TUX: Measuring Human–AI Tacit Understanding

Anonymous ACL submission

Abstract

As large language models (LLMs) increasingly act as collaborative partners, human–AI alignment is often evaluated through explicit task success, accuracy, or reward optimization. Yet many collaborative settings depend on tacit understanding: whether an agent can align with a human's evaluative stance or representational priors without clear objectives, communication, or feedback. To study this capacity, we develop a spectrum-placement task inspired by the social party game Wavelength, in which humans and agents independently place concepts along subjective spectra. We operationalize the Tacit Understanding Index (TUX) as a pairwise measure of similarity between human and agent judgments, and evaluate it with 241 human participants and 200 profile-conditioned LLM agents across four models. We find that nearest human–agent pairs in trait space achieve significantly higher TUX, suggesting that tacit alignment is structured by person-level characteristics rather than random similarity. Regression analyses show that TUX becomes more explainable as predictor sets become richer, with individual traits, decision-making styles, and confidence improving over aggregate trait-distance baselines. These findings suggest that tacit understanding between humans and LLMs is measurable, while revealing the limits of profile-based conditioning for capturing deeper representational alignment.

1 Introduction

As AI systems shift from tools to collaborative partners, a central challenge is not only whether AI can complete tasks, but whether humans and AI systems can become meaningfully aligned in shared situations. Yet alignment in human–AI systems is commonly operationalized through explicit, goal-directed collaboration: agents are considered aligned when they jointly optimize a shared objective, complete a task efficiently, or maximize an explicit reward function (Schmutz et al., 2024; Fragiadakis et al., 2024). That is, alignment is inferred through observable outcomes such as task accuracy, negotiation efficiency, or cooperative payoff (Sidji et al., 2024; Schmutz et al., 2024; Senoner et al., 2024). While these measures are valuable, they capture only a limited form of collaboration.

Importantly, many forms of collaboration depend on something less visible: whether collaborations develop a tacit sense of each other's perspective. Theories of *common ground and grounding* argue that communication depends on the accumulation of shared assumptions about what participants know, believe, and take for granted (Clark and Brennan, 1991). In the case of human–human collaborations, people routinely rely on mutual understanding, shared expectations, and implicit calibration to coordinate action and make sense of each other's judgments. Research on *shared reality* suggests that people often seek not only agreement over external facts, but also a sense that another person shares their internal orientation toward the world—their interpretations, feelings, and evaluative stance (Baek and Parkinson, 2022). Relatedly, focal-point reasoning shows that people may coordinate around mutually salient interpretations even without direct communication (Schelling, 1960). Together, these theories suggest that collaboration is not reducible to explicit instruction-following or task completion; it also involves representational alignment between minds.

However, human–AI alignment has not yet fully incorporated the above tacit dimension. In fact, less is known about whether AI systems can approximate the implicit representational priors that make another agent's judgments feel understandable or socially attuned. This gap matters because many collaborative settings are not organized around a specific objective. In creative brainstorming, aesthetic judgment, interpersonal advising, and emotional companionship, a collaborator may need to infer what someone means by phrases such as "this feels too cold," "this seems trustworthy," or "this does not fit my style." In these settings, success de-

pends not only on producing a useful response, but also on whether the collaborator can recognize and align with the latent frames of reference through which the other person interprets the situation.

This gap in human–AI collaborations has practical and theoretical consequences. Practically, an AI may appear successful because they produce fluent or high-performing outputs, while still failing to understand the user's implicit framing, values, or expectations. Theoretically, the absence of a tacit-understanding construct limits our ability to compare human–human and human–AI collaborations. Without a way to measure tacit alignment, we cannot distinguish between agents that merely optimize visible outcomes and agents that exhibit deeper representational calibration with human collaborators. Moreover, little empirical work examines how individual human traits shape alignment capacity, or whether LLMs exhibit stable behavioral signatures that influence tacit calibration (Deng et al., 2024; Park et al., 2026).

In this paper, we ask whether tacit understanding between humans and AI can be measured as a form of representational calibration. Building on theories of common ground, shared reality, and individual-difference-based coordination (Clark and Brennan, 1991; Baek and Parkinson, 2022; Schelling, 1960; John and Srivastava, 1999; Frederick, 2005), we test the hypothesis (H): **human–agent pairs with more similar trait-level descriptors would exhibit higher tacit understanding.**

To test this hypothesis, we use a controlled spectrum-placement task inspired by the social party game *Wavelength* as a measurement setting for eliciting tacit representational judgments, where participants—both humans and profile-conditioned LLM agents—independently place cues along conceptual spectra.The task removes explicit shared goals, direct communication, and performance feedback, allowing us to estimate several pair-level tacit understandings through participants' placements. Accordingly, our work is guided by two research questions (RQs):

RQ1: How can tacit understanding be measured as a pair-level representational alignment construct in a weakly objective human–AI setting?

RQ2: How are personality and decision-making traits related to human–AI tacit understanding?

To address the above RQs, we conduct a user study with N=241 human participants and a bank of 200 profile-conditioned LLM agents instantiated across four model families (`Qwen3-14B`, `Qwen2.5-7B-Instruct`, `Mistral-Nemo`, and `Llama-3.1-8B`). Humans and agents complete the same spectrum-placement task under matched instructions, allowing us to compare independently generated judgments under weakly objective conditions. We compute pair-level tacit understanding using the Tacit Understanding Index (TUX), which measures similarity between human and agent placements across shared headers. We conduct regression modeling to examine how richer person-level descriptors—including personality traits, decision-making styles, and confidence—explain variation in TUX.

We find that nearest human–agent matches in trait space achieve significantly higher TUX than random matches (Cohen's d=0.29), supporting TUX as a measurable operationalization of tacit understanding. Further, TUX is increasingly explainable as predictor sets become richer: models incorporating individual traits, decision-making styles, and confidence outperform a single aggregate trait-distance baseline. Finally, coefficient analyses suggest that tacit understanding is shaped disproportionately by confidence in decisions and a subset of social and cognitive predictors. These findings position tacit understanding as a measurable but partial form of human–AI representational alignment. Our work makes the following contributions:

Operationalizing tacit understanding through TUX, a pair-level measure of tacit understanding as representational calibration between independent human and agent judgments.

Trait–Behavior modeling of tacit understanding by modeling pre-task trait with spectrum-placement behavior to examine how individual differences explain variation in tacit understanding.

Human–Agent data Bank for tacit understanding, which we construct and release[1], comprising de-identified human traits and spectrum-placement responses, alongside profile-conditioned LLM agents, persona profiles, and agent responses, to support future research.

2 Related Work

Tacit Understanding and Shared Representation in Collaboration. Research in social science and coordination theory has long emphasized that successful collaboration often depends on tacit understanding (Baek and Parkinson, 2022). In classic accounts of tacit coordination, participants converge on mutually expected behaviors without negotia-

[1]The data bank will be publicly released upon acceptance.

tion by relying on salient focal points (Schelling, 1960). A central mechanism underlying tacit coordination is common ground: the shared knowledge, beliefs, and assumptions that interlocutors treat as mutually recognized (Clark and Brennan, 1991). Common ground enables collaborators to interpret utterances, predict actions, and update expectations efficiently. These mechanisms explain how coordination can emerge even when communication is minimal Further, theories of shared reality and intersubjectivity argue that individuals are motivated to align internal representations with others for mutual understanding (Higgins et al., 2021; Ehkirch and Matsumae, 2024). Such alignment may be grounded in shared cultural background, common experiences, social values, or worldview (Kahan et al., 2011; Wang and Hu, 2021). Prior work characterized tacit knowledge in graphic design as experience-based, context-sensitive, and often embedded in situated design actions (Son et al., 2024). Building on the above body of research, our work refers to *tacit understanding* as the ability of humans and agents to align with one another's latent representational priors without communication, shared reward, or feedback.

Persona Conditioning in LLMs. As LLMs are increasingly deployed as social and collaborative agents, substantial effort has been devoted to conditioning models on persona or profile information to induce behavioral characteristics (Chen et al., 2024; Shimgekar et al., 2026; Kim et al., 2026). A common approach is direct persona prompting, in which the model is instructed to "act as" a particular character or adopt specific traits (Tseng et al., 2024; Chen et al., 2024). However, prior work suggests that such shifts may not reliably translate into consistent behavioral changes across tasks (Mannekote et al., 2025; Zou et al., 2025). Other approaches attempt to induce persona shifts by manipulating internal model representations, such as through vector steering or activation editing (Turner et al., 2023; Qiu et al., 2024), or by fine-tuning models on persona-specific data (Woźniak et al., 2024). Recent generative-agent work offers an alternative: structured contextual grounding through persistent memory and profile construction. These systems elicit richer characteristics through extended interaction and maintain them through memory architectures that retrieve task-relevant contextual fragments during inference (Park et al., 2023, 2026). Recent work proposed a knowledge ecosystem in which users create, share, and inject knowledge modules into LLM interactions (Zhao et al., 2025), and how a chatbot mediates organizational memory by retrieving knowledge embedded in communication practices (Lee et al., 2026). Further, personality and cognitive traits influence how people perceive social situations, anticipate others, and coordinate behavior (John and Srivastava, 1999; Frederick, 2005; Deng et al., 2024; Wang and Hu, 2021). However, in the case of human–AI collaborations, most methods optimize toward population-level preference consistency, as in reinforcement learning from human feedback, direct preference optimization, and related preference-based training paradigms (Christiano et al., 2017; Rafailov et al., 2023; Liu et al., 2025). This paper combines these two perspectives. We construct persona-conditioned LLM agents with reusable profile information, and we pair this with human trait profiling to examine whether person-level descriptors explain tacit understanding. We evaluate whether profile-conditioned agents can participate in a non-goal-oriented calibration task where the outcome is representational alignment.

Human–AI Collaboration Benchmarks. Research has introduced a variety of benchmarks for studying human–AI cooperation. CoBlock (Wu et al., 2024) is a multi-agent 3D blocks-world platform where human and LLM agents manipulate objects and communicate to achieve cooperative goals. The Constrained Human–AI Cooperation challenge (Du et al., 2024) evaluates embodied social intelligence, requiring an AI helper to infer a physically constrained human's intent and limitations while completing everyday tasks. Cooperative game settings have also been adapted: Codenames has been used as a testbed where an LLM codemaster gives clues to guide a human guesser under partial information (Sidji et al., 2024), multi-agent suites such as SPIN-Bench (Yao et al., 2025) and Concordia (Smith et al., 2026) evaluate planning, negotiation, social intelligence, trust, and mixed-motive interaction. Relatedly, human-AI interactions and alignment are explained by the mutual theory of mind (Wang et al., 2021) and mental models (Bansal et al., 2019). Beyond benchmark settings, recent work has also examined LLM-based collaborators for workplace settings (Das Swain et al., 2025; Mysore et al., 2025). While these benchmarks are valuable for testing social reasoning under interactive and strategically complex conditions, they typically retain explicit objectives, communication channels, shared rewards, or feed-

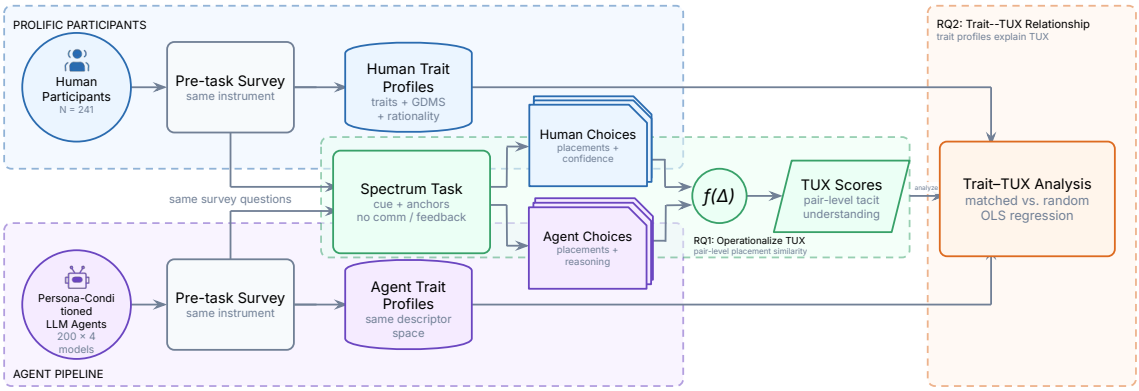

Figure 1: A schematic figure of our study design: Human participants and profile-conditioned LLM agents complete the same pre-task survey instrument and the same spectrum-placement task. Survey responses define human and agent trait profiles, while task responses define human and agent choices.

back loops that scaffold coordination. As a result, successful performance may reflect task-specific heuristics, following conversational signals, or optimizing a visible payoff rather than aligning with a partner's latent internal framing. Our work complements existing benchmarks by minimizing explicit scaffolding and foregrounding partner diversity. Our setting asks whether humans and agents can independently produce similar representations based on their underlying traits. This shifts the evaluation target from overt task success to tacit representational fit, providing a controlled setting for measuring when and how human–AI tacit understanding emerges.

3 Study Design and Methods

Figure 1 summarizes our study design. Human participants and profile-conditioned LLM agents complete the same pre-task survey instrument and spectrum-placement task. Survey responses define human and agent trait profiles, while task responses define human and agent choices. Pairwise placement similarity measures TUX, which is then analyzed against trait profiles through matched-versus-random comparison and regression.

3.1 Spectrum-Placement Task

Wavelength Game. Our task is inspired by the social party game *Wavelength*, a cooperative guessing game in which players coordinate around meanings along a conceptual spectrum (Morrison et al., 2025). In the original game, players are shown a conceptual spectrum defined by two opposing anchors, such as "hot–cold," "formal–casual," or "safe–risky." One player serves as the cue-giver and privately sees a hidden target location somewhere along the spectrum. The cue-giver then provides a verbal clue intended to help the other players infer the target placement, without directly naming a number or placement. For example, if the spectrum is "hot–cold" and the hidden target is closer to the "hot" side, the cue-giver might say "fresh coffee." The remaining players act as guessers: they interpret the clue and place a pointer on the spectrum where they believe the hidden target is located. Success is scored by how close their placement is to the hidden target. The game, therefore, rewards not only descriptive precision, but also shared conceptual framing: players must have similar assumptions about how examples map into a spectrum.

In our study, to operationalize tacit alignment without explicit communication or feedback, we adapt this structure into a symmetric single-response task, implemented as the interface shown in Figure 2. Each participant completes an independent session as a "guesser." Participants receive spectrum-cue headers that are predefined by the researchers. They are told that the target location has been selected by a like-minded peer, and their task is to estimate where they believe that peer would place the pointer along the spectrum.

This design removes asymmetry, overt coordination, and direct feedback. Participants respond independently and never see partner responses during the task. The perceived task frame remains collaborative—"align with a similar mind"—but without outcome feedback, making the task a controlled probe of tacit representational calibration.

Spectrum Headers. Our task consists of a number of *spectrum headers*. Each header defines a conceptual continuum through two anchors (e.g., *Subjective–Objective*) and a cue (e.g., *Beautifulness*).

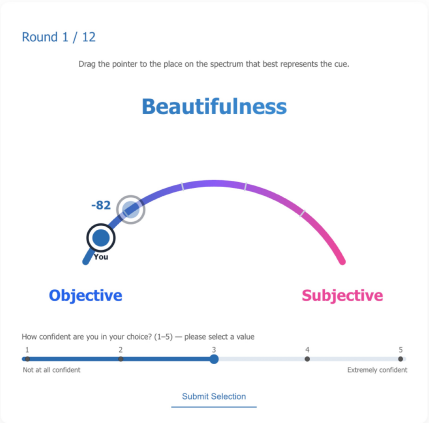

Figure 2: Interface of the spectrum placement task.

Cue	Left Anchor	Right Anchor
A coworker takes credit for your idea	Confront them directly	Let it go to avoid tension
A large party	Leave early and rest	Stay late and meet people
Planning a weekend trip	Detailed itinerary	Go with the flow
Quitting stable job for a new startup	Uncomfortable	Excited
Bungee jumping	Stressful	Thrilling
Beach	Calm	Crowded

Table 1: Example spectrum headers used in our task. These six headers were used as the fixed core set for all the participants. Each header combines a cue with a pair of conceptual anchors, designed to preserve semantic coherence while allowing interpretive variation.

We treat these headers as measurement items: each header must be coherent enough that participants understand the scale, but open-ended for capturing differences in priors and values.

Our goal in constructing the headers was to elicit structured variation in interpretation. We manually curated anchor pairs and cue items according to three criteria: 1) **Coherent anchor contrast:** the two endpoints define an interpretable conceptual continuum; 2) **Interpretive openness:** the cue admits multiple plausible placements rather than a single obvious answer; and 3) **Domain breadth:** the full header set spans semantic, affective, social, and judgment-oriented domains.

Under this design, individuals with similar priors or perspectives would likely tend to produce closer placements, while those with divergent priors may locate the same cue differently. This makes header responses useful behavioral traces for estimating tacit understanding through pair-level calibration. Table 1 shows the fixed headers.

3.2 Evaluating Tacit Alignment

To evaluate tacit alignment across human and language model (LM) agents, we constructed a dual-track experiment that mirrors task framing, response modality, and input prompts across human and AI agent participants. Both groups first complete individual-difference questionnaires, followed by the same spectrum alignment task under matched instructions. The goal is to assess behavioral calibration—how well a participant's placements align with that of a similarly profiled peer—under minimal communication. We investigate how traits are associated with independent judgment alignments between humans and agents. A key design decision is the inclusion of a large, structured agent bank with reusable memory profiles, enabling controlled human–agent comparison and trait–behavior analysis.

3.2.1 Human Participants

We recruited human participants through the Prolific platform. Participants were compensated at an average rate of $10.04 per hour, and the median completion time was 11 minutes and 57 seconds. Each participant's session consisted of two phases: a pre-task trait survey and the spectrum alignment task. All participants provided informed consent and completed the study via a hosted Qualtrics interface. Participants were included only if they passed all three sanity-check questions—two in the pre-task survey and one in the spectrum alignment game—and spent at least three minutes on the game task. After filtering out participants who failed attention check questions, the final retained dataset contains 241 human participants.

Pre-task Survey. Participants completed a survey questionnaire designed to elicit psychological and decision-making traits. This included: 1) Big Five Personality Inventory (BFI-10) (Rammstedt and John, 2007)—measuring openness, conscientiousness, extraversion, agreeableness, and neuroticism, 2) Empathy Scale–Cognitive Subscale (BES-A) (Carré et al., 2013)—assessing cognitive perspective-taking and awareness of others' emotional states, 3) Rational Ability Subscale of the Rational–Experiential Inventory (REI) (Pacini and Epstein, 1999)—measuring logical analysis and reasoning skill, 4) Narcissism Subscale of the Short Dark Triad (SD3) (Jones and Paulhus, 2014)—evaluating self-perceived social dominance and attention-seeking tendencies, and 5) General Decision-Making Style (GDMS) (Scott and Bruce, 1995)—capturing rational, intuitive, avoidant, de-

pendent, and spontaneous decision tendencies.

Spectrum Alignment Task. Participants then played the spectrum game described above. Responses were recorded on a continuous slider (0–100), with randomized non-core headers and a shared set of core headers.

Data Collection and Scale. The task used 22 spectrum headers in total. Human participants answered a subset of headers, at an average of 11.71 headers per participant: the shared core headers were administered broadly across participants, while additional non-core headers were sampled from the remaining header pool. Participants also reported confidence on a 1–5 scale for each round.

3.2.2 Agent Bank Construction

To enable controlled human–agent comparison, we constructed a bank of 200 profile-conditioned LLM agents. Each agent began with a demographic seed profile adapted from an open-source generative society simulation repository (Park et al., 2026). Because these seed profiles provide only limited context for inducing distinctive and diverse agent behavior, we further used an extended, simulated *profile-enrichment semi-structured interview script* to elicit richer autobiographical content from each seed persona. Agents were not given the full profile-enrichment transcript, and each task query triggered a top-K retrieval over the memory stream, selecting relevant memories based on semantic similarity, importance, and recency. Retrieved memory nodes were combined with demographic context and the current task header to construct the system prompt. This approach preserves rich persona context while avoiding context overflow; a detailed diagram of the persona-injection and memory workflow is provided in Appendix A.3 and A.4.

Behavioral inference across models. We instantiated the same 200 agent profiles across four models: *Qwen3-14B*, *Qwen2.5-7B-Instruct*, *Mistral-Nemo*, and *Llama-3.1-8B*. Each model received the same demographic seed profile, retrieved memory context, spectrum headers, and task instructions, and responded in zero-shot mode without fine-tuning or reward optimization. Unless otherwise noted, the main analyses use `Qwen3-14B` as the primary family because it was complete in the canonical data, showed the highest repeated-attempt consistency, had the strongest average trait–header expression summary, and produced the largest matched-versus-random TUX gap. The remaining families are reported as robustness analyses in Appendix B.2.

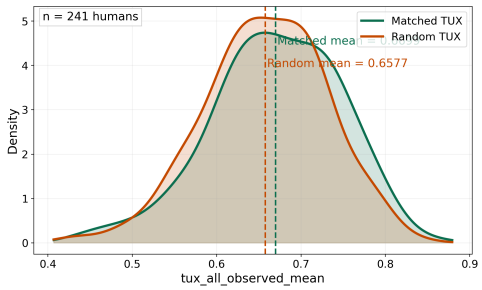

Figure 3: Matched-versus-random TUX distributions.

4 Results

4.1 RQ1: Measuring TUX

A core goal of this work is to operationalize tacit understanding in a weakly objective setting using a measurable pair-level quantity. For each human–agent pair and each overlapping observed header j, we define per-header similarity as $\mathtt{s_{haj}} = 1 - |\mathtt{y_{hj}} - \mathtt{y_{aj}}|$, where $\mathtt{y_{hj}}, \mathtt{y_{aj}} \in [0, 1]$ are the normalized human and agent choices on header j. We then define the final Tacit Understanding Index (TUX) as the mean similarity over all observed overlapping headers: $\mathrm{TUX}_{ha}^{\mathrm{all}} = \frac{1}{|O_h|} \sum_{j \in O_h} s_{haj}$, where O_h is the set of headers observed for human h and therefore overlapping with agent attempt a. Larger TUX indicates greater human–agent similarity.

Unless otherwise noted, our analyses use `Qwen3-14B` as the primary agent family, which showed the highest repeated-attempt consistency among the four agent families (`ICC`=0.72). Appendix B.2 reports robustness analyses across all four models and shows that the trait–TUX regression pattern is not unique to this primary family.

To test whether this operationalization captures meaningful tacit alignment, we compare the TUX of the nearest human–agent match in the 8-trait space against a random baseline defined as the mean TUX over all candidate agent attempts. The matched mean TUX is 0.67, compared with 0.65 for the random baseline, with a small–medium effect size (Cohen's d=0.29) and statistically significant paired t-test. Within this primary family, these results indicate that the proposed TUX operationalization captures a statistically reliable, though modest, aspect of tacit human–agent similarity.

4.2 RQ2: Explaining TUX with Traits

We fit OLS regression models predicting $\mathrm{TUX}_{ha}^{\mathrm{all}}$ from increasingly rich human–agent descriptor gaps. In the broad Model 4, the predictor set con-

tains eight primary trait gaps—extraversion, agreeableness, conscientiousness, neuroticism, openness, cognitive empathy, rational ability, and narcissism—five pre-task GDMS gaps—rational, intuitive, avoidant, dependent, and spontaneous—and the human's mean confidence on the shared core headers. The model can be written compactly as

$$\mathrm{TUX}_{ha}^{\mathrm{all}} = \alpha + \sum_{k=1}^{8} \beta_k d_{ha}^{(k)} + \sum_{\ell=1}^{5} \gamma_\ell g_{ha}^{(\ell)} + \delta c_h^{\mathrm{core}} + \varepsilon_{ha}. \quad (1)$$

Here, $d_{ha}^{(k)}$ denotes one of the eight primary trait gaps, $g_{ha}^{(\ell)}$ denotes one of the five GDMS gaps, each gap is an absolute human–agent difference in pooled-z standardized space, and c_h^{core} is the human's mean confidence on the shared core headers. Model 1 uses a single aggregated 8-trait distance. Model 2 replaces this aggregate with the eight individual primary trait gaps. Model 3 adds the five pre-task GDMS gaps. Model 4 adds mean confidence on the shared core headers.

As shown in Table 2, explanatory power increases monotonically as the predictor set becomes richer. Full-dataset R^2 rises from 0.05 in Model 1 to 0.15 in Model 4, and the corresponding Pearson correlation between fitted and observed TUX rises from r=0.23 to r=0.39. This progression suggests that TUX is not merely a noisy proxy: it becomes increasingly explainable as more concentrated person-level information is included. In other words, the latent ingredients of tacit understanding are not exhausted by one aggregate trait distance, and richer predictor sets recover more of the variation in the pair-level TUX signal.

The same pattern appears in predictive sensitivity, also reported in Table 2. Under repeated 80/20 human-level train/test splits (5 repeats), mean SMAPE remains low across all models, ranging from 9.60% to 9.38%, while mean training R^2 improves from 0.05 in Model 1 to 0.16 in Model 4. When held-out predictions are pooled across all five splits, Pearson correlation improves from r=0.24 in Model 1 to r=0.35 in Model 4. Thus, the richer models do not merely fit the full dataset more closely; they also improve held-out predictive performance under repeated human-level splits.

Target-definition ablation. As a target-definition ablation, we repeated the same regression family using the core-only mean TUX score instead of the all-observed mean TUX. Across all four models, the core-only target produced a higher full-dataset fit than the all-observed target: for example, R^2 increased from 0.05 to 0.07 in Model 1 and from 0.15 to 0.19 in Model 4, while the corresponding fitted–observed Pearson correlation increased from r=0.23 to r=0.26 in Model 1 and from r=0.39 to r=0.44 in Model 4. Therefore, we interpret the core-only result as a coverage-density sensitivity: when the predictand is restricted to a more uniformly shared header subset, the signal becomes cleaner, suggesting that denser shared-header coverage may sharpen future human–AI tacit-understanding benchmarks.

Predictor	M1	M2	M3	M4
Aggregated distance				
Agg. 8-trait distance	-0.2274***	–	–	–
Primary trait gaps				
Extraversion gap	–	-0.1625***	-0.1634***	-0.1300***
Agreeableness gap	–	-0.0191***	-0.0231***	-0.0464***
Conscientiousness gap	–	0.0066**	0.0076**	0.0138***
Neuroticism gap	–	-0.0422***	-0.0322***	-0.0391***
Openness gap	–	0.0372***	0.0399***	0.0433***
Cognitive empathy gap	–	-0.1041***	-0.0886***	-0.0470***
Rational ability gap	–	-0.0832***	-0.0802***	-0.0548***
Narcissism gap	–	-0.1042***	-0.0974***	-0.0922***
GDMS gaps				
GDMS rational gap	–	–	-0.1035***	-0.0780***
GDMS intuitive gap	–	–	-0.0224***	0.0001
GDMS avoidant gap	–	–	0.0326***	-0.0032
GDMS dependent gap	–	–	-0.0467***	-0.0394***
GDMS spontaneous gap	–	–	0.0330***	0.0313***
Confidence				
Mean confidence, core headers	–	–	–	-0.2650***
Model-level metrics				
Full-data R^2	0.0517***	0.0743***	0.0899***	0.1524***
Held-out SMAPE	9.5969	9.5892	9.6188	9.3777
Held-out Pearson's r	0.2399***	0.2707***	0.2599***	0.3477***

Table 2: OLS regression models with standardized coefficients. SMAPE and Pearson's r are computed by pooling held-out predictions across k-fold splits (k=5) (* p<0.05, ** p<0.01, *** p<0.001).

Comparing Predictors. Beyond overall model fit, the coefficient structure shows that some human–agent differences matter much more than others. To make predictors directly comparable, we inspect standardized coefficients from the broad model that includes the eight trait distances, five GDMS distances, and mean confidence on the shared core headers. Table 2 reports the standardized coefficients across the model sequence.

The dominant predictor is human mean core confidence, which enters with a negative coefficient (β=-0.27). Excluding confidence, the largest trait-side predictors in the broad model are extraversion (β=-0.13), narcissism (β=-0.09), and GDMS rational style (β=-0.08); among the primary non-GDMS trait variables, extraversion, narcissism, and rational ability form the strongest trio. This pattern suggests that tacit understanding is not uniformly distributed across the full trait set, but instead is driven disproportionately by a smaller subset of social and cognitive predictors.

The negative confidence effect is particularly notable. Follow-up diagnostics suggest that confidence is partly linked to response extremity: more

confident participants tended to make more extreme or idiosyncratic placements, which were less likely to align with agent placements and were therefore associated with lower TUX.

Qualitative interpretation of the dominant predictors. Excluding confidence, the strongest trait-side predictors are extraversion, narcissism, and rational ability, which together point to a plausible decomposition of tacit understanding into *social expressivity*, *self-oriented interpersonal stance*, and *deliberative calibration*. A cautious interpretation is that extraversion may matter because it shapes how legible or consistently projected a person's choices are to a partner; narcissism may matter because stronger self-oriented or competitive interpersonal tendencies can reduce supplementary fit and make tacit matching harder; and rational ability may matter because tacit coordination under ambiguity requires internally calibrated judgments rather than only social compatibility.

This interpretation is broadly consistent with prior work linking extraversion to some team-performance contexts, narcissism to less cooperative team conflict processes, and rationality to collective cognitive competence in group decision-making, while also suggesting that tacit understanding may not reduce to the same trait profile typically associated with overt teamwork performance (Peeters et al., 2006; Kichuk and Wiesner, 1997; Lynch et al., 2022; Curseu et al., 2013).

5 Discussion and Conclusion

We designed a spectrum-placement task for studying tacit understanding between humans and AI under weakly objective conditions, and proposed TUX as a pair-level measure. Using independent placements by humans (N=241) and profile-conditioned agents (N=200), we found that nearest human–agent matches in trait space achieve significantly higher TUX than a random baseline, supporting TUX as a credible measure of tacit understanding. Further, richer descriptor sets improved the explanation and prediction of TUX, while a smaller subset of predictors carries much of the signal. Together, these results suggest that tacit understanding is measurable and structurally constrained, but not fully captured by simple compatibility heuristics. More broadly, they show that human–AI alignment should not be studied only through explicit goal achievement: representational fit itself can be quantified, compared, and modeled.

Operationalizing Tacit Understandings. Our results suggest that tacit understanding in human–AI settings can be operationalized as a measurable behavioral construct rather than treated as a purely qualitative intuition. The key contribution of this work is not only the spectrum placement task, but also an operationalizable pair-level TUX that makes tacit calibration observable and statistically testable. In weakly objective settings, where explicit task success is unavailable, this kind of operationalization is important: it provides a way to study whether two agents independently converge on similar internal framings. This has implications for alignment evaluation: human–AI collaboration should not be assessed only through external performance, but also through whether systems can approximate the latent evaluative frames through which humans interpret ambiguous situations.

Tacit Understanding is Structured but Partial. Our findings suggest that tacit understanding is real but partial. Nearest human–agent matches in the active 8-trait space achieved significantly higher TUX than a random baseline, indicating that human–agent tacit understanding is not reducible to chance. However, the absolute gap remains modest, and even the richest descriptor models explain only a limited share of TUX variance. We view this combination of reliability and incompleteness as theoretically meaningful. The implication is that profile-conditioned agents may approximate some aspects of human representational alignment, but current profiles remain too shallow to fully capture the tacit priors that shape human judgment.

Designing Richer Measures of Person-Level Alignment. The regression progression provides a complementary implication. As predictor sets become richer, both explanatory fit and predictive sensitivity improve. This pattern supports the interpretation that TUX is a meaningful pair-level construct: it responds systematically to psychologically and behaviorally relevant descriptors.For benchmark design, this suggests that future evaluations of tacit understanding can invest in richer person descriptors, including traits, states, preferences, lived experiences, and contextual styles, beyond what we considered in this study. The same lesson appears from the header-specific analysis: broader header sets carry substantially richer trait signal than the shared core subset alone, implying that future benchmarks may benefit from both richer *person descriptions* and richer *prompt sets*.

6 Limitations and Future Directions

This work has limitations, which also suggest interesting future directions. First, TUX is a starting point of operationalizing tacit understanding. While the matched-versus-random result supports it as a meaningful behavioral measure, tacit understanding is a broader construct than pairwise placement similarity in a single task family. Our task is built around a spectrum-placement paradigm inspired by Wavelength. This structure is useful for isolating internal calibration without explicit communication, but it also constrains the form that tacit understanding can take. Future work can examine whether the same conclusions hold under alternative operationalizations, including dialogic, affective, and co-creative settings.

Also, the current descriptor set is not comprehensive. Our work likely explains only a limited share of TUX variance, and inspires future directions of richer descriptors and contextual factors. These may include lived experience, cultural background, aesthetic preference, domain familiarity, and other forms of contextual prior not captured by standard survey scales. Future benchmarks can expand both the breadth and the granularity of person descriptors. Further, the task header coverage is unequal. Our main paper-facing TUX uses all observed overlapping headers, which better reflects the available behavioral data, but it also means that not all human participants contribute the same header set. The core-only ablation and the appendix header analysis help diagnose this issue, but they do not remove it. A future study with denser, fully shared header coverage would allow cleaner comparisons and may sharpen the resulting benchmark.

Further, the agent bank depends on a specific persona-construction pipeline. Our agents are not generic language models; they are profile-conditioned agents built from demographic seeds, interview-style enrichment, and memory retrieval. This is a strength for controlled experimentation, but it also means the findings should be interpreted as applying to this class of profile-conditioned LLM agents rather than to all AI systems. Different profile construction or conditioning methods may yield different tacit-understanding patterns. Additionally, our model-family comparison remains bounded. Although the study includes multiple models and a shared agent bank, it is still only a limited slice of the broader space of contemporary language models. Future work can test whether the present structural patterns replicate across other larger families, instruction-tuned variants, and non-LLM agent architectures.

Again, our analyses are explanatory and predictive, but not causal. The fact that richer descriptor sets improve TUX prediction does not imply that any individual trait directly causes tacit understanding. Likewise, the coefficient ranking should be interpreted as a structured association pattern rather than a definitive mechanism. Our work inspires future research in exploring causal pathways through targeted manipulations or intervention-based experimental designs.

7 Ethical Considerations

This study was reviewed and approved by the authors' institutional review board (IRB) at our research institution. Human participants were recruited through Prolific and completed the study through a hosted Qualtrics interface after providing informed consent. Participants were compensated at an average rate of $10.04 per hour, and participation was voluntary. We applied pre-specified quality-control criteria, including sanity-check questions and minimum task-duration filtering, to ensure data validity without penalizing attentive participants.

The human dataset was de-identified before analysis. We do not report participant names, contact information, or other directly identifying information. Survey responses were used only to construct aggregate trait descriptors for analysis, and individual-level results are reported only in aggregate. Because the task asks participants to make subjective judgments and report confidence, we treat these responses as behavioral measurements rather than evaluations of participant ability or correctness.

The agent portion of the study uses profile-conditioned LLM agents constructed from demographic seed profiles and simulated interview memories. These profiles are used to support controlled experimental comparison, not to make claims about real individuals or demographic groups. We therefore interpret agent behavior as behavior of model-conditioned personas rather than as representative of any real population. Any public release of the data bank will exclude personally identifying human information and will document the synthetic nature and limitations of the agent profiles.

8 AI Involvement Disclosure

AI-assisted tools were used only for language editing, formatting assistance, and improving clarity

during manuscript preparation. The study design, data collection, statistical analyses, interpretation of results, and experimental decisions were conducted by the authors.

References

Elisa C. Baek and Carolyn Parkinson. 2022. Shared understanding and social connection: Integrating approaches from social psychology, social network analysis, and neuroscience. *Social and Personality Psychology Compass*, 16(11):e12710.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11.

Arnaud Carré, Nicolas Stefaniak, Fanny D'Ambrosio, Leïla Bensalah, and Chrystel Besche-Richard. 2013. The basic empathy scale in adults (bes-a): Factor structure of a revised form. *Psychological Assessment*, 25(3):679–691.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

P.L. Curseu, R.J.G. Jansen, and M.M.H. Chappin. 2013. Decision rules and group rationality: Cognitive gain or standstill? *PLOS ONE*, 8(2):e56454.

Vedant Das Swain, Qiuyue" Joy" Zhong, Jash Rajesh Parekh, Yechan Jeon, Roy Zimmermann, Mary P Czerwinski, Jina Suh, Varun Mishra, Koustuv Saha, and Javier Hernandez. 2025. Ai on my shoulder: Supporting emotional labor in front-office roles with an llm-based empathetic coworker. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–29.

Xiaoqi Deng, Sarinasadat Hosseini, Yoshihiro Miyake, and Takayuki Nozawa. 2024. Cooperativeness as a personality trait and its impact on cooperative behavior in young east asian adults who synchronized in casual conversations. *Behavioral Sciences*, 14(11):987.

Weihua Du, Qiushi Lyu, Jiaming Shan, Zhenting Qi, Hongxin Zhang, Sunli Chen, Andi Peng, Tianmin Shu, Kwonjoon Lee, Behzad Dariush, and Chuang Gan. 2024. Constrained human-ai cooperation: An inclusive embodied social intelligence challenge. In *Advances in Neural Information Processing Systems*, volume 37.

Quentin Ehkirch and Akane Matsumae. 2024. Understanding the influence of interpersonal factors on interactions in co-design through intersubjectivity: A systematic literature review. *Design Science*, 10:e4.

George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. 2024. Evaluating human-ai collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*.

Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42.

E. Tory Higgins, Maya Rossignac-Milon, and Gerald Echterhoff. 2021. Shared reality: From sharing-is-believing to merging minds. *Current Directions in Psychological Science*, 30(2):103–110.

Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of Personality: Theory and Research*, 2 edition, pages 102–138. Guilford Press.

Daniel N. Jones and Delroy L. Paulhus. 2014. Introducing the short dark triad (sd3): A brief measure of dark personality traits. *Assessment*, 21(1):28–41.

Dan M. Kahan, Hank Jenkins-Smith, and Donald Braman. 2011. Cultural cognition of scientific consensus. *Journal of Risk Research*, 14(2):147–174.

Susan L. Kichuk and Willi H. Wiesner. 1997. The big five personality factors and team performance: Implications for selecting successful product design teams. *Journal of Engineering and Technology Management*, 14(3–4):195–221.

Jiwon Kim, Violeta J Rodriguez, Dong Whi Yoo, Eshwar Chandrasekharan, and Koustuv Saha. 2026. Pair-safe: A paired-agent approach for runtime auditing and refining ai-mediated mental health support. *arXiv preprint arXiv:2601.12754*.

Sangwook Lee, Adnan Abbas, Yan Chen, Young-Ho Kim, and Sang Won Lee. 2026. Choir: A chatbot-mediated organizational memory leveraging communication in university research labs. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, Yongbin Li, and Dacheng Tao. 2025. A survey of direct preference optimization. *arXiv preprint arXiv:2503.11701*.

Jennifer Lynch, Alexander McGregor, and Alex J. Benson. 2022. My way or the highway: Narcissism and dysfunctional team conflict processes. *Group Processes & Intergroup Relations*, 25(4):1157–1171.

Amogh Mannekote, Adam Davies, Guohao Li, Kristy Elizabeth Boyer, ChengXiang Zhai, Bonnie J. Dorr, and Francesco Pinto. 2025. Do role-playing agents practice what they preach? belief-behavior consistency in llm-based simulations of human trust. *arXiv preprint arXiv:2507.02197*.

Katelyn Morrison, Zahra Ashktorab, Gabriel Enrique Gonzalez, Djallel Bouneffouf, and Justin Weisz. 2025. Establishing the cooperative game wavelength as a testbed to explore mutual theory of mind. In *Proceedings of the Workshop on Advancing AI Through Theory of Mind at AAAI 2025*.

Sheshera Mysore, Debarati Das, Hancheng Cao, and Bahareh Sarrafzadeh. 2025. Prototypical human-ai collaboration behaviors from llm-assisted writing in the wild. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16830–16857.

Rosemary Pacini and Seymour Epstein. 1999. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, 76(6):972–987.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. 2026. Llm agents grounded in self-reports enable general-purpose simulation of individuals. *Preprint*, arXiv:2411.10109.

Miranda A.G. Peeters, Harrie F.J.M. van Tuijl, Christel G. Rutte, and Isabelle Reymen. 2006. Personality and team performance: a meta analysis. *European Journal of Personality*, 20(5):377–396.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2024. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.

Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1):203–212.

Thomas C. Schelling. 1960. *The Strategy of Conflict*. Harvard University Press.

Jan B. Schmutz, Neal Outland, Sophie Kerstan, Eleni Georganta, and Anna-Sophie Ulfert. 2024. Ai-teaming: Redefining collaboration in the digital era. *Current Opinion in Psychology*, 58:101837.

Susanne G. Scott and Reginald A. Bruce. 1995. Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement*, 55(5):818–831.

Julian Senoner, Simon Schallmoser, Bernhard Kratzwald, Stefan Feuerriegel, and Torbjørn Netland. 2024. Explainable ai improves task performance in human–ai collaboration. *Scientific reports*, 14(1):31150.

Soorya Ram Shimgekar, Vipin Gunda, Jiwon Kim, Violeta J Rodriguez, Hari Sundaram, and Koustuv Saha. 2026. Ai psychosis: Does conversational ai amplify delusion-related language? *arXiv preprint arXiv:2603.19574*.

Matthew Sidji, Wally Smith, and Melissa J. Rogerson. 2024. Human-ai collaboration in cooperative games: A study of playing codenames with an llm assistant. *Proceedings of the ACM on Human-Computer Interaction*, 8(CHI PLAY):1–25.

Chandler Smith, Marwa Abdulhai, Manfred Diaz, Marko Tesic, Rakshit Trivedi, Sasha Vezhnevets, Lewis Hammond, Jesse Clifton, Minsuk Chang, Edgar Duenez-Guzman, and 1 others. 2026. Evaluating generalization capabilities of llm-based agents in mixed-motive scenarios using concordia. *Advances in neural information processing systems*, 38.

Kihoon Son, DaEun Choi, Tae Soo Kim, and Juho Kim. 2024. Demystifying tacit knowledge in graphic design: Characteristics, instances, approaches, and guidelines. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16525–16548.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Gaofeng Wang and Weiwei Hu. 2021. Peer relationships and college students' cooperative tendencies: Roles of interpersonal trust and social value orientation. *Frontiers in Psychology*, 12:656412.

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 511–520. IEEE.

Guande Wu, Chen Zhao, Claudio Silva, and He He. 2024. Your co-workers matter: Evaluating collaborative capabilities of language models in blocks world. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4941–4957.

Jianzhu Yao, Kevin Wang, Ryan Hsieh, Haisu Zhou, Tianqing Zou, Zerui Cheng, Zhangyang Wang, and Pramod Viswanath. 2025. Spin-bench: How well do llms plan strategically and reason socially? *arXiv preprint arXiv:2503.12349*.

Dora Zhao, Diyi Yang, and Michael S Bernstein. 2025. Knoll: Creating a knowledge ecosystem for large language models. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–23.

Huiqi Zou, Pengda Wang, Zihan Yan, Tianjun Sun, and Ziang Xiao. 2025. Can LLM "self-report"?: Evaluating the validity of self-report scales in measuring personality design in LLM-based chatbots. In *Second Conference on Language Modeling*.

Table A.1: Pre-task survey instruments used to construct human and agent descriptors.

Construct	Source	Items used	Scale
Big Five personality	BFI-10 (Rammstedt and John, 2007)	Reserved; generally trusting; tends to be lazy; relaxed, handles stress well; has few artistic interests; outgoing, sociable; tends to find fault with others; does a thorough job; gets nervous easily; has an active imagination.	1–5
Cognitive empathy	BES-A cognitive empathy subscale (Carré et al., 2013)	Finds it hard to know when friends are frightened; can usually work out when friends are scared; can often understand how people are feeling before they say; not usually aware of friends' feelings; has trouble figuring out when friends are happy.	1–5
Rational ability	REI rational ability subscale (Pacini and Epstein, 1999)	Not good at complicated problems; not good at problems requiring careful logical analysis; not very analytical; reasoning things out carefully is not a strong point; does not reason well under pressure.	1–5
Narcissism	SD3 narcissism subscale (Jones and Paulhus, 2014)	People see me as a natural leader; I hate being the center of attention (reverse-coded); many group activities are dull without me; I know I am special because everyone keeps telling me so; I like to get acquainted with important people.	1–5
Decision-making style	GDMS (Scott and Bruce, 1995)	Logical and systematic decision-making; avoiding decisions until pressure is on; consulting other people before important decisions; relying on intuition; making snap decisions; exploring all options before deciding.	1–5

A Study Materials

This appendix summarizes the study materials used to collect human and agent responses. We include the pre-task survey instruments, the full set of spectrum headers, and the profile/interview structure used to instantiate persona-conditioned agents.

A.1 Pre-task Survey Instruments

Before the spectrum-placement task, participants completed a pre-task survey measuring personality, empathy, rational ability, narcissism, and decision-making style. Unless otherwise noted, all items used a five-point Likert scale from 1 = *Disagree strongly* to 5 = *Agree strongly*. Table A.1 summarizes the constructs used in the survey.

A.2 Full Spectrum-Placement Header Set

Each spectrum-placement trial presents a cue and two spectrum anchors. Participants place the cue along the spectrum according to where they believe a like-minded peer would place it. Table A.2 lists the full set of 22 headers used in the study.

A.3 Agent Persona Construction Materials

Agents were initialized from structured demographic seed profiles and then enriched through a simulated semi-structured interview. The seed profile provided stable background attributes, while the interview generated autobiographical memory content used for retrieval-conditioned downstream responses.

Seed profile schema. Each seed profile contained demographic and background fields that served as static identity anchors. The schema included name fields, age, sex, race/ethnicity descriptors, city/state, family background, education, work status, marital status, language background, financial background, and a short profile summary. For privacy and reproducibility, we report the schema categories rather than a complete raw seed profile. Table A.3 summarizes the fields used.

Interview script. The simulated interview contained 45 open-ended questions designed to elicit autobiographical memories, values, daily routines, relationships, health, stress, finances, work, family background, and future goals. Each question–answer turn was stored as a memory node and used as part of the agent's persistent memory stream. During task inference, relevant memory nodes were retrieved and included in the agent prompt.

Table A.4 summarizes the major interview domains, and Table A.5 gives representative questions. The full 45-question script is included in the supplemental repository.

Table A.2: Full set of spectrum-placement headers.

ID	Cue	Left anchor	Right anchor
0	A coworker takes credit for your idea	Confront them directly	Let it go to avoid tension
1	A large party	Leave early and rest	Stay late and meet people
2	Planning a weekend trip	Detailed itinerary	Go with the flow
3	Quitting stable job for a new startup	Uncomfortable	Excited
4	Bungee jumping	Stressful	Thrilling
5	Beach	Calm	Crowded
6	Your music taste	Quiet	Bombastic
7	Public speaking	Uncomfortable	Energizing
8	Smartphone	Useful	Harmful
9	Alarm Clock	Helpful	Annoying
10	Neighbor	Friendly	Nosy
11	Beautifulness	Objective	Subjective
12	Vacation	Familiar place	New destination
13	Drinking	Cringe	Cool
14	Competition	Motivating	Stressful
15	Glass	Durable	Fragile
16	Aircraft engine	Distraction	Focus
17	Coffee	Relaxing	Energizing
18	Surprise party	Overwhelming	Delightful
19	Texting style	Brief and direct	Expressive and detailed
20	Genius	Lucky	Hardworking
21	Deadline	Early completion	Last-minute burst

Table A.3: Agent seed profile schema. We report field categories rather than raw profile values to avoid exposing sensitive or identifying example details.

Category	Fields
Basic identity	first name, last name, age, sex, city, state
Race and ethnicity	ethnicity, race, detailed race, Hispanic origin
Family background	residence at age 16, same residence since age 16, family structure at age 16, family income at age 16, parents' education, mother's work history
Education and work	highest degree received, work status, occupation-related summary when available
Household and social status	marital status, military service, citizenship status, language background
Financial background	total wealth / financial-status field
Profile summary	short natural-language summary of the agent's background, values, and life context

A.4 Agent Memory Architecture

Figure A.1 illustrates the persona-injection and memory-stream workflow used for profile-conditioned agents. Each agent begins from a demographic seed profile. A simulated interview then elicits life-history and value-related responses, which are segmented and written into a persistent memory stream. During spectrum-placement inference, the current header retrieves top-K relevant memory nodes, which are combined with demographic context and the current task into a system prompt header. The LLM agent then generates a spectrum-placement response consisting of reasoning and a numerical placement.

B Additional TUX Robustness and Diagnostic Analyses

B.1 Separate Human and Agent Descriptor Regression

The main regression analysis models TUX using human–agent descriptor gaps, which directly tests whether trait similarity predicts pair-level tacit understanding. As an exploratory extension, we also fit models using human and agent descriptors as separate predictors. This specification asks whether TUX is predictable from the directional positions of the human and agent in descriptor space, rather than from absolute gaps alone. A-family models use separate human and agent descriptors; B-family models additionally include same-trait human$\times$agent interactions.

The separate-descriptor extension provides two diagnostic observations. First, modeling human and agent descriptors separately improves fit relative to the main gap-based model: the best separate-descriptor model reaches $R^2 = 0.1835$ ($r = 0.4284$, $\rho = 0.4128$), compared with $R^2 = 0.1524$ ($r = 0.3904$, $\rho = 0.3614$) for the main gap-based model. Second, most of this gain comes from separate human/agent descriptors and core confidence rather than interaction terms: adding same-trait interactions increases full-dataset R^2 only from 0.1804 to 0.1835 when confidence is included. We therefore treat this as a

Table A.4: Major domains covered by the agent interview script.

Domain	Approximate question IDs
Life history and turning points	1–4
Family and close relationships	5–7
Neighborhood and living environment	8–10
Daily routines and responsibilities	11–15
Health and wellbeing	16–22
Religion, social media, and coping	23–26
Finances and household budget	27–32
Work and workplace context	33–40
Parents' work background	41–43
Future goals and values	44–45

Table A.5: Representative questions from the 45-question simulated interview script.

ID	Question
1	Tell me the story of your life, from childhood to education, family, relationships, and major life events.
2	Was there a crossroads in your life where multiple paths were available and your choice made a significant difference in defining who you are?
5	Tell me more about family who are important to you. Do you have a partner or children?
8	Tell me about the neighborhood and area in which you are living now.
11	Across a typical week, how do your days vary?
16	Tell me all about your health.
20	Tell me about how you have been doing over the past year.
27	Tell me about how you make ends meet and what the monthly budget looks like for you and your family.
33	What is or was your occupation, and what kind of work do you do?
40	Overall, how do you feel about your financial situation?
44	Imagine yourself a few years from now. What do you hope your life will look like?
45	What do you value the most in your life?

diagnostic result: TUX is partly predictable from directional human and agent profiles, but the main gap-based model remains the more direct test of pairwise trait-similarity alignment.

B.2 Agent Family Ablation

The main results use `Qwen3-14B` as the primary agent family. Table B.3 summarizes reliability and response-distribution diagnostics. `Qwen3-14B` had the highest repeated-attempt ICC (0.7235), the highest binned response entropy (2.7642), and the strongest average trait–header expression summary (mean absolute trait–header correlation $= 0.0909$). We therefore treat `Qwen3-14B` as a pragmatic primary family because it provides the strongest combination of repeated-attempt consistency, non-collapsed response diversity, trait-conditioned expressivity, and matched-versus-random TUX signal.

Table B.4 compares the regression specification across families. The Model 4 specification includes eight primary trait gaps, five pre-task GDMS gaps, and human mean confidence. The trait–TUX relationship is not unique to `Qwen3-14B`: all four families yield significant fitted–observed correlations and significant pooled held-out correlations. Interestingly, `Qwen2.5-7B` achieves the strongest full-data and held-out regression metrics ($R^2 = 0.2233$, held-out $r = 0.4378^{***}$), indicating that the regression result is robust across agent families even though the primary family was selected for broader reliability and matched-alignment criteria.

Table B.5 reports matched-versus-random TUX comparisons across families. `Qwen3-14B` produces the largest matched-minus-random gap (0.0122; $t = 4.5498$, $d_z = 0.2931$), followed by `Llama-3.1-8B` (0.0106; $t = 3.8601$, $d_z = 0.2487$). The effect is weaker and not significant for `Qwen2.5-7B` and `Mistral-Nemo`. These results suggest that model-family choice affects the strength of different TUX diagnostics: some families yield stronger regression predictability, while others yield stronger matched-versus-random alignment signals.

C When TUX Matters

This section reports diagnostics that ask when the matched-versus-random TUX advantage is largest and which headers appear most informative for benchmark construction.

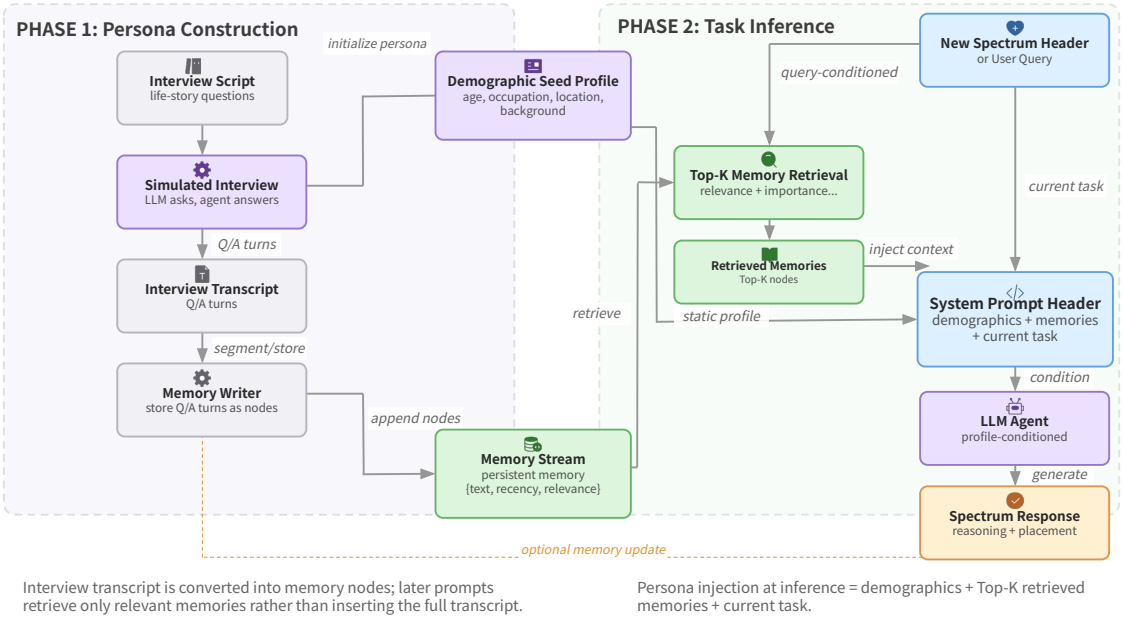

Figure A.1: Generative-agent persona and memory workflow. Demographic seed profiles initialize a simulated interview that produces autobiographical Q/A transcripts. These transcripts are segmented into persistent memory nodes rather than inserted wholesale into later prompts. During spectrum-placement inference, the current header retrieves top-K relevant memories, which are combined with demographic context into a system prompt header that conditions the LLM agent's response.

C.1 Header-Level TUX Component Gaps

To examine where the matched-versus-random TUX advantage is concentrated, we decomposed the human-level TUX gap into per-header components. For each header j, we computed the signed component gap

$$g_j = \frac{1}{|H_j|} \sum_{h \in H_j} \left[1 - |y_{hj} - y_{m(h)j}| - \frac{1}{N_a} \sum_{a} (1 - |y_{hj} - y_{aj}|) \right], \tag{2}$$

where H_j is the set of humans who answered header j, $m(h)$ is the trait-matched agent for human h, and N_a is the number of candidate agent attempts. Positive values indicate that the trait-matched agent is closer to the human response than the random candidate-pool average on that header.

The largest positive component gaps appeared on headers such as *Bungee jumping* ($g_j = 0.0757$), *Public speaking* ($g_j = 0.0751$), and *Surprise party* ($g_j = 0.0718$). Because normalized TUX distances are defined on a $[0, 1]$ scale mapped from the original $[-100, 100]$ slider, these correspond to average raw-spectrum reductions of approximately 15.14, 15.02, and 14.36 points, respectively (Figure C.1).

We also examined whether header-level response variance predicted where the matched advantage was largest. Human response variance was not clearly associated with the absolute component gap ($r = 0.2607$, $p = 0.2412$; $\rho = 0.1237$, $p = 0.5835$). In contrast, agent response variance was positively associated with absolute component gap ($r = 0.4635$, $p = 0.0298$; $\rho = 0.5280$, $p = 0.0116$). This suggests that TUX is most informative on headers that induce differentiated agent responses, rather than on headers that merely produce broad human disagreement. For benchmark construction, this points toward selecting headers that elicit meaningful variation in agent-side interpretations while preserving semantic coherence.

C.2 Header-Specific Trait Diagnostics

We also examined whether some game headers are more informative than others for tacit-understanding benchmark construction. The results suggest substantial heterogeneity across headers. At the header level, trait explainability varied markedly: for example, under the descriptor-gap model, *A large party* and *Surprise party* reached $R^2 = 0.1609$ and 0.1625, respectively, whereas *Neighbor* and *Beach* were

Table B.1: Full-dataset OLS regression extension models predicting TUX using human and agent descriptors as separate predictors. Predictor cells report signed standardized coefficients with significance stars. A-family models use separate human and agent descriptors; B-family models additionally include same-trait human×agent interactions. Significance legend: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Predictor / metric	A1	A2	A3	B1	B2	B3
Human primary traits						
Human Extraversion	0.3011^{***}	0.2892^{***}	0.2449^{***}	0.2820^{***}	0.2702^{***}	0.2259^{***}
Human Agreeableness	-0.0100^{***}	-0.0117^{***}	0.0006	-0.0222^{***}	-0.0239^{***}	-0.0116^{***}
Human Conscientiousness	0.0031	0.0383^{***}	0.0363^{***}	0.0017	0.0369^{***}	0.0349^{***}
Human Neuroticism	0.0516^{***}	0.0136^{***}	-0.0207^{***}	0.0528^{***}	0.0147^{***}	-0.0196^{***}
Human Openness	0.0031	-0.0201^{***}	-0.0107^{***}	0.0034	-0.0198^{***}	-0.0104^{***}
Human Cognitive empathy	-0.1641^{***}	-0.1460^{***}	-0.0919^{***}	-0.1641^{***}	-0.1462^{***}	-0.0921^{***}
Human Rational ability	0.0240^{***}	0.0448^{***}	0.0353^{***}	0.0169^{***}	0.0378^{***}	0.0283^{***}
Human Narcissism	0.0124^{***}	0.0335^{***}	0.0215^{***}	0.0085^{**}	0.0296^{***}	0.0176^{***}
Agent primary traits						
Agent Extraversion	-0.0339^{***}	-0.0266^{***}	-0.0266^{***}	-0.0124^{***}	-0.0051	-0.0051
Agent Agreeableness	-0.0129^{***}	-0.0052	-0.0052	-0.0037	0.0040	0.0040
Agent Conscientiousness	0.0158^{***}	0.0132^{***}	0.0132^{***}	0.0197^{***}	0.0172^{***}	0.0172^{***}
Agent Neuroticism	-0.0015	-0.0055^{*}	-0.0055^{*}	-0.0030	-0.0071^{*}	-0.0071^{**}
Agent Openness	-0.0031	-0.0003	-0.0003	-0.0035	-0.0007	-0.0007
Agent Cognitive empathy	-0.0033	-0.0044	-0.0044	-0.0032	-0.0041	-0.0041
Agent Rational ability	0.0364^{***}	0.0306^{***}	0.0306^{***}	0.0501^{***}	0.0441^{***}	0.0441^{***}
Agent Narcissism	-0.0137^{***}	-0.0095^{***}	-0.0095^{***}	-0.0095^{***}	-0.0054	-0.0054^{*}
Human GDMS						
Human GDMS rational	–	-0.1389^{***}	-0.1029^{***}	–	-0.1384^{***}	-0.1023^{***}
Human GDMS avoidant	–	0.0810^{***}	0.0537^{***}	–	0.0807^{***}	0.0535^{***}
Human GDMS dependent	–	0.0100^{***}	0.0162^{***}	–	0.0108^{***}	0.0171^{***}
Human GDMS intuitive	–	-0.0180^{***}	-0.0112^{***}	–	-0.0174^{***}	-0.0106^{***}
Human GDMS spontaneous	–	-0.0589^{***}	-0.0291^{***}	–	-0.0551^{***}	-0.0253^{***}
Agent GDMS						
Agent GDMS rational	–	-0.0216^{***}	-0.0216^{***}	–	-0.0230^{***}	-0.0230^{***}
Agent GDMS avoidant	–	0.0012	0.0012	–	0.0015	0.0015
Agent GDMS dependent	–	-0.0096^{***}	-0.0096^{***}	–	-0.0115^{***}	-0.0115^{***}
Agent GDMS intuitive	–	-0.0109^{***}	-0.0109^{***}	–	-0.0118^{***}	-0.0118^{***}
Agent GDMS spontaneous	–	-0.0264^{***}	-0.0264^{***}	–	-0.0321^{***}	-0.0321^{***}
Same-trait interactions						
Extraversion H×A	–	–	–	0.0504^{***}	0.0502^{***}	0.0502^{***}
Agreeableness H×A	–	–	–	0.0201^{***}	0.0201^{***}	0.0201^{***}
Conscientiousness H×A	–	–	–	-0.0124^{***}	-0.0127^{***}	-0.0127^{***}
Neuroticism H×A	–	–	–	0.0132^{***}	0.0131^{***}	0.0131^{***}
Openness H×A	–	–	–	0.0026	0.0027	0.0027
Cognitive empathy H×A	–	–	–	-0.0002	-0.0005	-0.0005
Rational ability H×A	–	–	–	-0.0215^{***}	-0.0211^{***}	-0.0211^{***}
Narcissism H×A	–	–	–	0.0140^{***}	0.0139^{***}	0.0139^{***}
Same-GDMS interactions						
GDMS rational H×A	–	–	–	–	0.0036	0.0036
GDMS avoidant H×A	–	–	–	–	-0.0009	-0.0009
GDMS dependent H×A	–	–	–	–	-0.0037	-0.0037
GDMS intuitive H×A	–	–	–	–	0.0031	0.0031
GDMS spontaneous H×A	–	–	–	–	0.0122^{***}	0.0122^{***}
Confidence						
Mean confidence, core headers	–	–	-0.2356^{***}	–	–	-0.2356^{***}
Model-level metrics						
Full-dataset R^2	0.1120	0.1341	0.1804	0.1150	0.1373	0.1835
Full-data Pearson	0.3346^{***}	0.3662^{***}	0.4247^{***}	0.3391^{***}	0.3705^{***}	0.4284^{***}
Full-data Spearman	0.3279^{***}	0.3583^{***}	0.4087^{***}	0.3322^{***}	0.3626^{***}	0.4128^{***}

much lower at $R^2 = 0.0165$ and 0.0262. Adding aligned header-specific confidence often increased explainability substantially, with representative gains such as *Public speaking* from 0.1190 to 0.2276, *Alarm Clock* from 0.0697 to 0.2026, and *Deadline* from 0.0541 to 0.1881.

Reverse-prediction analyses further showed that broader header sets carry meaningfully richer trait signal than the shared core-6 subset. In the human–agent pair setting, using all 22 header distances outperformed the core-6 ablation for every primary trait (e.g., extraversion: $R^2 = 0.1834$ vs. 0.1379; agreeableness: 0.1111 vs. 0.0415). The same pattern held in the human population setting (e.g., extraversion: 0.4260 vs. 0.3641; rational ability: 0.2046 vs. 0.0818) and in the agent population setting (e.g., agreeableness: 0.2873 vs. 0.0528; extraversion: 0.2844 vs. 0.1230). Taken together, these diagnostics suggest that header informativeness is far from uniform, and that future tacit-understanding benchmarks may benefit from richer prompt sets rather than relying only on a small universally shared subset.

D Additional Human-Side Analyses

D.1 Human GDMS Pre/Post Shift

As a human-only appendix analysis, we examined whether participants' self-reported decision-making style shifted from the pre-task survey to the post-task reflection items. This analysis asks whether

Table B.2: Predictive sensitivity for separate human/agent trait regression extension models under repeated 80/20 human-level train/test splits. SMAPE is averaged across five held-out test folds; Pearson and Spearman are computed by pooling held-out predictions across all five splits. R^2_{train} is averaged across the corresponding training folds. Significance legend: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Model	Predictor set	SMAPE	Pearson	Spearman	R^2_{train}
A1	Human + agent primary traits	9.2880	0.3263^{***}	0.3090^{***}	0.1040
A2	A1 + human/agent GDMS	9.2939	0.3246^{***}	0.2983^{***}	0.1266
A3	A2 + core confidence	9.1692	0.3753^{***}	0.3551^{***}	0.1767
B1	A1 + same-trait interactions	9.2681	0.3315^{***}	0.3149^{***}	0.1069
B2	B1 + GDMS values/interactions	9.2746	0.3296^{***}	0.3027^{***}	0.1297
B3	B2 + core confidence	9.1515	0.3796^{***}	0.3586^{***}	0.1798

Table B.3: Agent-family reliability and response-distribution diagnostics. All families contain 200 profiles, 22 headers, and complete parsed placements.

Agent family	Parse rate	Within SD	ICC	Header SD	Entropy
`Qwen3-14B`	1.0000	0.0769	0.7235	0.1190	2.7642
`Qwen2.5-7B`	1.0000	0.0688	0.3875	0.0825	1.9864
`Mistral-Nemo`	1.0000	0.1339	0.4786	0.1735	2.5470
`Llama-3.1-8B`	1.0000	0.1230	0.3452	0.1523	1.9847

completing spectrum-placement task changed how participants described their decision process. For each GDMS dimension, we computed the paired change score as $\Delta_i = \text{post}_i - \text{pre}_i$ and tested whether the mean change differed from zero using paired t-tests, with Wilcoxon signed-rank tests as a nonparametric robustness check.

The clearest shifts were a decrease in avoidant style and increases in intuitive and spontaneous style (Figure D.1). Avoidant style decreased substantially from pre-task to post-task ($\Delta = -0.5270$, 95% CI $[-0.6807, -0.3733]$, $t = -6.7543$, $p < 0.001$, Cohen's $d_z = -0.4351$). Intuitive style increased by $\Delta = 0.5809$ (95% CI $[0.4457, 0.7162]$, $t = 8.4067$, $p < 0.001$, $d_z = 0.5450$), and spontaneous style increased by $\Delta = 0.2490$ (95% CI $[0.0891, 0.4088]$, $t = 3.0673$, $p = 0.0024$, $d_z = 0.1976$). Rational style was effectively unchanged ($\Delta = 0.0041$, $p = 0.9291$), while the dependent-style proxy decreased modestly ($\Delta = -0.2199$, $p = 0.0306$, $d_z = -0.1401$).

These results suggest that spectrum-placement task may prompt participants to describe their behavior as less avoidant and more intuition-driven after completing the task. However, the post-task items are not uniformly literal repeats of the pre-task GDMS scale. In particular, the avoidant item is proxy-comparable through time pressure, and the dependent-style item is based on reliance on task cues rather than interpersonal dependence. We therefore interpret these shifts as descriptive evidence about participants' perceived task strategy, not as evidence about human–agent alignment or TUX validity.

Table B.4: Final Model 4 TUX regression comparison across agent families. Full-data Pearson and Spearman are fitted–observed correlations on the full dataset. Held-out Pearson and Spearman are computed by pooling held-out predictions across five human-level train/test splits. Significance legend: $^{***}p < 0.001$.

Agent family	Full R^2	Full r	Full ρ	SMAPE	Held-out r	Held-out ρ	Train R^2
`Qwen3-14B`	0.1524	0.3904^{***}	0.3614^{***}	9.3777	0.3477^{***}	0.3023^{***}	0.1554
`Qwen2.5-7B`	0.2233	0.4726^{***}	0.4640^{***}	8.8740	0.4378^{***}	0.4311^{***}	0.2258
`Mistral-Nemo`	0.1732	0.4161^{***}	0.4013^{***}	9.3405	0.4209^{***}	0.4036^{***}	0.1661
`Llama-3.1-8B`	0.1663	0.4078^{***}	0.3942^{***}	8.9030	0.4087^{***}	0.3895^{***}	0.1621

Table B.5: Matched-versus-random TUX comparison across agent families. Matched agents are nearest neighbors in the active 8-trait space; random TUX is the mean over candidate agent attempts for each human. Significance legend: $^{***}p < 0.001$.

Agent family	Matched	Random	Gap	Rel. improv.	t	d_z
`Qwen3-14B`	0.6699	0.6577	0.0122	1.8575%	4.5498^{***}	0.2931
`Qwen2.5-7B`	0.6707	0.6677	0.0030	0.4522%	1.7967	0.1157
`Mistral-Nemo`	0.6723	0.6667	0.0056	0.8453%	1.6835	0.1084
`Llama-3.1-8B`	0.6883	0.6776	0.0106	1.5699%	3.8601^{***}	0.2487

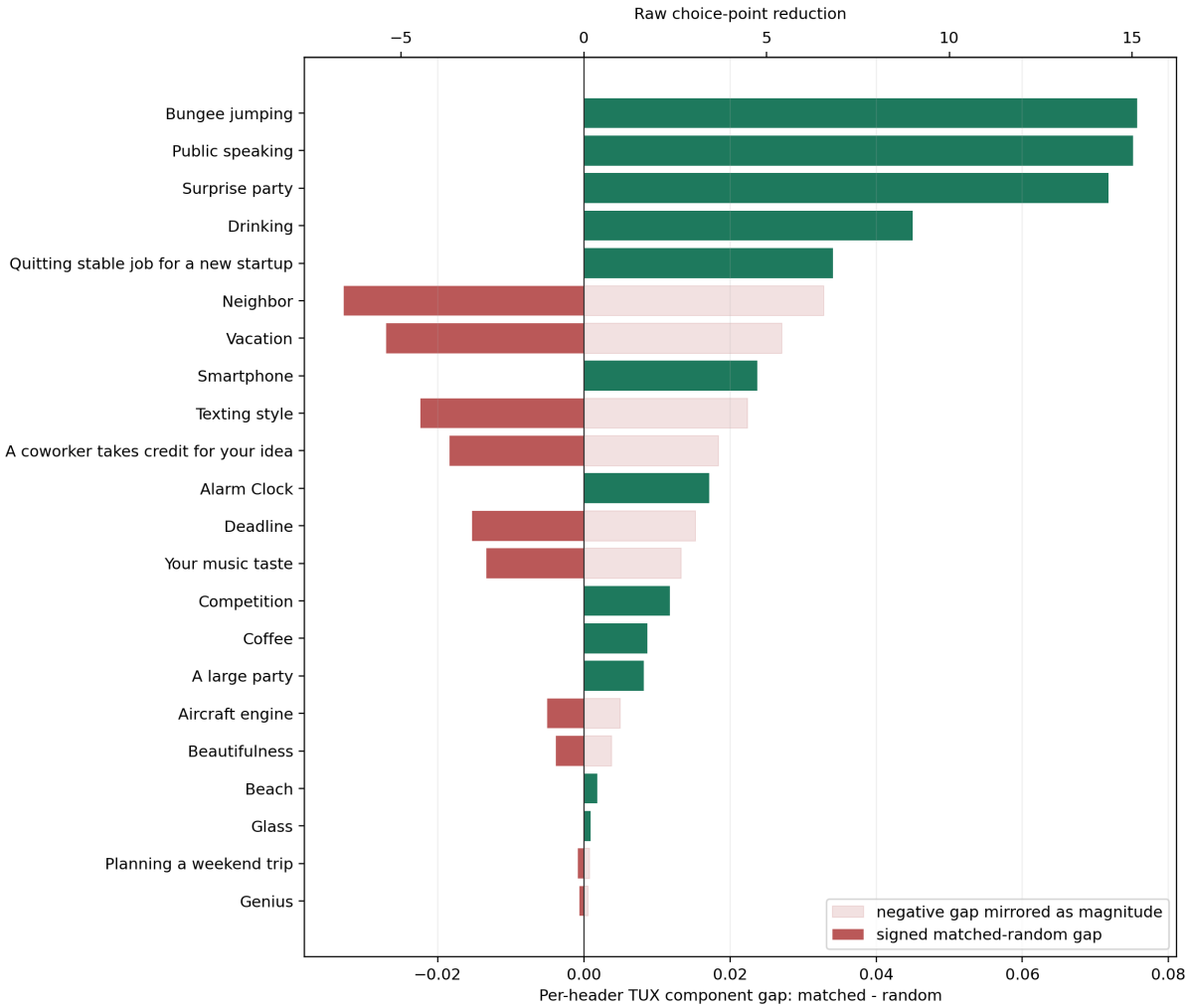

Figure C.1: Per-header matched-versus-random TUX component gaps. Positive values indicate headers where the trait-matched agent is closer to the human response than the random candidate-pool average. The top axis maps normalized TUX component gaps to raw slider-point reductions on the original $[-100, 100]$ response scale.

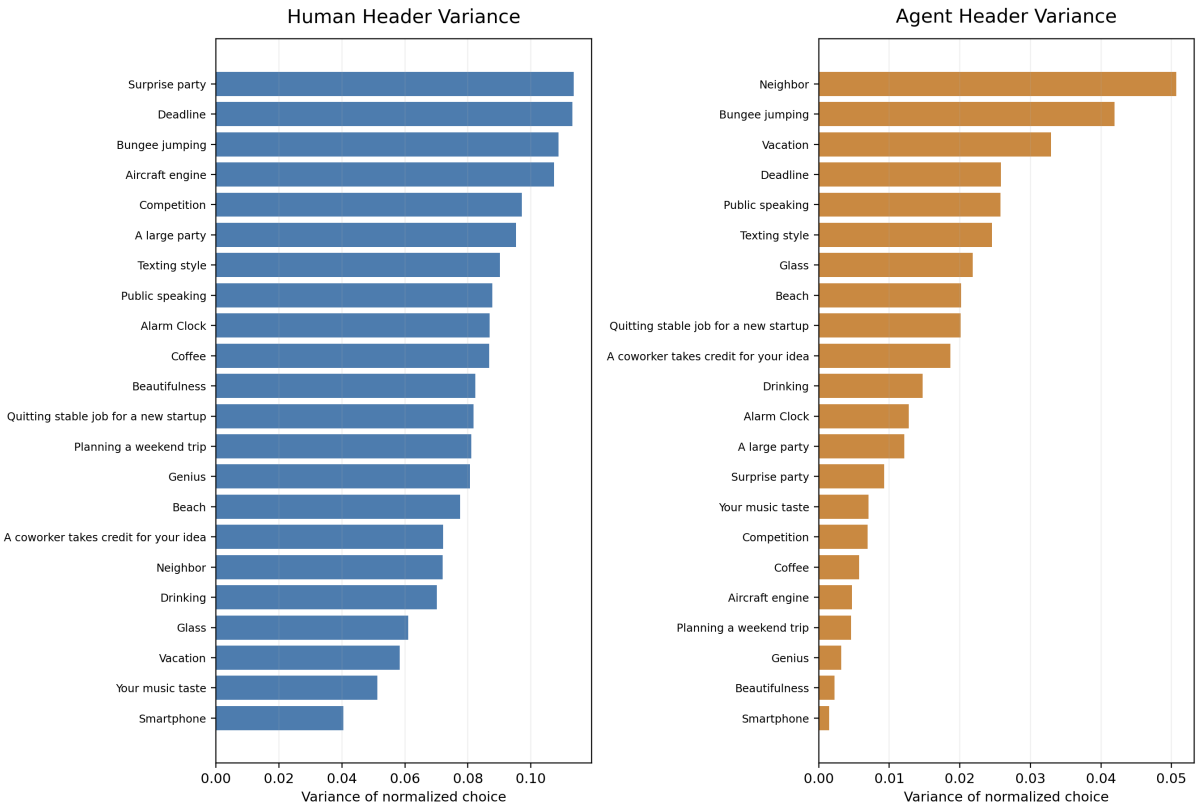

Figure C.2: Per-header response variance for humans and agents. Agent-side variance is positively associated with the absolute per-header matched-random TUX component gap, suggesting that headers with more differentiated agent responses carry more TUX-relevant signal.

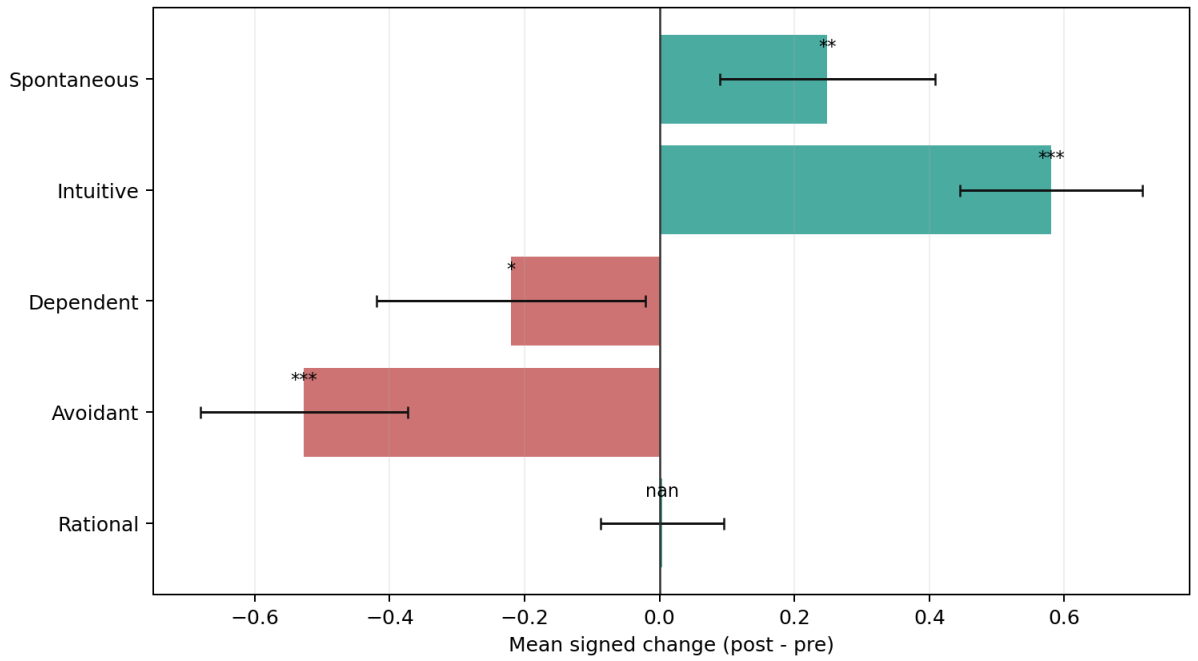

Figure D.1: Mean pre/post GDMS-style change among human participants. Positive values indicate higher post-task scores than pre-task scores; error bars show 95% confidence intervals.