to clear that confidence-bound rule, and 180 were selected. These calculations explain the design, not the substantive meaning of +0.40. Safety, fairness, blinding and independent statistical-reproduction requirements were evaluated separately.

Before a completed large-scale comparison, a clean-replication execution encountered a conflict between the prompt for concise generic guidance and a validator requiring at least 15 words in every domain. Of 180 attempted generation jobs, 164 completed and 16 exhausted their permitted attempts. The frozen failure rule stopped the execution before output standardization, judging or statistical analysis. Accordingly, this execution supplies no comparative outcome estimate. The failure and subsequent methodological correction were retained in the research record.

The first completed 180-case study applied the balanced design, producing 540 recommendations and 1,080 judgments under the frozen comparison and analysis procedures. Blinding, safety and reliability checks were followed by independent statistical reproduction from the retained evidence. A later boundary audit found that the structured-calibration condition had received supplemental Profile Builder material beyond the intended endpoint of the sixth questionnaire scenario (S6). The completed execution consequently measured an augmented workflow rather than the intended 28-item questionnaire alone.

The distinction was not repaired retrospectively by relabelling the intervention; a separate corrected replication was required. [12]

The corrected replication removed the supplemental Profile Builder information and ended the structured-calibration input at S6. The research question, questionnaire, condition roles, rubric, model assignments, statistical plan and practical criterion were otherwise retained. The candidate pool was extended to 294 cases because the earlier pool did not contain a subset for which the frozen exact selector could certify the required balance. Of these, 252 met eligibility rules; an outcome-blind selection procedure retained 180. Enlarging the pool made a balanced selection possible without increasing the number of analyzed cases.

The corrected replication produced 540 recommendations, each judged twice after condition labels were withheld. Analysis included 50,000 bootstrap draws, paired significance testing, sign-flip testing, multiplicity adjustment, judge-agreement checks, and separate safety and blinding reviews. Provider requests had a 300-second timeout: late responses were excluded under predetermined rules so that the research team could not selectively replace an output after learning its result. A narrow correction to retained citation information occurred before unblinding without altering scores or rationales. Independently reproduced statistical calculations were compared with the retained evidence; this is computational reproduction, not an independent human replication.

[See Figure 3]

D. Subsequent Diagnostics and Component Investigations

After the corrected replication, a six-case diagnostic investigated whether restricting the length of recommendations could obscure distinctions contained in the structured profile. It compared constrained and fuller CSA recommendations, but lacked a comparably expanded ordinary-personalization condition. A twelve-case factorial diagnostic addressed that omission by evaluating ordinary personalization and structured calibration in both compressed and fuller forms. The prespecified interaction compared the gain from fuller output under C with the corresponding gain under B. Both investigations were later explanatory diagnostics, not components of the original confirmatory decision.

After the principal comparisons and compression diagnostics, a source-linked audit addressed whether questionnaire answers could be traced to candidate instructions rather than collapsed into an opaque profile summary. For each questionnaire item, the implemented mapping recorded the relevant functional domain, research construct, permissible instruction, source, authorization condition and prohibited inference. The 180 retained corrected-study cases were compiled without new model calls. The audit measured mapping counts and the

handling of narrative responses, duplicates and explicit conflicts. Its purpose was to establish provenance and assess the volume of candidate rules, not to prove that each rule improved communication. A technically traceable instruction can still be unnecessary, intrusive or poorly timed; this unresolved question led to the component investigation.

[See Table II]

The developmental component investigation then examined whether including all questionnaire-derived instructions was useful for particular communication tasks. Forty-eight synthetic cases were assessed under ten fixed conditions: the full 28-item pack, a prespecified 12-item core, six conditions each omitting one of the six operational instruction domains defined in the frozen mapping registry, an intentionally mismatched full profile and no profile. Each case-condition combination produced two task messages, one concerning ambiguous interpersonal resistance and another concerning a missed deadline with uncertain supporting facts. The study therefore comprised 480 case-condition packets and 960 task messages. Two task scores were averaged within each case and condition, making the case—not the individual message—the unit of paired inference.

The first free-text judging execution in the component study was unusable because its exact-substring evidence validator rejected 575 of 576 scores. Those scores were excluded. A separately frozen recovery introduced deterministic evidence identifiers and rejudged the same 480 case-condition packets without regenerating messages. The recovered evidence comprised 480 primary judgments, 95 of 96 reliability judgments and three triggered safety-verification judgments; a provider response exceeding the frozen queue limit was quarantined without replacement. Analysis used case-paired resampling, Holm-adjusted domain comparisons, task-specific and sensitivity analyses, and independent statistical reproduction. Reliability and optimization rules were retained as decision criteria, so this remained an instrument-development test rather than a validation of an optimized form.

The component study tested the consequences of sending complete or reduced instruction packs to the model; the subsequent developmental pilot tested whether selecting a small number of task-relevant accommodations was a better approach. The final developmental pilot therefore examined task-based instruction selection. Twelve previously uninspected cases performed the two communication tasks under four conditions: ordinary personalization, task-routed structured calibration, the complete source-linked instruction pack and no personalization. The 96 resulting messages were assessed both for blinded quality and, separately, for adherence to the selected route. The route-aware review considered whether selected accommodations appeared, whether unrelated or unauthorized interventions were added, and whether the message represented the intended route. An invalid original judgment format led to exclusion of 73 judgments; a separately

frozen recovery rejudged the unchanged messages. This pilot tested selection and prompt-based delivery of accommodations. It did not implement or test a deterministic output-validation and repair layer.

The distinction between selecting an instruction and verifying its execution is central to the interpretation of this pilot. A selected accommodation could be omitted, and the generated message could add an unrelated intervention even when the intake mapping and routing decision were traceable. A product-level layer for checking those behaviours and rejecting or repairing noncompliant outputs remains a proposed engineering response, not a tested treatment in these studies.

E. Evidence Boundaries and Reproducibility

The studies tested related but nonidentical implementations. The 12-persona study was an exploratory feasibility test. The 60-case study improved independence of ordinary personalization. The original completed 180-case study included supplemental structured-condition information beyond the intended sixth-scenario boundary; the corrected replication removed that exposure and is the primary questionnaire-only comparison. The later compression, mapping, component and routing studies investigated mechanisms and implementation limits rather than forming a single pooled trial. Their numerical findings must be interpreted for the distinct procedures actually used.

[See Table III]

The research used synthetic individuals, model-generated recommendations and AI-based judging to permit controlled, paired comparisons under defined information boundaries. The work separated some model roles, retained evidence, froze decision rules where specified and independently reproduced statistical calculations, but none of these controls is equivalent to direct evaluation with people. The same features limit generalization: synthetic preference records may not capture the complexity of lived circumstances, and judges may reward characteristics that do not correspond to a person's experience of helpfulness, authenticity or autonomy. These limits were not removed by blinding, protocol freezes or independent statistical reproduction. The next sections report the principal comparisons and then examine the developmental investigations in the order in which their questions arose.

V. Results

A. Exploratory Feasibility Study

The initial twelve-persona study tested whether questionnaire-derived communication guidance showed enough promise to justify more rigorous investigation. CSA received a higher suitability score than ordinary personalization in eleven of

twelve cases, with a mean paired advantage of +0.571 points on the 0–4 scale (95% confidence interval [CI], +0.346 to +0.781). It also exceeded generic assistance in all twelve cases, by an average of +1.560 points. These results established a promising early signal, not proof of effectiveness. [12]

The comparison with ordinary personalization was limited because its briefs had been prepared with knowledge of the synthetic personas rather than derived through an independent intake. The principal lesson was therefore not the apparent size of the advantage but the need to test whether it survived when an ordinary user description was obtained independently.

B. Independent Sixty-Case Comparison

The second study corrected that limitation by obtaining and freezing the ordinary-personalization response before constructing the CSA condition. Among sixty independently constructed synthetic cases, mean suitability scores were 2.044 for generic assistance, 3.210 for ordinary personalization, and 3.452 for CSA. The paired CSA advantage over ordinary personalization was +0.243 points (95% CI, +0.143 to +0.342); forty-six of sixty cases favoured CSA, and the difference was statistically significant after adjustment for the two main comparisons. [12]

The advantage remained when the comparison became more credible, but it was markedly smaller than in the exploratory study. This result provided the first independent synthetic comparison that the structured questionnaire contributed useful information beyond a neutral open-ended description. It did not show that the improvement was large enough to meet the practical criterion later adopted for confirmation. A judging-information error required the original recommendations to be scored again without internal case identifiers; the invalid judgments were excluded, and the finding rests on the corrected scoring.

C. Large-Scale Comparison and Correction

An intervening clean-replication attempt stopped before scoring because its response requirements and validation rule conflicted. It produced no interpretable outcome and is not included in the evidence of effectiveness.

The first completed 180-case study found a stronger advantage for CSA over ordinary personalization: +0.589 points (95% CI, +0.528 to +0.653), with 171 of 180 cases favouring CSA. The observed difference exceeded the prespecified practical criterion, and no critical safety violations were verified in the CSA condition. A later audit, however, found that this condition contained additional Profile Builder information beyond the intended twenty-two questions and six scenarios. The result therefore describes an augmented workflow, not the isolated effect of the questionnaire.

This distinction matters because the apparent performance of structured calibration depends on what information the

intervention actually includes. The original result was retained as evidence about the augmented workflow, but a separate replication was needed before drawing conclusions about the questionnaire itself.

D. Corrected Questionnaire-Only Replication

The corrected study removed the additional Profile Builder material while retaining the 180-case paired comparison of individualized support recommendations. Mean scores were 1.113 for generic assistance, 3.059 for ordinary personalization, and 3.394 for questionnaire-only CSA. The paired advantage over ordinary personalization was +0.335 points (95% CI, +0.282 to +0.388), with 148 of 180 cases favouring CSA. The adjusted p-value was 3.65×10^{-26} . This is the principal synthetic estimate of how structured questionnaire information influenced the judged suitability of support recommendations, not the quality of ready-to-send interpersonal messages. [12]

The positive difference was statistically supported, but the study failed its separately specified provisional magnitude and safety gates. Its confidence interval did not clear the researcher-selected +0.40 criterion, and one critical safety violation occurred in CSA against a requirement of zero. Failure to meet this provisional threshold does not invalidate the effect estimate or imply that the measured improvement is without value; rather, the experiment did not establish the magnitude of advantage that its design convention required. Nor does the favourable mean establish that individual users would find the recommendations beneficial.

The difference between the original and corrected 180-case estimates reinforces the importance of defining the intervention precisely. Additional personalized information may improve judged output, but the stronger augmented result cannot be used to claim a larger effect for the 28-item questionnaire. The corrected estimate is narrower and more relevant to the research question.

[See Table IV]

[See Table V]

E. Interpretation of the Principal Findings

The exploratory, independent, and corrected questionnaire-only studies consistently favoured structured calibration over ordinary personalization when the model was asked to recommend appropriate individualized support. The independent and corrected studies are the strongest evidence because ordinary personalization was elicited without access to CSA questionnaire answers. Their contribution is evidence that structured intake can improve the suitability of recommendations about communication accommodations under controlled synthetic judging. This does not establish that a language model can reliably translate those accommodations into better outgoing messages.

That distinction is material. Recommending how to support a person permits an explicit description of preferences, relevant circumstances and boundaries. Producing a ready-to-send message requires a further sequence: translate answers into permitted accommodations, select those relevant to the task, resolve competing instructions and execute them without adding unwanted interpretation or changing the person's facts or voice. The later component experiments test this harder and different proposition. They cannot be treated as a straightforward replication of the recommendation comparisons; equally, their failures cannot be assumed to arise only from execution rather than from the instrument, task coverage or measurement.

The combined evidence therefore supports continued development of structured elicitation, while leaving the independent effectiveness of individual domains and reliable task-level execution unresolved. The corrected study did not meet its provisional magnitude or zero-violation gates, and all findings remain limited to synthetic cases and AI judging. The later studies should be read as a series of tests of the path from questionnaire answer to final communication: whether answers remain traceable, whether the right accommodations are selected, and whether a model actually applies them without unauthorized additions. Section VI reports these findings before Section VII develops an evidence-derived governance hypothesis; neither the proposed controls nor their benefits have yet been tested.

VI. Further Investigation of the Findings

The corrected questionnaire-only experiment established an advantage in judged recommendations about individualized support, but left an important question unanswered: whether the information obtained through calibration would improve a message that a person could actually send. A recommendation can describe a person's preferences and propose accommodations explicitly. A finished message must use the relevant accommodations without turning the message into an explanation of the profile. The investigations that followed therefore examined what might limit this translation, beginning with response length and proceeding to instruction mapping, domain contribution and task-specific delivery.

A. Does Response Compression Account for the Smaller Effect?

The first explanation considered was that short recommendation outputs might obscure distinctions captured by the questionnaire. In a six-case exploratory diagnostic, fuller CSA responses scored 0.483 points higher on average than compressed CSA responses. This indicated that length could matter, but did not establish a disadvantage specific to

structured calibration: ordinary personalization had not been given the same opportunity to produce a fuller response.

A subsequent 12-case factorial comparison included both ordinary personalization and CSA at compressed and fuller lengths. Allowing a fuller response increased the ordinary-personalization score by 0.298 points and the CSA score by 0.306 points. The difference between those gains was 0.008 points (95% CI, -0.246 to +0.246). CSA's estimated advantage over ordinary personalization was also similar under the two length conditions: 0.410 and 0.419 points, respectively. The evidence thus supports a general benefit from fuller expression in these cases, but does not support compression as a distinctive explanation for the reduced questionnaire-only effect. Because the two full-length conditions differed appreciably in average word count, the diagnostic is best interpreted as ruling out an obvious CSA-specific length interaction rather than establishing a universal effect of response length.

The inquiry consequently moved from how much the model could say to what information it was being asked to use. If additional space benefited both comparison methods, a more relevant concern was whether the questionnaire's individual answers produced specific instructions and whether those instructions improved the task for which they were activated.

B. Source-Linked Mapping and Instruction Load

The source-linked mapping audit examined the intermediate product of calibration: the proposed behavioural instructions derived from questionnaire responses. The retained 180 corrected-study cases were compiled using explicit mappings that connected each answer to a research construct, permitted accommodation, authorization condition and prohibited inference. All cases could be processed without new model calls. The compiler produced 48–64 candidate rules per case, with a median of 55. Of 1,980 narrative records, 71 produced actionable candidate instructions; the other 1,909 remained contextual information. The audit therefore demonstrated that the mapping could distinguish an expressed preference from material that merely described a circumstance. [12]

That result was valuable but narrow. The ability to identify the source and permitted scope of an instruction does not establish that the instruction is helpful when applied, or that all proposed instructions should be active at the same time. A median of 55 candidate rules also raised a practical concern: the model might encounter many individually defensible requirements with overlapping effects, competing priorities or little relevance to the immediate task. The audit established traceability of the proposed instruction set, not utility of each item, proof of human authorization, or reliable compliance by the message generator.

C. Domain Ablation on Ready-to-Send Communication

The 48-case component study tested whether the assembled instructions improved two concrete outputs: a response to ambiguous interpersonal resistance and a message about a missed deadline under uncertain evidence. Each case was evaluated with the full 28-item instruction pack, a prespecified 12-item core, six versions each omitting one operational domain, a mismatched full profile and no profile. These were task-level message comparisons, not further tests of how well an assistant could describe a person's desired support. Across ten conditions and two tasks, the study produced 960 messages.

The full instruction pack did not outperform no profile: the mean difference was -0.047 points (95% CI, -0.146 to $+0.056$). Nor did the correctly matched full pack outperform a mismatched profile (-0.034 ; 95% CI, -0.136 to $+0.067$). The 12-item core was essentially tied with no profile ($+0.007$; 95% CI, -0.097 to $+0.106$), while the full pack did not outperform the core (-0.054 ; 95% CI, -0.155 to $+0.050$). These comparisons provide no demonstrated advantage for the tested full, shortened or correctly matched profile in the two message tasks. In particular, correct matching did not yield the person-specific benefit that an individualized communication system ultimately needs to establish.

Removing some domains improved rather than reduced judged message quality. Relative to the corresponding domain-removed condition, the full pack scored 0.146 points lower when decisions and evidence was retained (95% CI, -0.234 to -0.061 ; Holm-adjusted $p = 0.012$), and 0.136 points lower when authenticity and rewriting was retained (95% CI, -0.233 to -0.039 ; adjusted $p = 0.045$). Removing reciprocal perspective showed a similar negative direction, though not a multiplicity-adjusted statistically clear result. Adverse differences were concentrated in the ambiguous interpersonal task, while the missed-deadline task showed much smaller effects. This task dependence is consistent with the possibility that applying additional analysis, challenge, alternatives or rewriting constraints can make a straightforward interpersonal message less suitable. It does not establish that the underlying preferences should no longer be elicited.

A leave-one-domain-out comparison measures the additional contribution of one domain while the other domains remain available. It cannot determine whether a domain contains useful information on its own, especially where presentation, uncertainty, interpretation and authenticity instructions overlap. The two tasks also covered only a narrow set of circumstances in which particular accommodations might be relevant. Finally, repeat-judgment reliability was below the study's prespecified adequacy point estimate (ICC = 0.596 against a requirement of 0.60). The domain study therefore identified a credible mapping and task-application problem, but did not meet its conditions for selecting a validated optimized questionnaire. Its

item matrix is a basis for further review, not a demonstrated final instrument.

[See Table VI]

D. Task-Routed Application of Accommodations

The next study tested a narrower explanation for the full pack's poor performance: perhaps giving the generator only a few instructions selected for the task would avoid unnecessary interventions. Twelve previously uninspected retained cases were evaluated on the same two tasks under ordinary personalization, task-routed CSA, the full source-linked pack and no personalization. Routing selected one to three accommodations intended to affect the final message. The study assessed both the quality of the resulting messages and whether the selected instructions had actually been expressed.

Task-routed CSA did not outperform ordinary personalization. The paired difference was -0.042 points (95% CI, -0.275 to $+0.192$; exact sign-flip $p = 0.75$); five cases favoured each approach and two tied. It was also essentially tied with the full instruction pack (-0.006) and showed no reliable advantage over no personalization ($+0.050$). With twelve cases, these estimates do not prove equivalence or absence of benefit. They do show that this specific routing method failed to demonstrate the intended quality improvement.

The accompanying delivery audit identified more direct limitations. Nine of 72 selected accommodations were not visibly expressed, leaving route expression at 87.5%. Six of 24 routed messages (25%) contained irrelevant or unauthorized personalization. Only 10 of 24 task messages were judged both the best and a distinguishable match for their intended route. Review also confirmed one invented fact and one instance in which the model substituted its decision for the individual's authority. These are not merely small differences on a composite rating scale; they identify concrete failures to carry permitted instructions into the message while preserving factual and personal boundaries. The pilot's manipulation and reliability gates failed, so it did not justify scaling the prompt-only route as a validated solution.

[See Table VII]

E. Interpretation of the Developmental Investigations

The later studies evaluated a different part of the proposed CSA method from that measured in the principal recommendation comparisons. The larger studies asked whether structured intake enabled a model to recommend better individualized support. The component and routing studies asked whether questionnaire-derived information could be translated into better finished communications. That second task depends on a sequence of decisions: an answer must be interpreted as a

permissible accommodation, translated into an instruction, selected when relevant and expressed correctly in the final message. Success at the first stages does not establish success at the later ones.

Taken together, the investigations narrow the possible explanations. Compression alone did not account for the difference between structured and ordinary personalization. Source-linked compilation demonstrated that the questionnaire could produce traceable candidate instructions, but neither the complete instruction pack nor the tested shorter core demonstrated improved task-level messages. The task-routing pilot then showed that selecting fewer instructions was insufficient: the generator still omitted selected accommodations and introduced unrelated or unauthorized changes. The failure of the full pack to beat a mismatched profile leaves open whether its content was genuinely person-specific in the tested tasks or whether broadly useful communication advice obscured individual differences. Overlap among domains, limited task coverage and judgment variability remain additional plausible explanations.

These findings support an evidence-derived requirement for further investigation of both interpretation governance and execution governance. Interpretation must retain the link from an individual's answer to a proposed accommodation, including clarification, contextual limits and the individual's authority to accept, modify or reject it. Execution must determine which authorized instructions apply to a task and verify that the resulting message expresses them without inventing facts, imposing unwanted intervention or substituting for the individual's decision. The present experiments identify these control problems; they do not establish that a particular governance architecture solves them. That engineering hypothesis, and the distinction between tested mechanisms and proposed safeguards, is examined in the discussion.

VII. Discussion

[See Figure 4]

A. What the Experimental Program Establishes

The principal finding is that structured communication intake can improve the information used to recommend individualized AI support, but this advantage did not reliably carry through to the production of ready-to-send messages. In the independent 60-case study, structured calibration exceeded ordinary personalization by 0.243 points on the weighted 0–4 suitability scale. The corrected 180-case questionnaire-only replication estimated a larger difference of 0.335 points (95% CI, 0.282 to 0.388), with structured calibration favoured in 148 cases. These results provide converging synthetic evidence that a defined intake can elicit information useful for recommending

accommodations beyond what a neutral open-ended question supplies. The estimate remains below the prespecified provisional 0.40-point design criterion, and one critical violation prevented the corrected study from meeting its full safety requirements. Neither result demonstrates that an individual would experience better communication with another person.

The later studies deliberately tested a different outcome. The questionnaire-derived profile was no longer the subject of a support recommendation; instead, it became a set of instructions intended to change the content and manner of two practical messages. This introduced additional decisions between knowing a preference and using it: whether a preference matters for the task, how it should be reconciled with other preferences, and whether the generated message expresses it without unnecessary changes. The full source-linked instruction pack failed to show a reliable advantage over no profile or a mismatched profile. This does not logically reverse the earlier recommendation findings, because recommendation suitability and task-level message quality are related but different measurements. It does establish that the demonstrated intake advantage cannot be treated as a demonstrated communication-outcome advantage.

B. Alternative Explanations of the Task-Level Results

Several explanations remain compatible with the domain-ablation results. First, the full instruction pack was not selective enough for each task. Domain removals sometimes improved message quality, especially in the ambiguous interpersonal task. Such improvements are consistent with the cost of applying instructions that are individually plausible but irrelevant, redundant, or in tension when combined. A removal contrast, however, measures a domain's incremental contribution in the presence of the remaining domains; it does not measure whether the information in that domain is inherently unnecessary. The six operational domains overlap in such functions as expressing uncertainty, preserving authenticity, asking for clarification, and limiting intervention. Removing one may leave comparable guidance active elsewhere.

Second, the two chosen tasks offered limited opportunities to exercise all 28 items. A preference concerning overload or rewriting authority can contribute little to an output if the task never activates the relevant condition. The observed differences between the interpersonal and missed-deadline tasks are consistent with this limitation. Third, the correctly matched profile did not outperform a mismatched profile. This raises the possibility that the messages were influenced more by generally helpful conventions—clarity, caution, respectful tone, and explicit options—than by person-specific instructions. It is also possible that genuine differences were obscured by the

generator's selection and execution behaviour. The study was not designed to distinguish those possibilities conclusively.

Fourth, measurement uncertainty matters. The domain study's repeated-judgment ICC of 0.596 narrowly missed its planned 0.60 reliability requirement, whereas the corrected recommendation study reported an overall ICC of 0.937. These values arise from different evaluation tasks and should not be interpreted as a controlled comparison of judge quality. Nevertheless, small domain-level contrasts are difficult to interpret confidently when the relevant instrument shows material scoring variation. The frozen rubric also defined behavioural anchors at scores 0, 2, and 4 while leaving 1 and 3 to judge interpolation. That limitation does not nullify the reported comparisons, but it makes the operational meaning of a small score difference less precise than the numerical result alone suggests.

C. Distinguishing Elicitation, Interpretation, Selection and Execution

The findings are more informative when the CSA process is separated into distinct engineering functions. Elicitation obtains a person's stated preferences and circumstances. Interpretation converts those statements into candidate accommodations while preserving the distinction between description and authorization. Selection identifies which authorized accommodations are applicable to a particular communication task. Execution expresses those accommodations in the completed message while respecting facts, context, and the person's control of their own words. A positive result at an earlier stage does not validate the subsequent stages.

Prior work on iterative language-model feedback, including the Self-Refine approach of Madaan and colleagues [13], establishes that generation and subsequent revision can be separated into technical stages. It does not demonstrate that self-feedback provides the source authorization, task-relevance checks or factual and autonomy safeguards proposed here. Those controls remain a CSA engineering hypothesis requiring direct tests.

The source-linked mapping audit demonstrated that the retained questionnaire answers could be associated with traceable candidate instructions; it did not establish that those instructions were necessary, beneficial, or followed. The task-routed pilot then tested whether supplying a smaller number of task-relevant accommodations improved actual message generation. It did not: routed CSA differed from ordinary personalization by -0.042 points (95% CI, -0.275 to 0.192). More importantly for explaining the mechanism, nine of 72 expected accommodations were not visibly expressed, six of 24 routed messages contained irrelevant or unauthorized personalization, and only ten of 24 messages were judged the best match for their intended route. One invented fact and one substitution for the person's decision-making authority were

independently confirmed. These observed failures establish that the tested prompt-only delivery mechanism did not consistently execute even its selected instructions. They do not establish how much of the broader domain-study result was caused by execution rather than by the instrument, task coverage, or evaluation design.

D. Governance as an Evidence-Derived Engineering Hypothesis

The research therefore supports investigating governance at two related boundaries. Interpretation governance would preserve the original answer, link any proposed accommodation to its source, distinguish explicit preferences from inferred needs, and require participant acceptance, modification, or rejection before a rule becomes authorized. Context-specific rules would not silently override general preferences, and conflicting instructions would be returned for clarification rather than combined by assumption. These safeguards follow from the instrument's design rationale, including its concern with autonomy, reciprocal understanding, and authenticity; they have not been evaluated through completed human calibration interviews.

Execution governance would apply after authorization. For each task, a selector would identify relevant accommodations, bound the number activated together, and define how priorities and conflicts are resolved. A separate check would then compare the generated message against the activated instructions, the original task, and explicit prohibitions. It would examine missing accommodations, irrelevant additions, unsupported facts, unrequested rewriting, and autonomy substitution. Any repair would be limited by prespecified rules, with an inspectable record linking answer, accommodation, instruction, task selection, and resulting output. Participants would retain the ability to review, revise, or withdraw their authorized preferences. These components describe a proposed architecture, not an implemented or experimentally verified capability of the current CSA system.

This proposal differs from adding further instructions to a prompt. The tested failure involved not only the content of instructions but also the absence of an independently checked boundary between generating a message and accepting it as compliant. An explicit verification step is therefore a falsifiable engineering response to observed omissions and unauthorized additions. The immediate experimental question is whether such a layer improves instruction expression, relevance, factual discipline, and respect for individual authority when compared with the existing prompt-only path. A favourable mechanism result would still not establish that recipients prefer the messages or that communication improves in real use.

E. Implications for Instrument Development and Evaluation

[See Figure 5]

The results do not support automatically discarding the domains whose removal improved certain messages, nor do they support treating the full 28-item form or proposed 12-item core as optimized. The ablation study did not meet its reliability and person-specificity preconditions for that decision. A defensible redesign requires examining what each item elicits, which response authorizes a behavioural change, whether its domain duplicates another function, and which tasks would make the accommodation useful. It also requires evaluating whether a shorter instrument could reduce burden while preserving fidelity, rather than assuming that more questions or more instructions will produce better support.

The appropriate progression is consequently developmental. Human review should determine whether questions and candidate accommodations are comprehensible, faithful, correctable, and acceptable. A bounded synthetic proof-of-mechanism should compare prompt-only execution with a frozen governed execution path using the same tasks and cases, placing delivery and safety checks before average quality improvement. After the instrument and execution process are revised and frozen, fresh human and synthetic comparisons can assess practical usefulness without reusing optimization cases as independent confirmation. Section VIII considers the limitations and requirements of that future work.

VIII. Limitations and Future Research

A. Limits of Synthetic Evaluation

The present studies evaluated artificial cases, model-generated recommendations and messages, and judgments made by language models. This design made it possible to compare conditions using the same stated preferences and tasks, but it did not measure how actual individuals interpret, experience, or use CSA assistance. A recommendation that closely reflects a synthetic profile may still be difficult to understand, burdensome to apply, or unwelcome to the person it is meant to assist. Neither the positive recommendation results nor the later task-level findings establish real-world changes in communication, relationships, decision quality, or individual autonomy.

The synthetic cases were constructed within a controlled research process rather than sampled from a population of human users. Their apparent diversity cannot establish representativeness across neurotypes, communication contexts, cultures, languages, or circumstances involving substantial stress. Model-generated cases and model-based judges may also share assumptions about what makes communication

appropriate. Blinding and computational reproduction strengthen the internal audit of the comparisons, but cannot remove these limitations or substitute for independent human assessment.

B. Measurement, Intervention and Task Coverage

The rubric assessed research justification, preference fidelity, contextual appropriateness, autonomy and authenticity, and conflict treatment. Its weights and anchors at scores 0, 2 and 4 were specified, but scores 1 and 3 required judges to interpolate between adjacent descriptions. This leaves some discretion in intermediate ratings. The later domain-ablation study also had lower repeat-judgment reliability than the corrected 180-case comparison, limiting confidence in small differences among instruction conditions. The provisional +0.40 threshold was set before confirmatory outcomes but was not validated as a minimum difference meaningful to human users. Consequently, the observed effect estimates and uncertainty intervals should be distinguished from both the study decision rule and any claim of practical benefit.

The studies did not isolate every component of the CSA process. The corrected large-scale comparison measured the suitability of recommendations about support; the domain-ablation and routing studies measured ready-to-send messages under more demanding operational conditions. The two communication tasks in those later studies provided limited opportunities to activate all six domains. Functional overlap between domains further complicates interpretation of leave-one-domain-out differences: a small marginal contribution does not prove that the underlying preference has no value, and a harmful full-pack contrast does not by itself identify whether the question, its mapping, its selection, or its execution caused the result. The absence of a demonstrated advantage for a correctly matched profile remains an important competing explanation that requires direct testing.

The work also evolved through methodological corrections and later diagnostic studies. The stopped execution yielded no comparative outcome. The original 180-case study measured an augmented workflow that included information beyond the intended questionnaire boundary and cannot be combined with the corrected questionnaire-only result as if the interventions were identical. The compression, mapping, ablation, and routing investigations were motivated by prior findings and served developmental rather than final confirmatory purposes. These distinctions constrain the scope of any pooled or general conclusion.

C. Testing the Proposed Governance Mechanism

The next engineering question is whether explicit controls can improve the faithful use of accommodations that have already

been elicited and authorized. A bounded proof-of-mechanism should compare the retained prompt-only task-routed path with a separately frozen governed path on unchanged development cases and tasks. The proposed path would retain the source of each accommodation, determine its relevance to the task, resolve conflicts and priority, check whether selected instructions were expressed, reject unsupported factual additions and unauthorized intervention, and permit only a prespecified correction before output acceptance. The primary measures should be route expression, irrelevant or unauthorized additions, factual and autonomy violations, reliability, and inspectability. Message quality should remain an important but separate outcome. Because this governance layer has not been implemented or tested in the reported research, improved performance cannot be assumed.

A successful mechanism test would not by itself validate the questionnaire. Item-level review must examine whether each question elicits a distinct, comprehensible preference, whether the mapped accommodation follows from the answer, whether its use imposes burden or unwanted inference, and whether it applies to tasks that genuinely engage the intended construct. Domain-level findings should guide this review, not automatically dictate deletion or retention. Any revised questionnaire, mapping, selector, or execution process must be frozen as a new intervention before a later independent comparison. Development cases used to select changes cannot serve simultaneously as unbiased confirmation of those changes.

D. Human Evaluation and Independent Confirmation

Human evaluation is required to determine whether people understand the questions, recognize their own preferences in the proposed accommodations, and can readily accept, modify, reject, or withdraw them. Controlled prototype studies should assess completion burden, accessibility, perceived fit, preservation of voice, interpretation accuracy, unwanted intervention, and the usefulness of resulting communications in realistic circumstances. Participants should retain authority over their original responses and every material instruction derived from them. Such research must distinguish the individual's judgment of usefulness from an external assessor's preference for polished or conventionally agreeable wording.

After the interpretation and execution mechanisms have been revised and developmentally evaluated, a fresh held-out comparison should test the resulting frozen CSA system against independently constructed ordinary personalization and an appropriate no-profile condition. Tasks should cover ambiguity, decision support, overload, authenticity, conflict, and reciprocal misunderstanding with enough breadth to evaluate whether particular accommodations help when their relevant circumstances arise. Outcomes should include task-level message quality, person-specific fit, safety, perceived

usefulness and correction burden, with decisions defined before outcome inspection. Only subsequent studies with people can establish whether improvements in judged messages translate into meaningful communication outcomes.

IX. Conclusion

This research developed an individualized communication-calibration instrument and evaluated its use across successive controlled synthetic studies. The exploratory comparison provided a reason to pursue structured intake, and the independent 60-case study found that its advantage persisted when ordinary personalization was constructed separately. In the corrected 180-case questionnaire-only replication, structured calibration produced more suitable support recommendations than ordinary personalization, with a mean difference of +0.335 points on the weighted 0–4 scale. That estimate was statistically supported, but it did not satisfy the provisional +0.40 confidence-bound criterion, and one verified critical violation prevented a complete safety success decision.

The later investigations changed the engineering question. Source-linked mapping established that responses could be traced to proposed accommodations; it did not establish that those accommodations would be selected or expressed correctly. On two ready-to-send communication tasks, the complete instruction pack did not demonstrate person-specific benefit, and the subsequent routing pilot identified omitted accommodations, irrelevant or unauthorized interventions, an invented fact, and an autonomy substitution. The recommendation and message studies therefore tested related but different propositions. Together, they support the value of continued investigation of structured intake while showing that the tested prompt-only delivery paths were not reliable implementations of individualized communication support.

The contribution is consequently both methodological and architectural. The work separates the elicitation of preferences from their interpretation, authorization, task-specific selection, execution, and verification. Its findings provide an empirical reason to investigate governance across those boundaries, while leaving the effectiveness of the proposed governed process unproven. The appropriate next stage is controlled mechanism testing and human evaluation, followed by independent confirmation of any revised system. The present evidence supports that research program; it does not establish that the current questionnaire or a future governance layer improves real-world communication.

Research Status and Evidence Basis

This manuscript reports a completed program of synthetic experiments, developmental audits and subsequent diagnostic studies. Experimentally measured findings include the 12-persona exploratory comparison, independent 60-case

comparison, original augmented-workflow 180-case study, corrected questionnaire-only 180-case replication, compression investigations, 48-case domain-ablation study, and 12-case task-routing pilot. The mapping audit provides evidence of source traceability and deterministic compilation, not human authorization or item-level efficacy. The stopped clean-replication execution is recorded as a methodological event, not as outcome evidence. The later component and routing studies are developmental and should not be treated as independent confirmation of an optimized intervention.

The proposed interpretation and execution-governance layer remains an engineering hypothesis, not an implemented or experimentally measured capability. All quantitative conclusions concern the synthetic cases, conditions, tasks and evaluation criteria used in the recorded studies. Human usefulness, real-world effectiveness, optimal questionnaire length, item-level utility, and the benefit of governed execution remain unresolved. Supporting repository protocols, frozen instruments, judgments, analysis records and correction logs provide the principal evidence basis; selected external literature supplies conceptual background rather than proof of CSA effectiveness.

Disclosure of Generative AI Use

Generative AI systems were used in the research workflow for synthetic case construction, communication-support generation, scoring and judgment, diagnostic investigations, and assistance with code, analysis and manuscript preparation. Research protocols specified model roles, information boundaries, judgment criteria and decision rules; retained outputs were subjected to validation, audit and computational reproduction where reported. The author directed the research questions, study design and revisions, assessed the interpretation of findings, and retains responsibility for the manuscript and its claims. AI-generated cases and model judgments were not represented as human participants, clinical validation or independent peer review.

Publication Information

Steven Boyle, Northline Advisory, Ontario, Canada. ORCID: 0009-0006-1607-7362. Research Paper, Version 1.0, published 19 October 2026. Supplementary instrument, mapping crosswalk and reproducibility materials are supplied or linked where available with the published version.

References

[1] H. Giles, D. M. Taylor, and R. Y. Bourhis, “Towards a theory of interpersonal accommodation through language:

Some Canadian data,” *Language in Society*, vol. 2, no. 2, pp. 177–192, 1973.

[2] C. Nicolaidis et al., “The development and evaluation of an online healthcare toolkit for autistic adults and their primary care providers,” *J. Gen. Intern. Med.*, vol. 31, no. 10, pp. 1180–1189, 2016. doi: 10.1007/s11606-016-3763-6.

[3] D. E. M. Milton, “On the ontological status of autism: The double empathy problem,” *Disability & Society*, vol. 27, no. 6, pp. 883–887, 2012. doi: 10.1080/09687599.2012.710008.

[4] C. J. Crompton et al., “Autistic peer-to-peer information transfer is highly effective,” *Autism*, vol. 24, no. 7, pp. 1704–1712, 2020. doi: 10.1177/1362361320919286.

[5] L. Hull et al., “Putting on my best normal: Social camouflaging in adults with autism spectrum conditions,” *J. Autism Dev. Disord.*, vol. 47, pp. 2519–2534, 2017. doi: 10.1007/s10803-017-3166-5.

[6] M. Doherty, S. McCowan, and S. C. K. Shaw, “Autistic SPACE: A novel framework for meeting the needs of autistic people in healthcare settings,” *Br. J. Hosp. Med.*, vol. 84, no. 4, pp. 1–9, 2023. doi: 10.12968/hmed.2023.0006.

[7] E. L. Deci and R. M. Ryan, “The ‘what’ and ‘why’ of goal pursuits: Human needs and the self-determination of behavior,” *Psychological Inquiry*, vol. 11, no. 4, pp. 227–268, 2000.

[8] S. Amershi et al., “Guidelines for human-AI interaction,” *Proc. CHI 2019*, 2019. doi: 10.1145/3290605.3300233.

[9] E. Tabassi, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023. doi: 10.6028/NIST.AI.100-1.

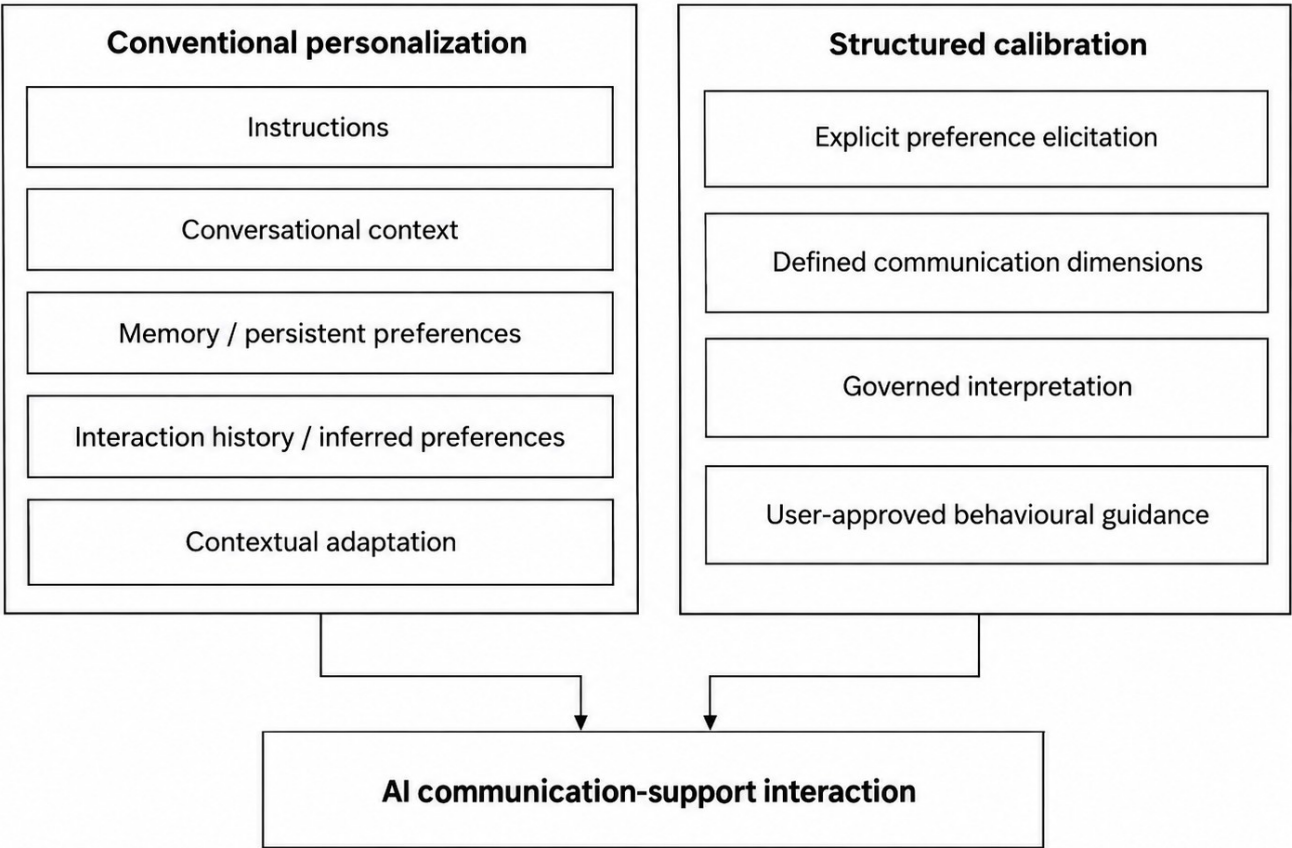
[10] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani, “LaMP: When large language models meet personalization,” in *Proc. 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 7370–7392. doi: 10.18653/v1/2024.acl-long.399.

[11] L. Zheng et al., “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023. arXiv:2306.05685.

[12] S. Boyle, “CSA Research Program: Why Each Study Was Run, What It Found, and How the Work Should Be Presented,” Northline Advisory, completed research-program synthesis, Oct. 16, 2026. Supporting materials: northlinedev1/northline-site, docs/research/individual-communication-support/, versioned research branches and experiment records.

[13] A. Madaan et al., “Self-Refine: Iterative refinement with self-feedback,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

Figures and Tables



Complementary approaches; structured calibration does not replace ordinary personalization

Figure 1. Conventional personalization and structured calibration are complementary approaches to AI communication support. Structured calibration adds explicit preference elicitation, defined dimensions, bounded interpretation and intended user authorization; it does not replace ordinary conversational personalization.

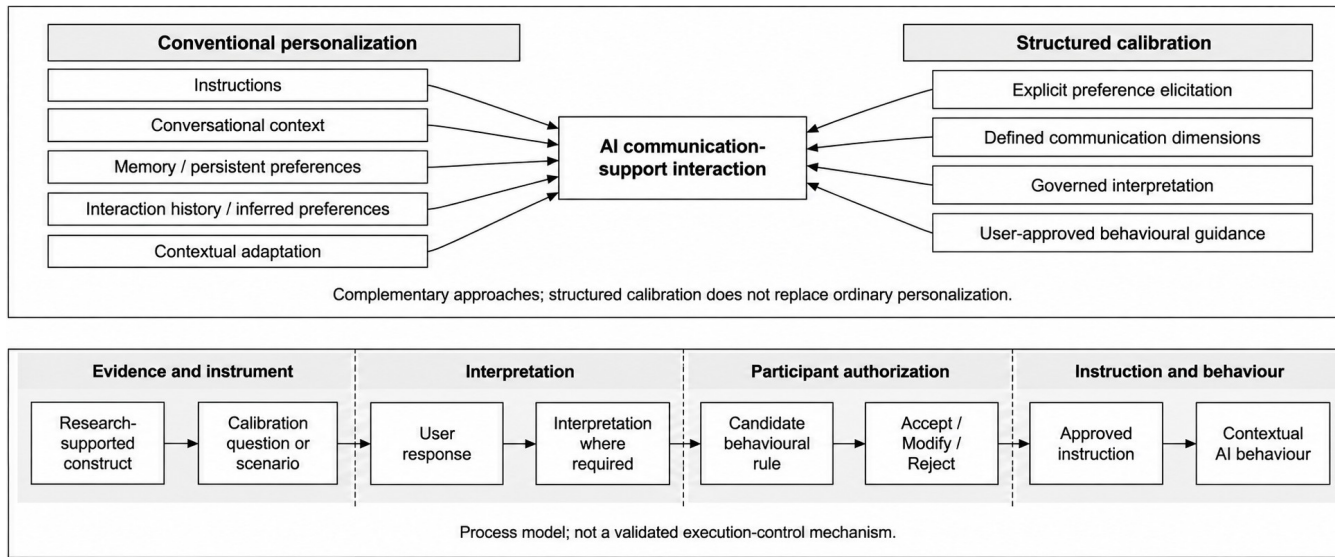


Figure 2. Proposed path from research-supported construct to questionnaire item, response, provisional interpretation, candidate behavioural rule and participant authorization. The depicted review and execution sequence is a process model, not a validated implementation.

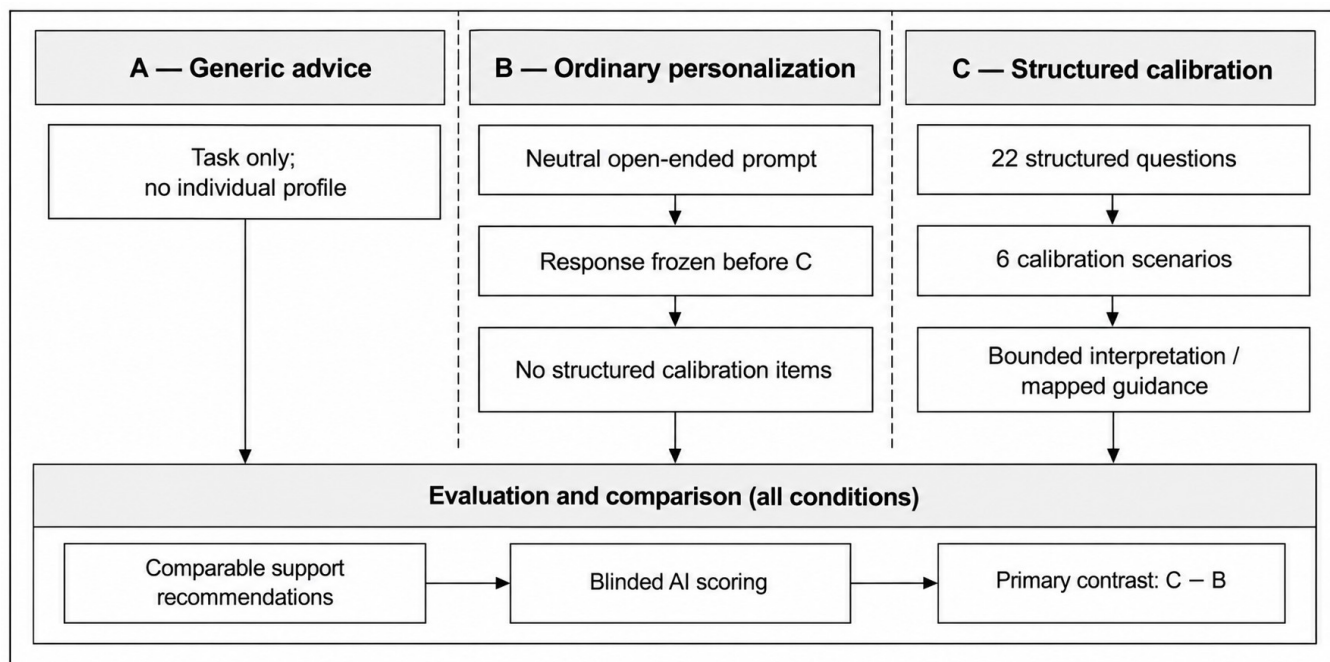


Figure 3. Principal comparison conditions: A received only the task, B received independently elicited ordinary preferences frozen before C, and C received questionnaire-derived information from 22 questions and six scenarios. Outputs were evaluated through blinded judgments, with C minus B the primary contrast.

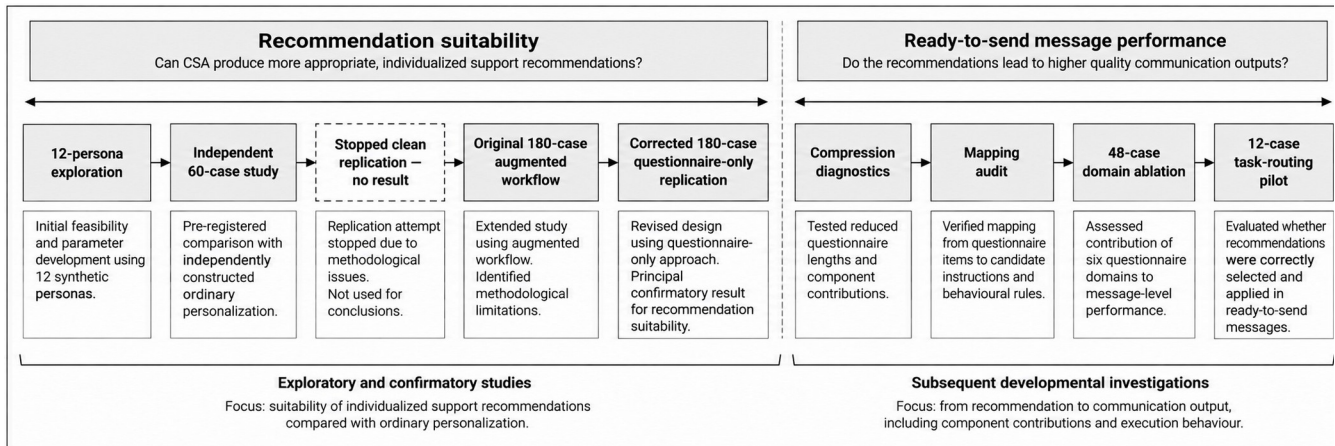


Figure 4. Chronological sequence of comparative recommendation studies and later developmental message-output investigations. The studies address related but distinct endpoints and are not pooled.

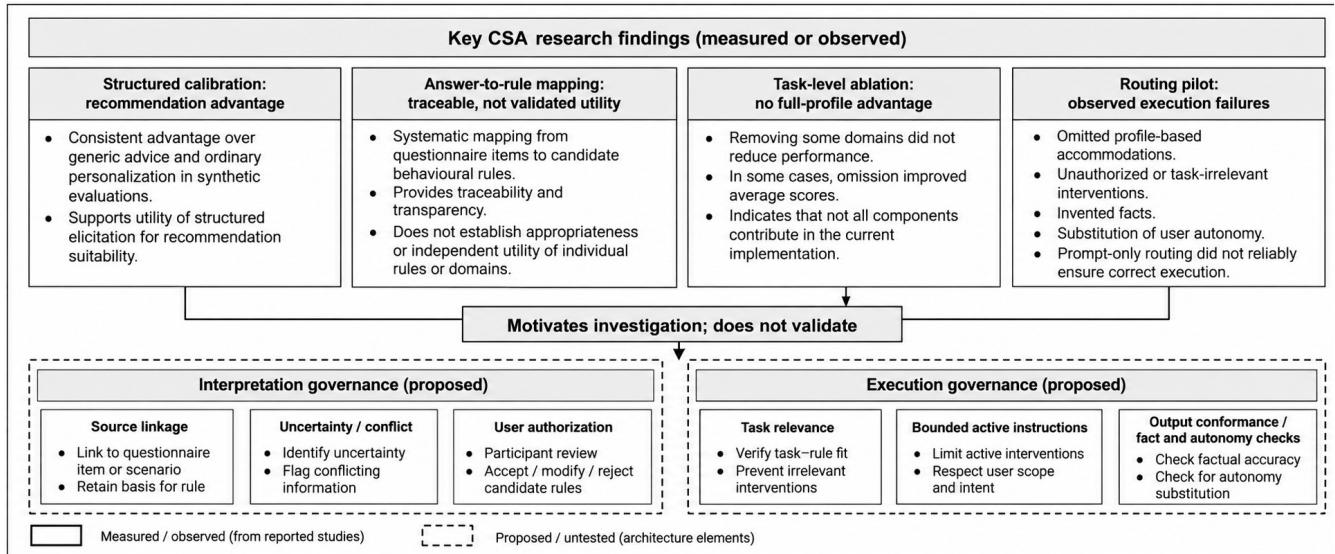


Figure 5. Measured recommendation, mapping, ablation and routing findings motivate proposed interpretation and execution governance. The proposed controls were not implemented or tested in the reported experiments.

Table I — Structure of the CSA Calibration Instrument

Component	Items	Purpose
Structured questions	Q1–Q22	Explicit communication-support preferences
Situational scenarios	S1–S6	Contextual preferences and ambiguity
Total	28 items	22 structured + 6 scenarios

Table I. Structure of the CSA calibration instrument: 22 structured questions (Q1–Q22) and six situational scenarios (S1–S6), comprising 28 inputs.

Table II — Operational Ablation Domains and Questionnaire Item Membership

Operational domain	Included items
Presentation and processing	Q1, Q2, Q5, Q6
Decisions and evidence	Q7, Q8, Q9, Q10, Q17, Q18, S3
State and context	Q11, Q12, Q13, Q14, S5
Interpretation and uncertainty	Q4, Q15, S1
Authenticity and rewriting	Q19, Q20, Q21, Q22, S2, S4
Reciprocal perspective	Q3, Q16, S6
28 items allocated once; operational groupings, not validated factors.	

Table II. Prespecified operational instruction domains used in the developmental ablation study and their questionnaire/scenario membership. These groupings are functional testing units, not validated psychological factors.

Table III — Evaluation Rubric Dimensions, Weights, and Behavioural Anchors

Dimension	Weight	Anchor specification
Research justification	25%	0, 2, 4 explicitly anchored
Preference fidelity	25%	0, 2, 4 explicitly anchored
Contextual appropriateness	20%	0, 2, 4 explicitly anchored
Autonomy and authenticity	20%	0, 2, 4 explicitly anchored
Conflict treatment	10%	0, 2, 4 explicitly anchored
Scores 1 and 3: judge interpolation between adjacent anchors. Scale: 0–4.		

Table III. Five dimensions of the frozen weighted scoring rubric. Behavioural descriptions were provided at scores 0, 2 and 4 on the 0–4 scale; judges interpolated scores 1 and 3.

Table IV — Summary of Principal Experimental Results Across the CSA Research Program

Study	Endpoint	C – B	95% CI	Interpretation
12-persona exploration	Support recommendations	+0.571	+0.346 to +0.781	Exploratory; B not independent
Independent 60-case study	Support recommendations	+0.243	+0.143 to +0.342	Independently constructed B
Original 180-case workflow	Augmented support recommendations	+0.589	+0.528 to +0.653	Supplemental Profile Builder
Corrected 180-case replication	Questionnaire-only support recommendations	+0.335	+0.282 to +0.388	+0.40 practical gate not met
Note: Developmental message-output contrasts appear in Table VI/VII, not this main confirmatory comparison.				

Table IV. Principal synthetic comparative evidence for recommendation suitability. Experimental conditions and information content differ across studies; values are not pooled.

Table V — Corrected 180-Case Confirmatory Comparison: Principal Result Summary

Measure	Finding
Design	
Eligible / analysed cases	252 eligible / 180 analysed
Recommendation outputs	540
Blinded judgments	1,080
Outcome	
Mean A	1.113
Mean B	3.059
Mean C	3.394
C – B	+0.335
95% CI (C – B)	+0.282 to +0.388
Cases favouring C	148 / 180 (82.2%)
Decision gates	
Provisional practical threshold	+0.40 — not met
Critical violation gate	One serious C violation — not met

Table V. Corrected 180-case questionnaire-only synthetic comparison. A statistically positive C minus B contrast did not satisfy the prespecified provisional practical-magnitude or zero-critical-violation gates.

Table VI — Domain-Ablation Findings for Task-Level Communication Outputs

Contrast Difference = first minus second	Mean difference	95% CI	Inference
Full – no profile	–0.047	–0.146 to +0.056	No advantage demonstrated
Full – mismatched profile	–0.034	–0.136 to +0.067	No advantage demonstrated
Core – no profile	+0.007	–0.097 to +0.106	No advantage demonstrated
Full – core	–0.054	–0.155 to +0.050	No advantage demonstrated
Full – drop decisions/evidence	–0.146	–0.234 to –0.061	Removal performed better; Holm-adjusted p = .012
Full – drop authenticity/rewriting	–0.136	–0.233 to –0.039	Removal performed better; Holm-adjusted p = .045
48 cases; 2 tasks/case; 10 conditions; 960 messages.			

Table VI. Task-level domain-ablation contrasts. Negative full-minus-dropout differences indicate that the corresponding dropout scored higher in the tested implementation; this does not establish intrinsic domain value.

Table VII — Task-Routing Pilot Outcomes and Execution Failures

Measure	Observed result	Interpretation
Routed – ordinary personalization	–0.042 (95% CI –0.275 to +0.192)	No advantage demonstrated
Routed – full	–0.006	No advantage demonstrated
Routed – no personalization	+0.050	No advantage demonstrated
Selected accommodations visibly expressed	63 / 72 (87.5%)	9 not expressed
Unauthorized / irrelevant personalization	6 / 24 messages (25%)	Output control failure
Best route match	10 / 24 messages	Selected route frequently not best
Invented fact	1 observed	Factual failure
Autonomy substitution	1 observed	User-agency failure
12 synthetic cases; 96 ready-to-send messages.		

Table VII. Task-routing pilot findings. Routing did not improve ready-to-send message quality and exhibited omission, task relevance, factual and autonomy errors.