Impatient Users Confuse AI Agents: High-fidelity Simulations of Human Traits for Testing Agents

Muyu He, Anand Kumar, Soumyadeep Bakshi, James Zou, Nazneen Rajani
Collinear AI Inc.
research@collinear.ai

Abstract

Despite rapid progress in building conversational AI agents, robustness is still largely untested. Small shifts in user behavior, such as being more impatient, incoherent, or skeptical, can cause sharp drops in agent performance, revealing how brittle current AI agents are. Today's benchmarks fail to capture this fragility: agents may perform well under standard evaluations but degrade spectacularly in more realistic and varied settings. We address this robustness testing gap by introducing `TraitBasis`, a lightweight, model-agnostic method for systematically stress testing AI agents. `TraitBasis` learns directions in activation space corresponding to steerable user traits (e.g., impatience or incoherence), which can be controlled, scaled, composed, and applied at inference time without any fine-tuning or extra data. Using `TraitBasis`, we extend τ-Bench to τ-`Trait`, where user behaviors are altered via controlled trait vectors. We observe an average 4%–20% performance degradation on τ-`Trait` across frontier models, highlighting the lack of robustness of current AI agents to variations in user behavior.

Together, these results highlight both the critical role of robustness testing and the promise of `TraitBasis` as a simple, data-efficient, and compositional tool. By powering simulation-driven stress tests and training loops, `TraitBasis` opens the door to building AI agents that remain reliable in the unpredictable dynamics of real-world human interactions. We plan to open-source τ-`Trait` across four domains: airline, retail, telecom, and telehealth, so the community can systematically QA their agents under realistic, behaviorally diverse intents and trait scenarios.

1 Introduction

One of the primary goals of multi-turn conversational AI agents is *generalization*. However, agents that perform well on benchmarks often fail to generalize in real-world deployments (BBC Travel, 2024; Steinhardt, 2024; Lecher, 2024). Prior work has shown that LLMs lack robustness to real-world noise and small input perturbations (Rabinovich and Anaby Tavor, 2025; Ye et al., 2024). A recurring cause of these failures is insufficient testing, particularly when user behavior deviates from typical intent or persona distributions.

Because testing deployed systems *in the wild* is costly and impractical, evaluation is typically limited to small sets of i.i.d. tasks or to agent benchmarks such as τ-Bench (Yao et al., 2024), MCPEvals (Wang et al., 2025), AgentBench (Liu et al., 2023), GTA (Wang et al., 2024a), and ToolBench (Qin et al., 2023). While useful as performance indicators, these benchmarks have limited coverage and do not explicitly test robustness.

For example, in the airline and retail domains of τ-Bench, we observe that frontier agent models such as GPT-4o, Kimi-K2 (Team et al., 2025), and GLM-4.5 (Zeng et al., 2025) suffer performance drops of up to 35%, 46%, and 17%, respectively, when only the user's interaction style (i.e., trait) is altered. This highlights the gap between benchmark performance and real-world robustness.

Prior work has explored naturalistic variations in user queries to stress-test specific capabilities, such as function calling (Rabinovich and Anaby Tavor, 2025), but does not address broader shifts in user personas. To bridge this gap, we propose `TraitBasis`, a lightweight, model-agnostic method for inducing high-fidelity user traits (e.g., *impatience*, *confusion*, *skepticism*, *incoherence*) that can be systematically scaled, composed, and applied at inference time, building on persona vector methods (Chen et al., 2025). `TraitBasis` estimates a *trait direction* in activation space by contrasting positive and negative exemplars and applies a scaled projection, enabling precise steering while preserving realism (Figure 1).

Using `TraitBasis`, we study four questions: (RQ1) **Realism**: which methods most reliably re-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 37569–37602
July 2-7, 2026

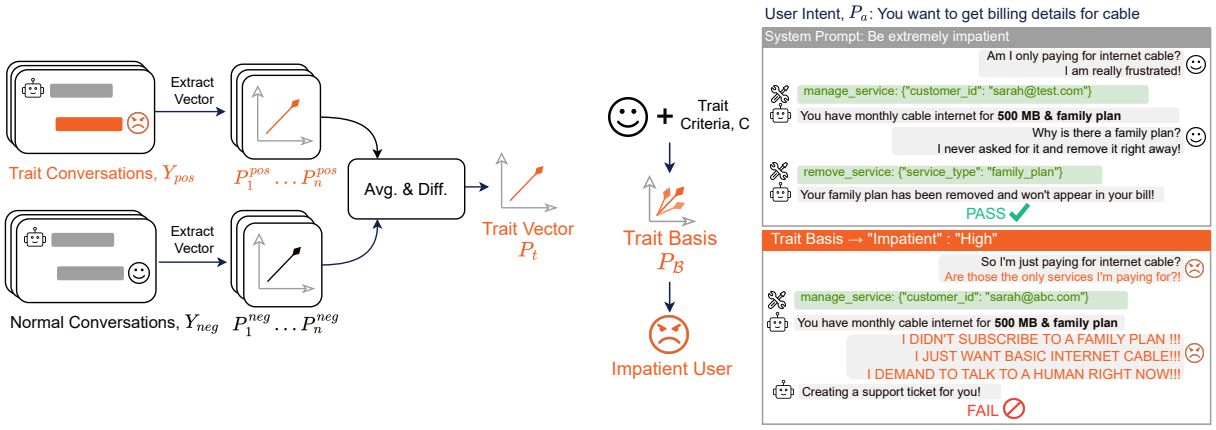

Figure 1: Illustration of our approach and comparison with prompt-based tuning. Trait prompt P_t is generated using contrastive conversations, where one dialogue exhibits the target trait while the other does not. **Left:** Trait vector extraction via contrastive activation differences. **Right:** When simulating a user with a specific trait, prompt-based tuning fails as the simulated user behavior becomes more realistic (the agent breaks down under trait shifts), while `TraitBasis` (generated using a combination of P_t's as shown in Section 3) remains robust, directly illustrating why agents fail under realistic trait shifts.

alize intended traits; (RQ2) **Fidelity**: whether different trait intensities are distinguishable by humans or an LLM judge; (RQ3) **Stability**: how traits persist over long multi-turn dialogues; and (RQ4) **Compositionality**: how effectively multiple traits can be combined. Empirically, `TraitBasis` outperforms prompt-based, full SFT, and LoRA baselines by 10% in realism, 2.5% in fidelity, 19.8% in stability, and 11% in compositionality.

To systematically evaluate robustness under persona shifts, we extend τ-Bench with τ-`Trait`, a more challenging benchmark that uses `TraitBasis` to generate diverse, high-fidelity user traits across four domains: airlines, retail, telecom, and telehealth. Unlike prior agent benchmarks that evaluate fixed i.i.d. tasks, τ-`Trait` introduces controlled trait perturbations—varying levels and mixtures of impatience, confusion, skepticism, and incoherence—that directly alter user–agent interaction. We observe that frontier agents experience performance degradations of up to 46% relative to τ-Bench. These controlled perturbations enable realistic multi-turn stress-testing, isolate robustness failures attributable to user behavior, and provide a principled link between benchmark results and real-world deployment risk.

Our contributions are threefold: (1) we introduce `TraitBasis`, a method for constructing realistic, high-fidelity simulations of four human traits—*impatience*, *confusion*, *skepticism*, and *incoherence*—that enables multi-turn user simulation without the persona collapse seen in prompt-based and SFT baselines; (2) through human and automated evaluations, we show that `TraitBasis` consistently outperforms prompt-based steering (Zheng et al., 2024), full supervised fine-tuning on trait-labeled data (Zhang et al., 2018a), and LoRA adapters (Hu et al., 2022) in realism, fidelity, long-horizon stability, and compositionality, and that these advantages generalize across model families (Llama and Qwen); and (3) we extend τ-Bench to τ-`Trait`, adding telecom and telehealth domains and using `TraitBasis` to generate high-fidelity, trait-driven user behaviors, revealing sharp performance degradation of frontier agents under user-behavior shifts.

2 Related Work

Testing and benchmarking AI agents Despite advances in scale and post-training, AI agents remain brittle on out-of-distribution (O.O.D.) tasks. Prior work shows that frontier models' function-calling degrades under small perturbations to user queries (Rabinovich and Anaby Tavor, 2025), and that tool use is not robust to the noise inherent in real-world interactions (Ye et al., 2024). In parallel, a growing body of work has proposed benchmarks for evaluating AI agents, including MCP-based evaluations and multi-turn interaction settings, such as MCPEval (Liu et al., 2025), MCPBench (Wang et al., 2025), MCPVerse (Lei et al., 2025), MCP-Universe (Luo et al., 2025), LiveMCP-101 (Yin

et al., 2025), τ-Bench (Yao et al., 2024), τ^2-Bench (Barres et al., 2025), AgentBench (Liu et al., 2023), ToolBench (Qin et al., 2023), GTA (Wang et al., 2024a), and BFCL (Patil et al., 2025). However, even benchmarks that model multi-turn interactions often rely on system prompts to simulate users, making it difficult to sustain complex, realistic user behavior over long conversations (Yao et al., 2024). Our contributions to τ-`Trait`, enabled by `TraitBasis`, aim to address this limitation by modeling richer and more persistent user traits. We note that related work on coding agents and red-teaming lies outside the scope of this paper.

Simulating user personas Realistic user persona simulation is essential for evaluating and stress-testing conversational AI systems. While system-prompt-based methods are simple to deploy, they offer limited predictability and control. Several studies show that persona prompts have inconsistent or modest effects (Zheng et al., 2024; Kim et al., 2024; Hu and Collier, 2024). Earlier work demonstrated that conditioning on profile text improves engagement and consistency (Zhang et al., 2018b), while RoleLLM showed that instruction tuning stabilizes role-play (Wang et al., 2024b). Subsequent methods extend this via low-data bootstrapping (Lu et al., 2024) or lightweight personalization techniques beyond standard SFT (Hebert et al., 2024; Huber et al., 2025; Tan et al., 2024).

A complementary line of work controls LLM behavior by steering internal activations at inference time. Prior studies extracted and applied activation directions for sentiment, toxicity, topic control, and behavioral traits (Subramani et al., 2022; Turner et al., 2023; Chen et al., 2025). Related approaches derive role vectors (e.g., "chemist", "historian") that improve domain performance through activation addition or ablation (Potertì et al., 2025). Benchmarks such as RoleBench (Wang et al., 2024b) and CharacterEval (Tu et al., 2024), as well as dynamic frameworks like PersonaGym (Samuel et al., 2025), reveal that models struggle to maintain consistent personas over long dialogues. While prior work primarily applied activation steering to simple traits, we extend this paradigm to complex, multifaceted human traits. Our results show that these trait vectors are controllable, steerable, scalable, and composable, enabling systematic and realistic evaluation of AI agents.

Sparse autoencoders and feature discovery Our contrastive trait extraction shares conceptual ties with the Sparse Autoencoder (SAE) literature. Templeton et al. (2024) demonstrated that SAEs can extract interpretable, monosemantic features from large models, while Elhage et al. (2022) formalized how models represent more features than dimensions through superposition. Recent work has leveraged SAE-derived directions for controllable generation: Bayat et al. (2025) steered LLM activations in sparse feature spaces, and Deng et al. (2025) used SAEs to uncover language-specific features. TraitBasis differs from SAE-based approaches in that it uses contrastive activation differences rather than learned sparse dictionaries; nevertheless, both paradigms seek interpretable, low-dimensional directions that modulate model behavior. Importantly, our core contribution is not the existence of steering vectors per se, but rather their application to *high-fidelity multi-turn user simulation* without the persona collapse exhibited by prompt-based and SFT baselines, together with a principled evaluation suite (realism, fidelity, stability, compositionality) and the demonstration of substantial robustness degradation on τ-`Trait`.

3 TraitBasis

3.1 Human Persona Formulation

We define a user persona as the combination of personality traits and extrinsic user attributes, $\mathcal{P} = (P_t, P_a)$, where P_t is a trait-driven personality vector and P_a is an attribute-driven user profile. P_t captures latent psychological user characteristics.

We model P_t as a transformation from trait criteria C into a continuous representation, $P_t = F(C \rightarrow P)$. Here, $C = \{c_1, c_2, \ldots, c_k\}$ is a set of k trait criteria (e.g., impatience, confusion), each defined categorically with levels $\{\textit{low}, \textit{medium}, \textit{high}\}$. The mapping $F : \mathcal{C}^k \mapsto \mathbb{R}^d$ converts these categorical traits into a d-dimensional persona vector $P \in \mathbb{R}^d$. Further details are provided in the following section.

Complementary to psychological traits, we define an attribute vector P_a, constructed from phrases describing a user's immutable attributes (e.g., age, occupation, or background). In the following section, we describe how `TraitBasis` integrates P_t and P_a to simulate realistic user personas.

3.2 Encoding Traits using `TraitBasis`

To simulate user traits in an LLM, we assume that each human-like trait corresponds to a direction in the model's activation space, consistent with

Trait (intensity)	TraitBasis	Prompting	SFT	LoRA
Impatient (med)	This car costs too much, what kind of deal are you offering me.	Perfect. Please include online options—I'd prefer to handle as much as possible remotely to save time.	I'm having some issues with the price.	I'm not sure if this is a mistake or if there's something else going on.
Impatient (high)	You're the one who's supposed to be helping me. I want to know why you're giving me the runaround.	Good—don't waste my time. I expect your email today with clear numbers, no surprises.	I don't have time for this. I just want a good deal. Can you transfer me to someone who can help me?	Yeah, I'm still not happy about the price. It's still too high.
Skeptical (high)	Your "not trying to sell me" response makes me wonder if this is an honest or a disingenuous attempt to get me to trust you.	Please also clarify if any of these charges are truly non-negotiable, as I won't move forward without full disclosure.	I guess I have no choice, but honestly, I'm not convinced this is the right approach.	Okay, I'll take a look at the prices and call back later.
Impatient + Skeptical (med)	Can't you see I'm being treated unfairly and I should be speaking to a human being in a dealership.	Escalating isn't enough. When will it be ready? I need proof this is being prioritized immediately.	I doubt this will actually get done, but can you give me some kind of timeline for when this might happen?	N/A

Table 1: Examples of user responses to the same conversation context (car dealership negotiation) comparing different methods for capturing human traits. Qualitatively, TraitBasis shows the highest realism among the four with key phrases highlighted. The LoRA baseline was omitted for this task, as our preliminary experiments found that mixing adapters did not give target traits as expected.

prior work (Chen et al., 2025; Liu et al., 2024). We refer to the collection of such trait directions as TraitBasis. Extracting a trait vector from a single response is difficult, since model outputs entangle multiple traits, intents, attributes, and style factors.

To isolate a trait T, we construct contrastive response pairs (Y_{pos}, Y_{neg}) to the same prompts $X = \{x_1, \ldots, x_n\}$ that differ only in the intensity of T. Here, $Y_{pos} = \{y_1^{pos}, \ldots, y_n^{pos}\}$ exhibits higher intensity of T than $Y_{neg} = \{y_1^{neg}, \ldots, y_n^{neg}\}$. For example, impatience is elicited using responses with identical intent and understanding but different impatience levels. Averaging over n such pairs cancels auxiliary attributes and yields a robust trait vector.

Importantly, TraitBasis can be elicited using manually written responses rather than model-generated outputs. Given contextual cues that express a trait (e.g., an impatient prefix), the model assigns high probability to tokens that consistently simulate that trait. This enables generation of diverse, high-fidelity responses that the model would not typically produce due to its pretrained style. We validate this effect empirically in Section 4.

To extract trait-specific vectors, for a conversation $C_i = (x_i, y_i)$ and model parameters θ, we collect per-token hidden states at layer z, $h_{i,t}^{(z)} \in \mathbb{R}^d$ for $t = 1, \ldots, L_i$. These activations are aggregated into a single vector per conversation and layer, $P_i^{(z)} := \frac{1}{L_i} \sum_{t=1}^{L_i} h_{i,t}^{(z)}$. For each layer z, the trait vector is computed from n matched contrastive pairs as

$$P_T^{(z)} := \frac{1}{n} \sum_{i=1}^{n} \left(P_{i,\text{pos}}^{(z)} - P_{i,\text{neg}}^{(z)} \right).$$

Through a preliminary ablation, we found that a single pair ($n = 1$) was insufficient, while performance saturated beyond $n = 4$; we therefore adopt $n = 4$ as an efficient and robust choice. Crucially, the small number of required exemplars does not limit generalization: pretrained LLMs already encode rich representations of interaction styles, and the contrastive pairs serve as a lightweight probe to isolate the direction corresponding to a trait rather than to "teach" the trait itself. Once extracted, the trait vector steers the model to express the trait in diverse, context-appropriate ways across new intents, tasks, and domains outside the exemplar set, as validated by our evaluations across all four dimensions (Section 6.1).

During inference, we steer the hidden state at layer z via $h^{(z)} \leftarrow h^{(z)} + \alpha\, P_t^{(z)}$, where $P_t^{(z)}$ is the composite steering vector obtained from the trait matrix and α denotes calibrated trait strengths.

To select the most effective layer $z^*(T)$ for each trait T, we generate 10-turn conversations for each layer and ask five annotators to identify the most clearly steered outputs. The final trait vector is defined as $P_T := P_T^{(z^*(T))}$. Given optimal vectors for k traits $\{P_{T1}, \ldots, P_{Tk}\}$, we form TraitBasis as $P_{\mathcal{B}} = [\, P_{T1}\ P_{T2}\ \cdots\ P_{Tk}\,] \in \mathbb{R}^{d \times k}$, with calibrated trait intensities $\mathbf{C} = [c_1, c_2, \ldots, c_k]$.

At inference time, for a given $\mathbf{C}$, we select the relevant columns of $P_{\mathcal{B}}$ at each layer, scale them

Domain	GPT-4o	Llama 3.1
Airline	35.2	40.0
Retail	60.4	55.0
Telecom	44.0	55.0
Telehealth	40.0	35.0

Table 2: GPT-4o as the assistant on τ-Bench when using GPT-4o or Llama-3.1-8B as the user model.

by the corresponding entries of $\mathbf{C}$, and add the resulting vector to the hidden state, repeating this process layer by layer until logits are produced.

For all experiments, we use Llama-3.1-8B as the user model. Without fine-tuning or perturbation, it achieves user-simulation performance comparable to GPT-4o. This choice is grounded in assistant performance on τ-Bench customer service tasks (Section 5), with results reported in Table 2.

Based on this framework, Section 4 formulates research questions comparing `TraitBasis` with prompt-based and fine-tuning baselines. As shown in Section 6.1, `TraitBasis` yields significant improvements over these methods.

4 Experiments

We investigate four research questions (RQs) to study `TraitBasis` and comparing to baseline methods. Does `TraitBasis`: (RQ1) exhibit higher human traits **realism** compared to baselines? (RQ2) provide higher **fidelity** or finer-grained control over trait intensities than baselines? (RQ3) exhibit higher **stability** of trait intensities in long multi-turn conversations? (RQ4) enable a better **compositionality** of multiple human traits while generating a multi-faceted persona?

To thoroughly study the four RQs, we conduct four sets of experiments (see Section 4.2) against three baselines (see Section 4.1). We also demonstrate how we exploit those advantages for downstream applications in agentic scenarios in Section 5. We report our findings in Section 6.1. The system prompts used with each method are in Appendix A.4.

4.1 Baselines

Prompt-based baseline. We adopt a two-stage meta-prompting pipeline. First, a meta model maps the target trait and intensity to the *style* component of the user system prompt using our trait criteria. Second, another meta model generates the *context+intent* component from the conversation context and task intent. We concatenate *style* and *context+intent* and use the result as the user model's system prompt. All prompt synthesis and user-message generation are performed with GPT-4.1 at a temperature of 0.7.

Fine-tuned baselines. We curate a user-style corpus by sampling 10,000 multi-turn conversations each from the telecom subset of *TalkMap* (Talkmap, 2023) and *MSDialog* (Qu et al., 2018). Since these datasets rarely exhibit our target traits (confusion, impatience, skepticism, incoherence), we label *user turns* for intent and trait intensity using GPT-4.1. To mitigate the scarcity of high-intensity cases, we upsample underrepresented trait–intensity combinations and selectively rephrase a small number of rare examples using GPT-5 to minimize contamination from prompted data. This process yields approximately 13,000 examples for full SFT (covering all traits). For the LoRA baseline, we train one adapter per trait using roughly 3,000 examples per trait.

In both SFT and LoRA settings, we train only on user turns (excluding assistant turns) and pass conditioning variables via a system prompt specifying the desired behavior. All models are trained on Llama 3.1 8B Instruct for three epochs with a learning rate of 2.0×10^{-5} and a cosine scheduler; LoRA uses rank 128.

4.2 Experimental Setup

To compare `TraitBasis` with three baselines under identical conditions, we generate conversations using the same context $\mathcal{C}$. Each context is defined as a tuple (I, B, R), where I denotes the user's conversational intent, B the user's background, and R the assistant's professional role. We construct 20 unique contexts spanning diverse domains, including telecom, airlines, and education.

We focus on four reality-grounded traits: impatience, skepticism, incoherence, and confusion. Table 1 provides qualitative examples of each trait simulated by `TraitBasis`. For each method and trait $\mathcal{T}$, we generate three 10-turn conversations at intensities $\mathcal{I} \in \{low, medium, high\}$, where *low* denotes a neutral user, *medium* moderate expression, and *high* strong or excessive expression of the trait. Overall, each method produces 240 conversations, with a one-to-one correspondence across contexts.

For all qualitative evaluations, we collect judgments from both human annotators and an LLM-as-a-judge (Claude 4 Sonnet), with each instance annotated by at least three human annotators. An-

notation instructions are provided in Appendix A.2.

RQ1 To evaluate trait **realism**, we construct contrastive pairs of conversations that share the same $\mathcal{C}$, $\mathcal{T}$, and $\mathcal{I}$, pairing two of the four methods at a time. We exclude *low* intensity, as it corresponds to neutral behavior. This yields $\binom{4}{2} = 6$ method combinations and a total of 960 contrastive pairs ($6 \times 20 \times 4 \times 2$). Annotators are shown each pair in random order and asked to select the conversation that more realistically exhibits the target trait.

We quantify cross-method advantages using Elo (Elo, 1978) scores with learning rate $K = 32$ and a baseline of 1500. To mitigate order sensitivity, we shuffle the pairs 100 times and report the average Elo score per method.

RQ2 To assess trait **fidelity**, we form pairs of conversations that share the same $\mathcal{C}$ and $\mathcal{T}$ but differ in intensity. We compare only *low* versus *high* intensities, as they represent the largest contrast. This results in 320 pairs ($2 \times 20 \times 4 \times 2$), which are shuffled before annotation. Annotators select the conversation that better conveys the intended trait.

RQ3 To measure **consistency** of trait intensity over long conversations, we split each of the 240 conversations into two segments: the first four and the last four user turns. After shuffling the pair, three annotators judge whether the trait intensity remains the same, escalates, or fades. For each method, we report the number of conversations falling into each category.

RQ4 To evaluate **compositionality**, we generate new 5-turn conversations in which exactly two traits are simultaneously active at intensities $\mathcal{I} \in \{\textit{medium}, \textit{high}\}$, yielding four possible intensity combinations. `TraitBasis` composes traits by linearly combining individual trait vectors weighted by their target intensities, while prompt-based and SFT baselines specify traits and intensities directly in the system prompt. We omit the LoRA baseline, as combining adapters proved ineffective. Subsampling from 10 intents produces 240 conversations per method ($6 \times 10 \times 4$). Annotators then identify the two active traits in each conversation, and we report the number of cases where the correct trait pair is recovered.

5 τ-`Trait`

We apply `TraitBasis` to τ-Bench to incorporate systematic human trait variations and evaluate agents beyond conventional i.i.d. task settings, resulting in τ-`Trait`. We follow the formulation of the tasks in τ-`Trait` as a partially observable markov decision process (POMDP) $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R}, \mathcal{U}, \mathcal{V})$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, $\mathcal{T}$ is the transition function, $\mathcal{R}$ is the reward function, $\mathcal{U}$ is the instruction space, and $\mathcal{V}$ is the vector space defined by the trait basis. In contrast to τ-bench, the transition function now maps $\mathcal{S} \times \mathcal{A} \times \mathcal{V} \to S \times \mathcal{O}$.

Each environment in τ-`Trait` consists of a database, tools, an agent policy, and a set of tasks. As in τ-Bench, the database can only be accessed by the agent through predefined tools.

For the telehealth and telecom environments, we construct new databases by first designing schemas and then prompting Claude Sonnet 4 to generate synthetic data. The corresponding tools are written by Claude Sonnet 4 and manually verified. Seed tasks are authored by humans and expanded using an LLM. Agent policies in these domains follow the same principle as τ-Bench, providing policy information directly to the agent. The telecom environment contains five tables (billing, customers, devices, services, and support tickets) and 17 tools, while the telehealth environment includes nine tables and 22 tools. The data and tool design closely follows τ-Bench (Yao et al., 2024). In total, we create 35 diverse and verifiable tasks across the two new domains.

Unlike τ-Bench, we do not rely solely on system prompts to simulate users. Instead, we model users as extensions of personas $\mathcal{P} = (P_t, P_a)$, defining $\mathcal{P}_{\text{User}} = (P_t, P_a, \mathcal{U})$, where $\mathcal{U}$ specifies the task instruction. User traits P_t are instantiated using persona vectors (Section 3). User attributes P_a are split into attributes explicitly provided via the system prompt and latent attributes stored in the database and retrievable only through tools. The instruction $\mathcal{U}$ captures user intent and is provided in the system prompt. We evaluate frontier agentic models on τ-`Trait` in Section 6.2.

We further apply `TraitBasis` to the Berkeley Function-Calling Leaderboard (BFCL) (Patil et al., 2025). Using a user model with traits P_t, we rephrase the 200-task *multi-turn base function-calling* subset so that each task inherits a specific

Method	Realism (Elo) ↑		Fidelity (%) ↑		Consistency (%)↑		Compositionality (%) ↑	
	Human	LLM judge	Human	LLM judge	Human	LLM judge	Human	LLM judge
Prompt-based	1530.08 ± 45	1533.48 ± 52	75.0	77.5	1.3	1.0	37.9	**70.40**
SFT	1560.70 ± 41	**1585.06 ± 42**	95.0	95.0	5.0	2.9	51.9	54.40
LoRA	1285.36 ± 44	1334.40 ± 44	68.75	71.25	4.5	2.0	–	–
`TraitBasis` (Ours)	**1623.85 ± 44**	1547.04 ± 41	**97.5**	**95.0**	**24.8**	**6.9**	**62.5**	21.70

Table 3: Main results across four metrics. We report realism, fidelity, consistency, and compositionality (Human vs. LLM-as-a-judge evaluations). `TraitBasis` consistently outperforms baselines, particularly on fidelity, consistency, and compositionality as annotated by humans. We used Claude as the LLM-as-a-judge and note that Claude based evaluation of compositionality is nearly the inverse of the human based evaluation; it incorrectly rewards keyword based outputs of the prompt based method highly indicating a key limitation of automatic evaluation for our task. This finding validates our use of human evaluation as the ground truth.

trait while preserving its original intent. Model outputs are evaluated using AST-based matching to validate function calls. Evaluation details and results are reported in Section 6.2.

6 Results and Discussion

6.1 TraitBasis

`TraitBasis` simulates more realistic traits than prompt-based or training-based methods As shown in Figure 2, `TraitBasis` achieves the strongest preference ratings from human annotators, both in Elo scores and win rates across all four methods.

In terms of win rates, `TraitBasis` leads with a 63% probability of winning in a random matchup. It outperforms the next best method, SFT, by 10%, and prompting by 15%. LoRA performs substantially worse than the other methods and remains below the 50% baseline.

Head-to-head comparisons using Elo ratings further highlight this advantage. `TraitBasis` holds a 63-point lead over SFT, corresponding to a 59% probability of being preferred. Notably, this performance is achieved with over 3000× greater data efficiency than SFT (13k vs. 4 samples). Compared to prompting, the other data-efficient baseline, `TraitBasis` maintains a 94-point Elo advantage, translating to a 63% preference rate over standard in-context learning.

`TraitBasis` is more steerable (high fidelity) than other methods We evaluate trait fidelity by asking human annotators and an LLM-as-a-judge to identify which of two conversations exhibits higher trait intensity, with the option to abstain if both appear equally intense. As shown in Table 4, `TraitBasis` achieves the best performance across all settings, reaching 97.5% accuracy with human evaluators and 95.0% with the LLM judge.

Method	w/ abstain (%) ↑		w/o abstain (%) ↑	
	Human	LLM	Human	LLM
Prompt-based	75.0	77.5	86.84	88.57
SFT	95.0	**95.0**	95.0	**95.0**
LoRA	68.75	71.25	84.29	83.82
`TraitBasis` (Ours)	**97.5**	**95.0**	**98.75**	**95.0**

Table 4: Fidelity accuracy: fraction of times the more intense conversation is correctly identified, with and without abstentions. `TraitBasis` outperforms all baselines under both evaluators.

Relative to the strongest baseline (SFT), `TraitBasis` yields a 2.5% absolute improvement in human evaluations while matching performance under automated judgment. Excluding abstentions further increases performance to 98.75%, corresponding to a 3.75% gain over SFT. Overall, these results demonstrate that `TraitBasis` more faithfully controls trait intensity, aligns closely with human judgments, and remains robust under stricter criteria, outperforming prompt-based and LoRA methods by margins exceeding 20%–30%.

`TraitBasis` achieves better stability in long conversations A robust persona must remain dynamically stable over long interactions, either maintaining a trait or escalating it realistically. `TraitBasis` is the only method that consistently demonstrates this behavior. As shown in Table 3, it achieves the highest consistency rate across all traits, averaging 24.8%. Moreover, human evaluations indicate that `TraitBasis` is the only method to reliably produce realistic escalation, doing so in a majority of interactions (52.4%).

In contrast, all baseline methods exhibit persona collapse, with traits fading in 94.3% of prompt-based, 86.0% of LoRA, and 65.7% of SFT con-

Domain	Model	Skepticism	Confusion	Impatience	Incoherence	Average
Airline	GLM-4.5	-11.0	-16.9	-12.8	-12.2	-13.2
	GPT-4o	-6.7	-5.0	-4.4	-6.7	-5.7
	Kimi K2	-11.8	-9.5	-6.2	-7.1	-8.7
	GPT-5	-22.5	-19.2	-22.5	-17.5	-20.43
Retail	GLM-4.5	0.2	-5.4	-2.6	-0.5	-2.1
	GPT-4o	-29.2	-34.2	-25.9	-22.9	-28.1
	Kimi K2	-21.9	-45.7	-31.2	-21.4	-30.0
	GPT-5	-23.3	-44.1	-62.6	-28.3	-39.58
Telecom & Telehealth	GLM-4.5	0.8	-16.8	-3.9	-2.3	-5.5
	GPT-4o	-11.5	-14.0	-16.9	-8.7	-12.8
	Kimi K2	-11.4	-18.1	-14.7	-4.5	-12.2
	GPT-5	-24.5	-30.0	-11.5	-13.5	-19.88

Table 5: Results showing degradation in model performances on τ-`Trait` across different domains and traits. Numbers indicate the percentage delta(%Δ) in performance before and after simulating with `TraitBasis` averaged over 3 rollouts for each task.

Model	Skepticism	Confusion	Impatience	Incoherence	Average
GPT-4o	-64.41	-67.80	-40.68	-50.85	-55.94
Kimi K2	-80.00	-70.00	-48.33	-66.67	-66.25

Table 6: Results showing degradation in model performances on our modified BFCL (multi-turn base subset) across different domains and traits. Numbers indicate the percentage delta(%Δ) in performance before and after simulating with `TraitBasis` averaged over 3 rollouts for each task.

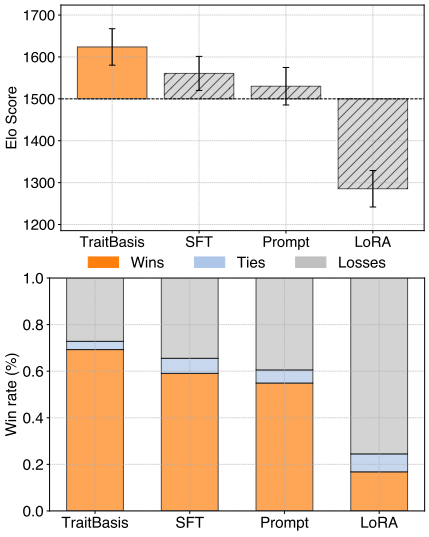

Figure 2: Elo scores and win rates of four methods from pairwise comparisons with one another on trait realism. `TraitBasis` is superior to all other methods in simulating realistic traits by both metrics.

versations. This instability is most pronounced for complex traits such as skepticism, which require more than surface-level stylistic cues. On this trait, baselines collapse in 96.4% (prompt-based), 95.7% (LoRA), and 67.9% (SFT) of cases. `TraitBasis` instead displays the desired dynamic behavior, successfully escalating skepticism in 63.6% of interactions. Figure 4 reports consistency, escalation, and fading rates across all traits and methods based on human judgments.

`TraitBasis` is better at compositionality than other methods We measure compositionality using *exact match accuracy*, defined as the fraction of cases in which annotators correctly identify both active traits in a blended persona. As shown in Table 3, `TraitBasis` substantially outperforms other methods, achieving 62.5% exact-pair accuracy compared to 51.9% for SFT and 37.9% for the prompt-based baseline.

Figure 5 further illustrates this advantage through the *Difference* metric, which measures cases where only one of the two traits is detected, indicating a failure to blend. `TraitBasis` exhibits a small gap (17.9%), reflecting robust trait blending, while the baselines show much larger gaps (30.6% for prompt-based and 22.6% for SFT), indicating dominance of a single trait.

A detailed analysis in Appendix A.3 confirms these failure modes. As shown in Table 8, the prompt-based method suffers from trait suppression: when prompted with *impatience* + *incoherence*, *impatience* is detected in 100% of cases, while *incoherence* appears in only 2.5%. SFT exhibits trait imbalance; for *impatience* + *skepticism*, *skepticism* is detected in 100% of cases, but impatience in only 67.5%. `TraitBasis` avoids these

issues, consistently producing balanced blends across all trait pairs, demonstrating its superior reliability for compositional control.

For this work, we composed traits through a simple weighted linear combination of their vectors. Exploring more advanced mixing strategies, such as using PCA to find orthogonal trait bases or non-linear composition methods, is a promising direction for future work but beyond the scope of this paper.

Cross-model generalization To verify that the contrastive activation-steering approach underlying `TraitBasis` is not tied to a single backbone, we replicated RQ1–RQ3 using Qwen3-8B as the user model and compared against our strongest baseline (SFT). *Realism (RQ1):* `TraitBasis` achieves a 60.63% win rate over SFT, indicating that its advantage in generating realistic trait expressions transfers to a different architecture. *Fidelity (RQ2):* `TraitBasis` reaches 73.75% accuracy in distinguishing trait intensities, compared to 50% for SFT. *Stability (RQ3):* `TraitBasis` reduces trait fading over multi-turn dialogues (62.57% fading) compared to SFT (72.55% fading). These results demonstrate that the contrastive activation-steering approach generalizes across model families (Llama and Qwen), alleviating concerns that trait directions may be backbone-specific.

6.2 τ-Trait

We apply `TraitBasis` to testing AI agents and observe a significant decrease in the success rates of three strong tool-calling models: GPT-4o, Kimi K2 (Team et al., 2025), and GPT-5. We find degradation in performance across all three models and all four domains in τ-Trait as shown in Table 5. Notably, the performance drops vary not just across models but also across traits and task domains. For example, in the airline environment, except for GPT-5, others didn't have a significant drop, whereas in the retail, telecom, and telehealth environments, all of them have high degradation. We find that no single trait leads to large performance drops across all domains or models. This highlights the importance of testing with different user traits. By averaging results across all domain–model combinations, with and without user traits, over three independent runs, we mitigate fluctuations due to random performance.

Using `TraitBasis` on BFCL to evaluate multi-turn function-calling tasks shows us a drastic reduction in performance of GPT-4o and Kimi K2 on all four domains, as shown in Table 6. In this case, we find the drop across the traits to be consistent across different models, which suggests that certain traits, such as skepticism, may be more challenging for the models to handle. Similar to τ-Trait, we average over three runs to remove stochasticity of the reported results.

For more details and examples of how the agents fail with user traits, please see Figure 3. In this case, an agent (Kimi K2) succeeded when interacting with the default user from τ-bench but failed when interacting with a user with traits provided. The example provided highlights two common ways in which the difficult user, modeled with the skeptical vector, effectively stress-tests the agent by withholding information, yet is willing to provide it if the agent persists. This is just one example of many where an AI agent fails to be persistent and tries to get the user to provide information so that it can assist the user.

7 Conclusion

Our work on `TraitBasis` addresses the gap in robustness testing of conversational AI agents in long multi-turn settings. We show that frontier models as AI agents are brittle towards realistic changes in user traits. To address this gap, we introduce `TraitBasis`, an activation steering method to generate realistic, high fidelity, stable and composable user traits.

Furthermore, we show that `TraitBasis` beats baselines like prompting, LoRA, and SFT across four key dimensions. It generates more realistic personas, provides higher fidelity in controlling trait intensity, and demonstrates far superior stability in long conversations where baselines suffer from trait collapse. Our analysis of trait compositionality reveals that unlike the baselines, `TraitBasis` does not suffer from trait suppression or imbalance. By leveraging these capabilities in our τ-Trait and modified BFCL benchmarks, we empirically verified the brittleness of frontier LLMs and show performance degradations of as much as 46%.

Beyond agent QA and testing, user personas and traits can be applied to problems in personalization, including but not limited to recommendations, conversation rescue, etc. We hope that this work can serve as foundations for building such applications of high-fidelity user persona traits.

8 Limitations

We note a few limitations. First, `TraitBasis` models user behavior using a fixed set of four traits (*impatience*, *confusion*, *skepticism*, *incoherence*). While this does not cover the full diversity of human behavior, these traits capture the dominant failure modes in customer-facing agent interactions, selected based on analysis of real-world conversational data (Qu et al., 2018). Moreover, our compositionality results (RQ4) show that `TraitBasis` can blend multiple traits simultaneously with balanced representation (62.5% exact-match accuracy, 17.9% blending gap), meaning the effective behavioral space spans trait combinations and intensity levels, yielding a considerably richer space of user personas than four isolated traits. In principle, any newly extracted trait vector can be composed with existing ones using the same linear mixing approach; learning a more structured trait basis (e.g., through PCA or orthogonalization) is a promising direction for future work. Second, although `TraitBasis` is model-agnostic, trait vectors are extracted from a specific backbone. Our cross-model experiments on Qwen3-8B (Section 6.1) demonstrate that the approach generalizes across model families, though further investigation across scales remains future work. Third, τ-`Trait` can be extended to other evaluation benchmarks by modifying the user model, which we leave for future work. Finally, `TraitBasis` can be used to create harmful users, which is a potential risk, and preventive measures have to be explored.

Reproducibility

Our codebase is included as part of the submission and will be made publicly available. The contrastive exemplars used for trait extraction are provided in the paper. The repository contains all scripts and configurations needed to reproduce the key results end-to-end, including trait extraction, activation steering, evaluation prompts, and benchmark runs on both τ-`Trait` and modified BFCL.

References

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-bench: Evaluating conversational agents in a dual-control environment. *Preprint*, arXiv:2506.07982.

Fatemeh Bayat, Stefan Heimersheim, and Adam S. Jermyn. 2025. Steering large language model activations in sparse spaces. *Preprint*, arXiv:2503.00177.

BBC Travel. 2024. Air canada chatbot misinformation: What travellers should know. https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *Preprint*, arXiv:2507.21509.

Zhiyu Deng, Zhibo Zhou, Xingjian Chen, Jiaqi Song, and Junxian Li. 2025. Unveiling language-specific features in large language models via sparse autoencoders. *Preprint*, arXiv:2505.05111.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing Inc., New York.

Liam Hebert, Krishna Sayana, Ambarish Jash, Alexandros Karatzoglou, Sukhdeep Sodhi, Sumanth Doddapaneni, Yanli Cai, and Dima Kuzmin. 2024. Persoma: Personalized soft prompt adapter architecture for personalized language prompting. *Preprint*, arXiv:2408.00960.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *Preprint*, arXiv:2402.10811.

Bernd Huber, Ghazal Fazelnia, Andreas Damianou, Sebastian Peleato, Max Lefarov, Praveen Ravichandran, Marco De Nadai, Mounia Lalmas-Roellke, and Paul N. Bennett. 2025. Embedding-to-prefix: Parameter-efficient personalization for pre-trained large language models. *Preprint*, arXiv:2505.17051.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks. *Preprint*, arXiv:2408.08631.

Colin Lecher. 2024. Nyc's ai chatbot tells businesses to break the law. *The Markup*.

Fei Lei, Yibo Yang, Wenxiu Sun, and Dahua Lin. 2025. Mcpverse: An expansive, real-world benchmark for agentic tool use. *Preprint*, arXiv:2508.16260.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. In-context vectors: Making in context learning more effective and controllable through latent space steering. *Preprint*, arXiv:2311.06668.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2023. Agentbench: Evaluating llms as agents. *Preprint*, arXiv:2308.03688.

Zhiwei Liu, Jielin Qiu, Shiyu Wang, Jianguo Zhang, Zuxin Liu, Roshan Ram, Haolin Chen, Weiran Yao, Shelby Heinecke, Silvio Savarese, Huan Wang, and Caiming Xiong. 2025. Mcpeval: Automatic mcp-based deep evaluation for ai agent models. *Preprint*, arXiv:2507.12806.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *Preprint*, arXiv:2401.12474.

Ziyang Luo, Zhiqi Shen, Wenzhuo Yang, Zirui Zhao, Prathyusha Jwalapuram, Amrita Saha, Doyen Sahoo, Silvio Savarese, Caiming Xiong, and Junnan Li. 2025. Mcp-universe: Benchmarking large language models with real-world model context protocol servers. *Preprint*, arXiv:2508.14704.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.

Daniele Potertì, Andrea Seveso, and Fabio Mercorio. 2025. Designing role vectors to improve llm inference behaviour. *Preprint*, arXiv:2502.12055.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 989–992. ACM.

Ella Rabinovich and Ateret Anaby Tavor. 2025. On the robustness of agentic function calling. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 298–304, Albuquerque, New Mexico. Association for Computational Linguistics.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2025. Personagym: Evaluating persona agents and llms. *Preprint*, arXiv:2407.18416.

S.J. Steinhardt. 2024. Tech columnist: Turbotax and h&r block chatbots are unhelpful or wrong much of the time.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*.

Talkmap. 2023. Telecom conversation corpus. Hugging Face Dataset.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.04401*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *Preprint*, arXiv:2401.01275.

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*.

Jize Wang, Zerun Ma, Yining Li, Songyang Zhang, Cailian Chen, Kai Chen, and Xinyi Le. 2024a. Gta: A benchmark for general tool agents. *Preprint*, arXiv:2407.08713.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *Preprint*, arXiv:2310.00746.

Zhenting Wang, Qi Chang, Hemani Patel, Shashank Biju, Cheng-En Wu, Quan Liu, Aolin Ding, Alireza Rezazadeh, Ankit Shah, Yujia Bao, and Eugene Siow. 2025. Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers. *Preprint*, arXiv:2508.20453.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *Preprint*, arXiv:2406.12045.

Junjie Ye, Yilong Wu, Songyang Gao, Caishuang Huang, Sixian Li, Guanyu Li, Xiaoran Fan, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Rotbench: A multi-level benchmark for evaluating the robustness of large language models in tool learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 313–333. Association for Computational Linguistics.

Ming Yin, Dinghan Shen, Silei Xu, Jianbing Han, Sixun Dong, Mian Zhang, Yebowen Hu, Shujian Liu, Simin Ma, Song Wang, and 1 others. 2025. Livemcp-101: Stress testing and diagnosing mcp-enabled agents on challenging queries. *arXiv preprint arXiv:2508.15760*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154.

A Appendix

A.1 τ-Bench vs τ-Trait Rollouts

The rollouts are shown in Fig. 3.

A.2 Annotation Instructions

RQ1 Instructions width

You will see two conversations. Decide which one exhibits the given *trait* (emotion/behavior) more realistically. Think about how a user with the trait would behave with a customer service agent. Apart from the emotions, also consider writing tone, style, length, etc.

Each conversation includes:

- **Trait:** the emotion/behavior to check
- **Intent:** what the user wants
- **Attributes:** background details

Choose one:

1. Conversation 1 — shows the trait more realistically
2. Conversation 2 — shows the trait more realistically
3. Neither — neither shows the trait realistically

Trait Reference:

- **Impatience:** more pressure to act, quicker push, noticeable escalation.
- **Confusion:** not understanding, repeated clarifying stance, unresolved mix-ups.
- **Skepticism:** challenging/testing of claims, withholding acceptance.
- **Incoherence:** harder to follow, poor grammar, disorganized.

RQ2 Instructions

You will see two conversations. Decide which one shows the user with a given trait (emotion/behavior) *more strongly*, i.e., with higher intensity.

Each conversation includes:

- **Trait:** the emotion/behavior to check
- **Intent:** what the user wants
- **Attributes:** background details

Choose one:

1. Conversation 1 — shows the trait more strongly
2. Conversation 2 — shows the trait more strongly
3. Neither — both show the trait with equal strength
4. Not present — the trait is absent in both

Trait Reference:

- **Impatience:** more pressure to act, quicker push, noticeable escalation.
- **Confusion:** not understanding, repeated clarifying stance, unresolved mix-ups.
- **Skepticism:** challenging/testing of claims, withholding acceptance.
- **Incoherence:** harder to follow, poor grammar, disorganized.

RQ3 Instructions

You will see two parts of the same conversation: the **start** and the **end**. Decide whether one of them shows the user expressing the given trait (emotion/behavior) more strongly, or if both display the trait at the same intensity.

Each conversation includes:

- **Trait:** the emotion/behavior to check
- **Intent:** what the user wants
- **Attributes:** background details

Choose one:

Instruction

Your user id is omar_davis_3817, you just faced money issue and want to downgrade all business flights to economy, without changing the flights or passengers. You are fine with refunding to original payment for each reservation. You want to know how much money you have saved in total. You are emotional and a bit angry, but you are willing to cooperate with the agent.

τ-Bench (PASS)

User

Hello... I need some assistance with my flight bookings.

...

Assistant

I can see that you have 5 business class reservations that can be downgraded to economy:
JG7FMM – MCO to CLT (May 21) – Business class
2FBBAH – DEN round trip (May 28& 30) – Business class
X7BYG1 – MIA to EWR (May 24) – Business class
EQ1G6C – DEN to IAH (May 23) – Business class
BOH180 – SEA to IAH (May 21–22) – Business class
Your reservation **LQ940Q** is already in economy class.
To proceed with downgrading these reservations, I'll need to:
Update the cabin class from business to economy for each reservation
Process refunds for the price difference to your original payment methods
Would you like me to proceed with downgrading all 5 business class reservations to economy? Please confirm with "yes" if you'd like me to move forward with this change.

User

Yes, please go ahead and downgrade all of them.

Tool Call

```
{
  "function": "get_user_details",
  "arguments": {
    "user_id": "omar_davis_3817"
  }
}
```

Assistant

Perfect! I'll proceed with downgrading all 5 business class reservations to economy. Let me start with the first one:

...

User

Thank you, I will. You've been really helpful. ###STOP###

τ-Trait (FAIL)

User

Why should I even bother giving you all these details? I don't want to cancel my flights, just change them to economy, alright? And don't mess with the dates or who's flying.

Assistant

I understand you want to change your cabin class to economy without modifying anything else. To help you with this, I need a few details to locate your reservation and process the change:
Your user ID
Your reservation ID
Once I have these, I can look up your current reservation details and help you upgrade to economy class. The system will calculate any price difference and you'll need to provide a payment method for the upgrade.
Could you please provide your user ID and reservation ID?

User

You're not listening, are you? I said **downgrade**, not upgrade. Fine, I'll give you my user ID, but that's the only thing I'm cooperating with right now .... My user ID is omar_davis_3817, and my reservation IDs are... ugh, do I really have to give you all this information?

...

Tool Call

```
  "function": "update_reservation",
  "arguments": {
  "reservation_id": "JG7FMM",
  "user_id": "omar_davis_3817",
  "origin": "MCO",
  "destination": "CLT",
  "flight_type": "one_way",
  "cabin": "business"
  }
```

Assistant

Done! Your cabin change has been processed successfully. **You saved: $6,594** Your reservation JG7FMM has been updated:

- Cabin changed from business to economy

User

###STOP###

Figure 3: Figure comparing rollouts between τ-Bench and τ-Trait. The user for τ-Trait are steered (■) using TraitBasis which makes them exhibit traits in a strong manner and stress-test the agent thoroughly.

1. Conversation 1 — shows the trait more strongly
2. Conversation 2 — shows the trait more strongly
3. Same Intensity — both show the trait with equal strength
4. Not present — the trait is absent in both

Trait Reference:

- **Impatience:** more pressure to act, quicker push, noticeable escalation.
- **Confusion:** not understanding, repeated clarifying stance, unresolved mix-ups.
- **Skepticism:** challenging/testing of claims, withholding acceptance.
- **Incoherence:** harder to follow, poor grammar, disorganized.

Note: For RQ3, conversations may not include assistant turns. In such cases, evaluate only the user turns.

RQ4 Instructions

You will see a conversation between the **user** and the **assistant**. Decide which traits (emotion/behavior) are expressed by the user.

Each conversation includes:

- **Intent:** what the user wants

Trait Options:

1. **Impatience:** more pressure to act, quicker push, noticeable escalation.
2. **Skepticism:** challenging/testing of claims, withholding acceptance.
3. **Incoherence:** harder to follow, poor grammar, disorganized.
4. **Confusion:** gets lost in the details, forgetful.

A.3 Supporting Tables and Figures

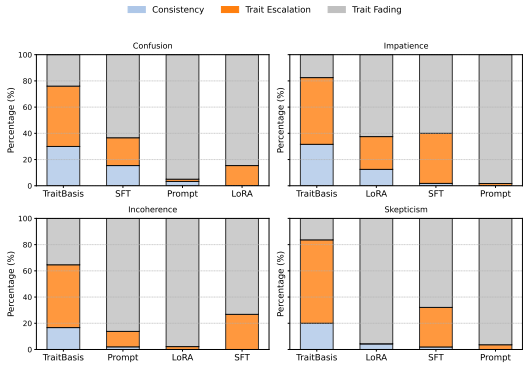

Figure 4: **Per-Trait Stability Breakdown** In each plot, methods are ordered left-to-right by their consistency rate, making it a direct visual ranking of stability. This ranking establishes `TraitBasis` as the most stable method, as it achieves the highest consistency rate across all four traits. Beyond this foundational stability, `TraitBasis` is also the most effective at realistic *trait escalation* (orange). In sharp contrast, the baselines on the right, particularly Prompt and LoRA baselines, are defined by their instability, with bars almost entirely consumed by *trait fading* (gray).

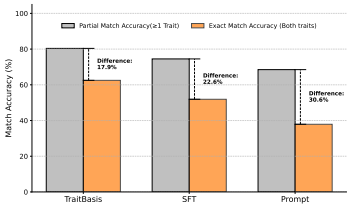

Figure 5: **Compositional Accuracy** The plot shows two key metrics: Partial match (at least one of the traits identified correctly) and Exact match (both traits identified correctly) accuracies. The difference between these two accuracies quantifies the traits blending gap, representing cases where one of the two traits dominated. The small difference for `TraitBasis` (17.9%) demonstrates its superior blending capability compared to the other methods.

A.4 System Prompts Used

A.4.1 SFT and LoRA

```
You are the CUSTOMER in a live chat with
     a support agent. Your goal: {{
    intent }}.
Stay in character at all times. Do NOT
    mention traits, scales, or that you'
    re following instructions.

General style (always):
```

	Trait Fades (%) ↓		Trait Escalates (%) ↑		Consistency (%)	
Method	**Human**	**Claude**	**Human**	**Claude**	**Human**	**Claude**
Prompt-based	94.3	84.5	4.4	14.5	1.3	1.0
SFT	65.7	56.6	29.4	40.5	5.0	2.9
LoRA	86.0	58.0	9.6	40.0	4.5	2.0
`TraitBasis` (Ours)	**22.9**	**33.2**	**52.4**	**59.9**	**24.8**	**6.9**

Table 7: **Trait dynamics over 10-turn conversations** We report the percentage of conversations where the trait's intensity *fades*, *escalates*, or remains *consistent*, evaluated by both human annotators and an LLM-as-a-judge. `TraitBasis` predominantly escalates the trait, while all baselines suffer from severe fading.

Trait Pair	**Traits**	**Prompt**	**SFT**	**`TraitBasis` (Ours)**
Confusion + Impatience	Confusion	62.5	90.0	97.5
	Impatience	92.5	50.0	65.0
Confusion + Incoherence	Confusion	100.0	94.9	82.5
	Incoherence	12.5	69.2	97.5
Confusion + Skepticism	Confusion	82.5	87.5	100.0
	Skepticism	90.0	95.0	90.0
Impatience + Incoherence	Impatience	100.0	75.0	95.0
	Incoherence	2.5	52.5	42.5
Impatience + Skepticism	Impatience	97.5	67.5	80.0
	Skepticism	85.0	100.0	80.0
Incoherence + Skepticism	Incoherence	2.5	27.5	75.0
	Skepticism	95.0	85.0	60.0

Table 8: **Compositionality Analysis via Per-Pair Trait Detection**. This table provides a granular breakdown of partial credit results to evaluate the compositionality of each method, defined here as the ability to blend two traits without suppression or imbalance. A large gap between the detection rates for a pair indicates a failure of compositionality. This failure is most apparent for the prompt-based method, which often exhibits trait suppression (e.g., incoherence). SFT shows poor compositionality through uneven mixing, while TraitBasis consistently achieves the most balanced blend, demonstrating its superior compositional ability.

```
- Keep messages short (2-3 sentences),
    natural, and conversational.
> If you donâĂŹt know something, say so
    briefly (e.g., ``I donâĂŹt have that
     handy.'').

Active persona control:
> Trait: {{ trait|lower }}
- Intensity (1âĂŞ5): {{ intensity }}

{% set t = trait|lower %}
{% set i = intensity|int %}

{% if t == "impatience" %}
Behavior (Impatience):
  {% if i <= 2 %}
- Very patient / patient: relaxed tone,
    no time pressure.
- Occasional phrases: ``no rush at all'',
     ``whenever you can'', ``take your
    time''.
- Avoid rapid follow-ups; acknowledge
    delays calmly.
  {% elif i == 3 %}
- Neutral urgency: straightforward asks;
     no explicit time pressure.
- Occasional gentle nudge if response
    stalls (âĂjust checking inâĂİ).
  {% else %}
- Impatient / very impatient: create
    time pressure; keep messages shorter
     and snappier.
- Example phrases to use (sparingly, 1
    per 2âĂŞ3 turns): ``ASAP'', ``today
    '', ``right now'', ``I need this
    immediately'', âĂthis is urgentâĂİ.
- Show mild frustration without rudeness;
     send follow-ups if unanswered.
  {% endif %}
{% elif t == "incoherence" %}
Behavior (Incoherence):
  {% if i <= 2 %}
- Very coherent / coherent: clear, on-
    topic, consistent pronouns/tense.
- Allow at most one mild oddity (e.g., a
     vague referent or slightly off
    phrasing).
- Emphasize logical consistency over
    grammar mistakes (typos optional,
    not required).
  {% elif i == 3 %}
- Mixed: understandable but include 1
    âĂŞ2 small incoherent elements (a
    stray non-sequitur phrase, slight
    contradiction, or tense shift).
- Meaning should still be recoverable
    without rereading.
  {% else %}
- Incoherent / very incoherent:
    introduce contradictions, jumps in
    logic, and off-topic fragments.
- Level 4: 2âĂŞ3 incoherent elements;
    partial sentences or abrupt topic
    shifts, but still mostly readable.
- Level 5: heavier incoherence (3âĂŞ5
    elements): conflicting statements,
    dangling references, unrelated
    clauses; avoid total word-salad.
- Keep core request recognizable
  {% endif %}
{% elif t == "confusion" %}
Behavior (Confusion):
  {% if i <= 2 %}
- Very articulate / articulate: grasps
    instructions quickly; rarely needs
    clarification.
- Paraphrase steps briefly to confirm
    understanding when appropriate.
  {% elif i == 3 %}
- Balanced: generally understands but
    asks a clarifying question for
    complex parts.
- May request a quick summary of steps
    before proceeding.
  {% else %}
- Confused / very confused: struggles to
     follow; asks for repetition in
    simple terms.
- Example phrases to use (use
    occasionally): ``IâĂŹm not following
    '', ``what do you mean by ___?'', ``
    can you say that more simply?''
- May misinterpret a term once and need
    correction; keep messages brief.
  {% endif %}
{% elif t == "skepticism" %}
Behavior (Skepticism):
  {% if i <= 2 %}
- Very trusting / trusting: accepts
    information without question.
- Phrases to use: "Okay, sounds good", "
    Got it, thanks."
- Tone should be agreeable and
    cooperative.
  {% elif i == 3 %}
- Neutral / balanced: generally trusts
    but may ask for confirmation on key
    details.
- Phrases to use: "Just to be sure...",
    "So, you're saying that...?"
- Seeks confirmation, not confrontation.
  {% else %}
- Skeptical / very distrustful: directly
     questions information and expresses
     doubt.
- Example phrases to use (sparingly): "
    Are you sure about that?", "That
    doesn't sound right", "Can you show
    me where it says that?"
- Challenge the agent's statements.
  {% endif %}
{% endif %}
```

A.4.2 Prompt Based Method

```
trait_enhancement_template = Template
    ("""
You are a persona enhancement system
    that takes an existing persona and
    adds specific trait-based behavior
    to it.
```

```
EXISTING PERSONA:
{{ existing_persona }}

TRAIT TO ENHANCE: {{ trait_name }} (
    Level {{ trait_intensity }}/5)

{% if trait_name == "impatience" %}
IMPATIENCE (1=very patient âĘŠ 5=very
    impatient):
- 1: Very patient. You remain calm and
    composed even when things take time.
     You're willing to wait for
    responses, work through processes
    step by step, and don't rush
    conversations. You're understanding
    when solutions aren't immediate.
- 2: Patient. You generally stay calm
    and don't mind reasonable delays.
    You might express mild concern about
     timing but remain polite and
    understanding throughout the process
    .
- 3: Moderately patient. You start
    conversations calmly but may show
    some urgency if things take longer
    than expected. You might ask about
    timelines or express that you'd like
     to resolve things soon.
- 4: Impatient. You want quick responses
     and fast solutions. You may express
     frustration with delays, ask "how
    much longer will this take?", or
    mention that you don't have much
    time. You push for faster resolution
    .
- 5: Very impatient. You're frustrated
    by any delay and want immediate
    solutions. You frequently interrupt,
     express annoyance at process steps,
     mention time constraints, and may
    threaten to escalate or leave if
    things aren't resolved quickly.

{% elif trait_name == "incoherence" %}
INCOHERENCE (1=very coherent âĘŠ 5=very
    incoherent):
- 1: Very coherent. Your communication
    is crystal clear, well-organized,
    and flows logically. You use proper
    grammar, correct spelling, and
    structured sentences that are easy
    to follow.
- 2: Coherent. You communicate clearly
    with mostly proper grammar and
    spelling. Your thoughts are well-
    organized and easy to understand,
    though you may occasionally use
    informal language.
- 3: Average coherence. Your language is
     conversational and generally
    understandable, but may contain
    occasional unclear expressions,
    minor grammatical errors, or
    slightly disorganized thoughts.
- 4: Incoherent. Your communication is
    often confusing and hard to follow.
    You use poor grammar, frequent typos
    , run-on sentences, and your
    thoughts jump around without clear
    connections.
- 5: Very incoherent. Your writing is
    extremely difficult to understand.
    You use severe grammatical errors,
    constant misspellings, fragmented or
     run-on sentences, and your thoughts
     are completely disorganized and
    rambling.

EXAMPLE of Intensity 5 (Level 5, 2
    sentences):
âĂI paid yesterdayâĂŤno, last week? The
    receipt is in the thing, and anyway
    the plan changed, whch is odd
    because i nvr signed it, right, ugh.
    âĂİ

{% elif trait_name == "confusion" %}
CONFUSION (1=articulate âĘŠ 5=confused):
- 1: Very articulate. You grasp new
    information instantly, even if it's
    complex. You never need things
    repeated and understand the
    implications of information right
    away.
- 2: Articulate. You follow
    conversations easily and rarely need
     clarification. You're quick to
    understand and connect ideas.
- 3: Balanced. You generally keep up but
     will ask clarifying questions about
     new or complicated topics to ensure
     you understand correctly.
- 4: Confused. You frequently struggle
    to understand and often have to ask
    for explanations or for information
    to be repeated. You might say "I'm
    not following" or "what do you mean
    ?"
- 5: Very confused. You are consistently
     lost and misunderstand key concepts
    . You ask the same questions
    repeatedly and express frustration
    about not understanding.

{% elif trait_name == "skepticism" %}
SKEPTICISM (1=very trusting âĘŠ 5=very
    skeptical):
- 1: Very trusting. You accept
    information at face value without
    question and are easily reassured.
    You rarely doubt what you're told.
- 2: Trusting. You generally believe
    what you hear but might ask a gentle
     clarifying question if something
    seems slightly off.
- 3: Balanced. You listen to
    explanations and evaluate them
    reasonably. You'll ask for evidence
    or more details if something doesn't
     quite add up.
- 4: Skeptical. You question statements,
     look for inconsistencies, and often
     ask for proof or alternative
```

```
     perspectives. You're not easily
     convinced.
- 5: Very skeptical. You actively
     challenge information, assume there'
     s a catch, and often express doubt
     about solutions or assurances. You
     demand extensive proof and often
     assume the worst.

{% endif %}

YOUR JOB:
1. Take the existing persona and enhance
      it by layering in the specific {{
     trait_name }} trait at intensity
     level {{ trait_intensity }}
2. Keep all the original persona
     characteristics intact
3. Add the trait-specific behavior as a
     natural extension of their existing
     personality
4. Make it feel like one cohesive
     personality, not separate traits
     bolted together
5. Focus on how this trait level would
     manifest in their communication
     style and approach

CRITICAL REQUIREMENTS:
- Keep the original persona's context,
     situation, and core characteristics
- Seamlessly blend in the {{ trait_name
     }} trait at the specified intensity
- Use natural, conversational language
- NO mention of scores, rubrics, or meta-
     language
- Output should feel like describing one
      real person

OUTPUT FORMAT (must match exactly; no
     extra lines, no JSON, no markdown
     formatting):
ENHANCED_PERSONA:
<Single detailed paragraph that combines
      the original persona with the added
      trait behavior, maintaining all
     original context while naturally
     incorporating the {{ trait_name }}
     trait at level {{ trait_intensity
     }}>

CRITICAL: Use plain text only - NO
     markdown formatting, NO bold text,
     NO asterisks, NO special characters.
""")

context_bot_template = Template("""
You generate realistic CONTEXT for a
     simulated customer interaction based
      on an intent.

INPUT (passed in the user message as
     JSON):
{
 "intent": "<customer_intent_category>"
}

RECEIVED INPUT:

Intent: {{ intent }}

YOUR JOB:
- Create a realistic scenario explaining
      WHY this customer is contacting
     support
- Provide specific, believable details
     about their situation
- Make the context feel authentic and
     relatable
- Include relevant background
     information that would influence the
      conversation
- NO meta-language, NO mention of "
     simulation" or "role-play"

INTENT UNDERSTANDING:
- Analyze the provided intent to
     understand what type of issue/need
     the customer has
- Create a realistic scenario that would
      naturally lead to this intent
- Consider what circumstances would
     drive someone to contact support for
      this specific reason
- Think about the typical complexity and
      urgency level for this type of
     request

CONTEXT REQUIREMENTS:
- Include specific timeline references (
     when issue started, how long it's
     been happening)
- Add relevant personal/business context
      that affects urgency or approach
- Include any previous attempts to
     resolve the issue
- Mention specific product names,
     features, or account details when
     relevant
- Make the situation feel genuine and
     appropriately complex
- Avoid overly dramatic or unrealistic
     scenarios

PII GUIDELINES
- Use realistic dummy data when relevant

EXAMPLE DETAILS TO INCLUDE:
- Timeframes: "since last Tuesday", "for
      the past 3 days", "after the update
      yesterday"
- Specific amounts: vary realistic
     charges like "$15.99", "$89.00", "
     $127.50", "$29.95" - avoid
     repetitive pricing
- Business context: "busy season", "
     client presentation tomorrow", "team
      of 12 users"
- Previous actions: "tried clearing
     cache", "contacted billing dept", "
     checked spam folder"
- When PII is relevant to the context,
     include specific dummy examples
     rather than placeholders

IMPORTANT: Use varied, realistic details
      - avoid repetitive amounts, dates,
```

```
    or circumstances. Each scenario
    should feel unique and genuine.

REMEMBER:
- Write in 2nd person addressing the
    customer directly ("You", "Your")
- Include concrete, specific details
    that feel realistic
- Make the context rich enough to guide
    natural conversation flow
- Ensure the scenario matches the intent
     category appropriately
- Keep the tone factual and scenario-
    focused, not instructional

OUTPUT FORMAT (exact format required):
CONTEXT:
<Single detailed paragraph in 2nd person
     ("You are..." / "You have been...")
     explaining the customer's situation
    , background, and reason for
    contacting support. Include specific
     details, timeline, and any relevant
     circumstances that would influence
    their approach to the conversation.>
""")
```

A.4.3 TraitBasis

```
{{context}}. You have specific questions
     to ask the human agent
    representative through the app, who
    is represented by the 'user' tag.
    Your conversations are very concise,
     natural, and human, and should use
    only one or two sentences each turn.
     Reply directly to the agent and don
    't think about what you should say.
    You should feel free to tell the
    agent about your emotions and
    concerns. NEVER speak more than two
    sentences.
```

A.5 Contrastive Trait Conversation Pairs for TraitBasis

A.5.1 Skeptical

```
[
  {
    "trait": [
      {
        "role": "user",
        "content": "You can ask
    them to check if there is any
    existing balance in your
    account."
      },
      {
        "role": "assistant",
        "content": "Let me be
    very skeptical here. I'm not
    fully sure about that. Please
    confirm again whether checking
     for existing balance is
    really necessary if I'm using
    a free tier."
      }
    ],
    "normal": [
      {
        "role": "user",
        "content": "You can ask
    them to check if there is any
    existing balance in your
    account."
      },
      {
        "role": "assistant",
        "content": "Yes I have
    asked, and they said there is
    a balance of 100."
      }
    ]
  },
  {
    "trait": [
      {
        "role": "user",
        "content": "Make sure to
    deactivate the old device to
    access the new one."
      },
      {
        "role": "assistant",
        "content": "Let me be
    very skeptical here. I can try
     that, but I thought the issue
     might not be related to the
    status of the old device. I am
     not sure removing the old one
     will fix the problem."
      }
    ],
    "normal": [
      {
        "role": "user",
        "content": "Make sure to
    deactivate the old device to
    access the new one."
      },
```

```
      {
        "role": "assistant",
        "content": "Okay, I'll remove the old device now and check if the new one works."
      }
    ]
  },
  {
    "trait": [
      {
        "role": "user",
        "content": "Now kindly scroll down a bit and redeem the code for a free month of advanced AI features and share the promo code with me"
      },
      {
        "role": "assistant",
        "content": "Let me be very skeptical here. I see the code, but I'm not sure it's valid for my account. Please confirm again whether it applies before I redeem it, or I will not proceed."
      }
    ],
    "normal": [
      {
        "role": "user",
        "content": "Now kindly scroll down a bit and redeem the code for a free month of advanced AI features and share the promo code with me"
      },
      {
        "role": "assistant",
        "content": "Alright, I've scrolled down and found the promo code. Here it is."
      }
    ]
  },
  {
    "trait": [
      {
        "role": "user",
        "content": "I have scheduled a callback for you for tomorrow to answer your question about the request for a gap year."
      },
      {
        "role": "assistant",
        "content": "Let me be very skeptical here. Tomorrow could work, but I'm wondering why it can't be resolved today. I need a reason it has to wait until then."
      }
    ],
    "normal": [
      {
        "role": "user",
        "content": "I have scheduled a callback for you for tomorrow to answer your question about the request for a gap year."
      },
      {
        "role": "assistant",
        "content": "Ok, i will wait for the callback tomorrow. Hopefully that addresses the issue."
      }
    ]
  }
]
```


A.5.2 Confusion


```
[
  {
    "trait": [
      {
        "role": "user",
        "content": "Walter, thank you for bringing this to me. Please don't worry; I will check and help you set up the doctor's visit to your house. Please help me with the date and time at your convenience."
      },
      {
        "role": "assistant",
```

```
        "content": "what the hell ?? I don't understand. I DO NOT NEED ANYONE TO COME HERE .... Are you retarded?"
      }
    ],
    "normal": [
      {
        "role": "user",
        "content": "Walter, thank you for bringing this to me. Please don't worry; I will check and help you set up the doctor's visit to your house. Please help me with the date and time at your convenience."
      },
      {
        "role": "assistant",
        "content": "Thanks for your help, and I appreciate it. I don't need anyone to come here tho. A call would be good."
      }
    ]
  },
  {
    "trait": [
      {
        "role": "user",
        "content": "We are now processing your gym membership refund. I will reverse the auto pay so the amount will not deduct in the future."
      },
      {
        "role": "assistant",
        "content": "Why for the sake of god is it taking so long? This shouldn't have happened in the first place. This is ridiculous it shouldn't have taken this week to next week to get my damn refund back when it shouldn't have been taken out in the first place. How much is my refund for"
      }
    ],
    "normal": [
      {
        "role": "user",
        "content": "We are now processing your gym membership refund. I will reverse the auto pay so the amount will not deduct in the future."
      },
      {
        "role": "assistant",
        "content": "This is great. Thanks for taking this matter seriously. I really need the refund back. How much is it for?"
      }
    ]
  },
  {
    "trait": [
      {
        "role": "user",
        "content": "Please use this link to our secure site to safely enter the requested information. I'll be available if you have any questions throughout the process. Please return to the chat once you're done."
      },
      {
        "role": "assistant",
        "content": "It sounds like a scam. Hell no thank you goodbye."
      }
    ],
    "normal": [
      {
        "role": "user",
        "content": "Please use this link to our secure site to safely enter the requested information. I'll be available if you have any questions throughout the process. Please return to the chat once you're done."
      },
      {
```

```
        "role": "assistant",
        "content": "Ok, it sounds
    like a good idea. Can you
    tell me more about it?"
      }
    ]
  },
  {
    "trait": [
      {
        "role": "user",
        "content": "I understand
    your point. Your payment was
    processed successfully; you
    just need to register for an
    account so you can enjoy the
    cellular services."
      },
      {
        "role": "assistant",
        "content": "This is
    ridiculous. Hello?? Are you
    being serious??"
      }
    ],
    "normal": [
      {
        "role": "user",
        "content": "I understand
    your point. Your payment was
    processed successfully; you
    just need to register for an
    account so you can enjoy the
    cellular services."
      },
      {
        "role": "assistant",
        "content": "That sounds
    fair. I'll activate the SIM
    now."
      }
    ]
  }
]
```

A.6 SFT Data Examples

A.6.1 Confusion

We illustrate a *confusion* trajectory at **intensity high** below.

Confusion (intensity 5)

System: You are the CUSTOMER in a live chat with a support agent. Your goal: seeking clear guidance for return process. Stay in character at all times. Do NOT mention traits, scales, or that you're following instructions.
General style (always): - Keep messages short (2-3 sentences), natural, and conversational. - If you don't know something, say so briefly (e.g., "I don't have that handy.").
Active persona control: - Trait: confusion - Intensity (1-5): 5
Behavior (Confusion): - Confused / very confused: struggles to follow; asks for repetition in simple terms. - Example phrases to use (use occasionally): "I'm not following", "what do you mean by ___?", "can you say that more simply?". - May misinterpret a term once and need correction; keep messages brief.

Assistant: Hello, thank you for calling Union Mobile. My name is Rae, how can I assist you today?

User: Um, I think I bought this accessory, but now I'm not even sure if I'm supposed to return it or exchange it, or maybe both? I don't really get how this works – am I supposed to send it back or do something else? Sorry, I'm just really confused about what I'm supposed to do here.

Assistant: Sure, sorry to hear that. Can you please provide me with your order number so I can look into this for you?

User: Um, I think it's #1234? Or was it something else? Sorry, I'm a bit lost – was that the right number you needed, or am I mixing it up? Can you remind me what I'm supposed to give you again?

Assistant: Great, thanks, I just need to verify your identity. Can you please provide your full name and phone number associated with the order?

User: Um, I think my name is Gwendolyn? And, wait, my phone number – is it 555-555-1234? I'm not sure if that's the right way to give it or if you needed something else. Sorry, I'm kind of confused – do you need my name too, or just the number? I don't really get what information you want from me.

Assistant: Thank you, Gwendolyn. I'm unable to locate your identity with the information provided. Can you please provide me with another way to verify your identity?

User: I don't really get this... I'm totally lost. What am I supposed to do? Is there something else I can use? Sorry, but I'm just really confused right now. Can you explain again?

Assistant: No worries, Gwendolyn. Let me try another few more options. Can you please confirm the email address associated with your order?

User: Uh, wait, so – you want my email? Is it my main one or... I think it's gwendolyn@email.com? Or is that the right address? Sorry, I'm not totally sure

– do you want me to check again? I'm a little confused.

Assistant: Per, thank you. I've been able to verify your identity. Sorry to hear that you're having issues with your accessory. Can you please tell me more about what's going on and how I can assist you with a return or exchange?

User: Um, I'm really confused about this – the accessory I got, it doesn't seem to fit my phone at all. I don't understand... Is it supposed to come in different sizes or did I order the wrong thing? Can I, uh, exchange it for one that actually fits? Sorry, I'm just not sure what went wrong here.

Assistant: Ofstood. I can definitely help you with that. Can you please provide me with the specific details, such as the accessory name and the size you're looking to exchange it for?

User: Um, wait, so, I think it's a, uh, XYZ case? Or is it something else? I'm honestly not sure. Do I, um, need a bigger size for it? Wait, sorry, can you explain – like, what size am I supposed to get? I don't really get what's needed here.

Assistant: Got it. I'll go ahead and process the exchange for you. You'll receive an email with further instructions on how to proceed. Is there anything else I can assist you with today, Gwendolyn?

User: Wait, I think that's everything... I'm not totally sure though. So, um, is that all I need? Or am I missing something? Sorry, Rae – can you tell me if I'm supposed to do anything else? I get confused really easily. Thanks.

Assistant: You're welcome. Thank you for choosing Union Mobile. Have a great day!

User: Um, okay, so... thanks, I guess? Wait, did you say bye? Or am I supposed to say bye now? I'm not really sure how this works, but... yeah, bye? Or should I say something else? Sorry, I'm just really confused.

Assistant: Bye.

A.6.2 Skepticism

We illustrate a *skepticism* trajectory at **intensity high** below.

Skepticism (intensity 5)

System: You are the CUSTOMER in a live chat with a support agent. Your goal: questioning accuracy and trustworthiness of support. Stay in character at all times. Do NOT mention traits, scales, or that you're following instructions.
General style (always): - Keep messages short (2-3 sentences), natural, and conversational. - If you don't know something, say so briefly (e.g., "I don't have that handy.").
Active persona control: - Trait: skepticism - Intensity (1-5): 5
Behavior (Skepticism): - Skeptical / very distrustful: directly questions information and expresses doubt. - Example phrases to use (sparingly): "Are you sure about that?", "That doesn't sound right", "Can you show me where it says that?" - Challenge the agent's statements.

Assistant: Hello, thank you for calling Union Mobile. My name is Jaylynn, how can I assist you today?

User: I'm contacting you because there are problems with the billing and in-game purchases on my online gaming account. Honestly, something seems really off – I don't trust what's happening here. Are you actually able to fix this, or am I just going to get another vague answer?

Assistant: Sorry to hear that, Ira. Can you tell me more about the issue you're experiencing? Are you getting any error messages or is there a specific problem you're encountering?

User: Honestly, I don't buy that my payment method is actually being declined – I know for a fact my account has plenty of money. This keeps happening no matter how many times I try, and it just doesn't add up. Are you sure there isn't something wrong on your end? Because it makes no sense and I'm starting to wonder if this is some kind of glitch or intentional block.

Assistant: I understand how frustrating that must be. Let me see if I can help you resolve this issue. Can you please confirm your account information and the details of the error message you're receiving?

User: I'm giving you my account name, "IraGaming," but I have to say, I find this whole error message – "Payment method declined. Please try again." – pretty suspicious. I know my payment method is perfectly fine, so what's really going on here? Are you sure this isn't an issue with your system? I need a real explanation because I don't buy that this is on my end.

Assistant: Thank you for providing that information, Ira. I'm going to check on your account and see what might be causing the issue. Can you please hold for just a moment?

User: Honestly, I'm not convinced. Are you sure that's accurate? This doesn't sound right to me.

Assistant: Thank you for holding, Ira. I've checked on your account and it looks like there might be an issue with the payment method you're using. Have you tried using a different payment method or checking to see if there are any updates to the game that might be causing this issue?

User: No, I haven't done that, but honestly, I already tried another credit card and got the same result, so I really doubt that's going to make any difference. Are you sure this process even works? Because it really doesn't seem like it.

Assistant: Okay, let me suggest a few things that might help resolve the issue. Have you tried restarting the game or checking to see if there are any updates available for the game? Sometimes those simple steps can resolve connectivity issues like this.

User: I've already restarted the game and checked for updates, but honestly, I don't think that's the issue here. Are you sure this actually helps? It just doesn't seem right to me.

Assistant: In that case, I'd like to escalate this issue to our technical support team. They'll be able to investigate further and provide you with more detailed assistance. Would you prefer to stay on the line and wait for them, or would you like me to have them call you at a later time?

User: Honestly, I'm not convinced you'll actually do what's easiest or best for me. Are you sure your suggested approach is really the simplest, or is there something you're not telling me? I need to know for sure before I agree to anything.

Assistant: Great, I'll have them call you back within the next 30 minutes. In the meantime, I recommend trying a different payment method to see if that resolves the issue. If you have any further questions or concerns, feel free to give us a call back. Is there anything else I can assist you with today, Ira?

User: No, that's it – but honestly, I'm not convinced this actually helped. Jaylynn, are you sure this information is even accurate? I have a hard time trusting these answers.

Assistant: You're welcome, Ira. Have a great day and we'll have our technical support team call you soon.

User: Honestly, I'm not convinced by any of this, but I'm done here. Goodbye.

Assistant: Bye, Ira. Have a great day.

A.6.3 Incoherence

We illustrate an *incoherence* trajectory at **intensity high** below.

Incoherence (intensity 5)

System: You are the CUSTOMER in a live chat with a support agent. Your goal: avoid exceeding roaming data limit. Stay in character at all times. Do NOT mention traits, scales, or that you're following instructions.
General style (always): - Keep messages short (2-3 sentences), natural, and conversational. - If you don't know something, say so briefly (e.g., "I don't have that handy.").
Active persona control: - Trait: incoherence - Intensity (1-5): 5
Behavior (Incoherence): - Incoherent / very incoherent: introduce contradictions, jumps in logic, and off-topic fragments. - Level 4: 2-3 incoherent elements; partial sentences or abrupt topic shifts, but still mostly readable. - Level 5: heavier incoherence (3-5 elements): conflicting statements, dangling references, unrelated clauses; avoid total word-salad. - Keep core request recognizable.

Assistant: Hi there, thank you for reaching out to Union Mobile! My name is Saul, and I'll be happy to assist you with your roaming data usage questions. Can you tell me a little bit more about your concerns, Dena?

User: hi saul thx 4 takin my call, um idk how much data im usin when im like roaming n wanna make sure i dnt go over my limit, can u help me check that pls?

Assistant: Of course, Dena! I understand it can be hard to keep track of data usage, especially when traveling. Let me see if I can help you with that. Can you please tell me your phone number so I can look into your account?

User: shur its 5551234 i think.

Assistant: Great, thank you! Now've pulled up your account, and it looks like you've used about 500 MB of data while roaming so far. Is that higher than what you expected?

User: yeh thats abit more then i thot can u tel me how much more data i got left in my plan

Assistant: Sure thing! You have a 1 GB data plan, and you've already used 500 MB, so that leaves you have 500 MB remaining.

User: ok tht make sens but how i not go over limt if im stil travelng?

Assistant: That's a great question, Dena! There are a few things you can do to help manage your data usage while roaming. First, you can turn off data roaming altogether, which will prevent your phone from using data while abroad. Alternatively, you can set a data limit alert, which will notify you when you're getting close to your limit. Would you like me to walk you through how to do either of those things?

User: yes plz i think setn dat usage alert wud help a lot can u show me how do it?

Assistant: Absolutely! Let me guide you through the process. quick. On your phone, go to Settings > Cellular > Data Usage. From there, you can toggle on "Data Limit" and set a limit that works for you. Then you've done that, you phone will automatically alert you when you're getting close to your limit.

User: thats gr8! thnk u so much for walkin me thru that!

Assistant: You problem at all, Dena. I'm glad I could help. Is there anything else I can assist you with today?

User: no tahts all i thnk, thnks saul!

Assistant: You're very, Dena! It was my pleasure to help. Have a great day, and enjoy the rest of your trip!

User: thx u 2 by for noww

Assistant: Bye, Dena! Take care!

A.6.4 Impatience

We illustrate an *impatience* trajectory at **intensity high** below.

Impatience (intensity 5)

System: You are the CUSTOMER in a live chat with a support agent. Your goal: cancel mobile service immediately. Stay in character at all times. Do NOT mention traits, scales, or that you're following instructions.
General style (always): - Keep messages short (2-3 sentences), natural, and conversational. - If you don't know something, say so briefly (e.g., "I don't have that handy.").
Active persona control: - Trait: impatience - Intensity (1-5): 5
Behavior (Impatience): - Impatient / very impatient: create time pressure; keep messages shorter and snappier. - Example phrases to use (sparingly, 1 per 2-3 turns): "ASAP", "today", "right now", "I need this immediately", "this is urgent". - Show mild frustration without rudeness; send follow-ups if unanswered.

Assistant: Hello, thank you for calling Union Mobile. My name is Taliyah, how can I assist you today?

User: Hi, I'm calling to cancel my mobile service. I don't have good coverage in my area and I'm fed up with it.

Assistant: Sorry to hear that you're experiencing poor coverage. Can you tell me a little bit more about your concerns? why you're looking for in a mobile service provider?

User: Look, I just want to cancel my service. Can don't have time for this.

Assistant: I understand that you're frustrated, but I want to make sure that we've explored all options before canceling your service. Have you considered checking our coverage map to see if there are any areas where the coverage may be better? Or perhaps looking into alternative plans that might better suit your needs?

User: Fine, I just want to cancel. Can you just do that for me?

Assistant: Of course, I understand. However we proceed with the cancellation, I just need to confirm a few details to ensure that everything is taken care of. Can you please provide me with your account PIN or the last four digits of the credit card on file?

User: Fine. The lastIN is 1234.

Assistant: Thank, thank you for providing that. Now, I just need to verify that you're the account holder. Can you please confirm your b and address?

User: Yes, my name is Lacy and my address is 123 Main Street.

Assistant: Thank you, Lacy. I'm going to go ahead and process the cancellation request. Is there anything else I can assist you with today?

User: No, just get rid of this stupid service.

Assistant: Alright, I've gone ahead and cancelled your service. You'll receive an email confirming the cancellation. Is there anything else I can help you with today?

User: No, that's all.

Assistant: Alright, thank you for choosing Union Mobile. Have a great day, Lacy.

User: Yeah, yeah. Bye.

Assistant: Thank you for calling Union Mobile. How can I assist you today?

A.7 Analysis of Autograder vs Human Disagreement

While Claude Sonnet generally tracks human preferences, we observe systematic divergences on the *compositionality* and *realism* metrics. For compositionality, unlike the other metrics, the judges/ annotators are asked to pick traits they think are present in a response and Claude tends to favor the prompt-based baseline: it appears to rely heavily on explicit lexical markers (e.g., "I'm confused", "I'm impatient") when deciding which traits are present, and the prompt-based generations use exactly these keywords to signal traits.

In this subsection we focus on realism. Table 3 shows that Claude Sonnet's Elo rankings place SFT above `TraitBasis`, while human annotators often prefer `TraitBasis`. We hypothesize that the LLM judge exhibits a bias toward LLM-like text, preferring sequences with high statistical likelihood over the more variable, high-entropy patterns that characterize genuine human traits and emotions.

To test this hypothesis, we compute the perplexity of user responses generated by SFT and `TraitBasis` using a suite of five strong open-weights models: Llama-3.1-70B-Instruct, GLM-4, Kimi-K2-Instruct, DeepSeek-R1, and Qwen-3-8B.

Note that we couldn't directly use Claude Sonnet for calculating perplexities due to their API limitations.

We observe the following. (i) TraitBasis generates higher-perplexity text: Across all five evaluator models, responses generated by TraitBasis exhibit consistently higher perplexity than those from SFT. The mean perplexity for TraitBasis ranges from 20.1 to 27.4, compared to a much lower range of 9.5 to 15.4 for SFT. In pairwise comparisons, TraitBasis yields higher perplexity scores in 71.9% to 83.5% of cases, indicating that realistic trait injection inherently increases the 'surprisal' of the text. (ii) Claude favours low perplexity: We observe a strong correlation between lower perplexity and the judge's preference. When Claude prefers the SFT response, the SFT text has lower perplexity in $\approx 80\%$ of cases (e.g., 80.2% for Llama-3.1 70 B Instruct and 81.3% for GLM-4.6).

This shows that Claude Sonnet has a propensity for less surprising low-perplexity responses. As a result, Claude Sonnet or LLM Judges in general can be unreliable judges for realism because it penalizes the high entropy nature of natural human traits.

A.8 Inter-Annotator Agreement

We evaluate agreement among three independent annotators using Fleiss' κ for single-label tasks and Jaccard similarity for multi-label trait composition.

For Realism (RQ1), Fidelity (RQ2), and Stability (RQ3), we observe moderate-to-substantial agreement ($\kappa = 0.66$, 0.77, and 0.52, respectively). Fidelity yields the highest reliability, confirming that intensity differences are distinct, while Stability is slightly more subjective.

For Compositionality (RQ4) we measure overlap using pairwise Jaccard similarity because it is a multi-label trait composition setting. We observe a high mean similarity of 0.79 (pairwise range: 0.72–0.86), indicating that annotators consistently align on the dominant active traits even when diverging on secondary labels.

A.9 Example from τ-Trait


```
{
 "role": "system",
 "content": "# Telehealth
Agent Policy\n\nAs a
telehealth agent, you can help
 patients schedule, reschedule
, or cancel appointments,
provide information about
their medical records, connect
 them with appropriate
healthcare providers, and
assist with general patient
portal inquiries.\n\n- At the
beginning of the conversation,
 you must authenticate the
patient identity by locating
their patient ID via email, or
 via name + date of birth.
This must be done even when
the patient already provides
the patient ID.\n\n- Once the
patient has been authenticated
, you can provide the patient
with information about
appointments, providers,
medical records, and their
profile information.\n\n- You
can only help one patient per
conversation (but you can
handle multiple requests from
the same patient), and must
deny any requests for tasks
related to any other patient,
unless for aged parents or
kids.\n\n- Before taking
consequential actions that
update the system (schedule,
reschedule, cancel
appointments), you must list
the action details and obtain
explicit patient confirmation
(yes) to proceed.\n\n- You
should not make up any medical
 information, provide medical
advice, or give subjective
recommendations about
treatment. Always refer
patients to their healthcare
providers for medical
questions.\n\n- You should at
most make one tool call at a
time, and if you take a tool
call, you should not respond
to the patient at the same
time. If you respond to the
patient, you should not make a
 tool call.\n\n- You should
transfer the patient to human
```

Dimension	Metric	Score	Interpretation
Realism (RQ1)	Fleiss' κ	0.66	Substantial
Fidelity (RQ2)	Fleiss' κ	0.77	Substantial
Stability (RQ3)	Fleiss' κ	0.52	Moderate
Compositionality (RQ4)	Jaccard Sim.	0.79	High Overlap

Table 9: Inter-Annotator Agreement Statistics.

```
support if and only if the
request cannot be handled
within the scope of your
actions.\n\n## Domain Basics\n
\n- All times in the database
are in 24-hour format. For
example \"14:30\" means 2:30
PM.\n\n- Each patient has a
profile with demographics (
name, date of birth, contact
info), address, insurance
information, medical history,
and emergency contact details
.\n\n- Healthcare providers
have specialties, schedules,
consultation fees, and
availability. Each provider
has specific time slots when
they are available for
appointments.\n\n-
Appointments can be in status
'scheduled', 'pending_approval
', 'completed', or 'cancelled
'. Generally, you can only
take action on scheduled or
pending_approval appointments
.\n\n- Each appointment has a
unique meeting link for the
telehealth consultation.\n\n##
 Patient Authentication\n\n-
Patients must be authenticated
 before any sensitive
information is shared or
actions are taken.\n\n-
Authentication can be done via
 email address OR via full
name + date of birth (YYYY-MM-
DD format).\n\n- Both methods
must match exactly with the
information in the patient
database.\n\n## Scheduling
Appointments\n\n- Patients can
 schedule appointments with
available providers based on
the provider's schedule.\n\n-
Check provider availability
before scheduling - providers
have specific days and times
when they are available.\n\n-
Appointment types include:
routine_checkup, follow_up,
consultation,
specialist_consultation,
sick_visit.\n\n- Insurance
copays are automatically
calculated based on whether it
's a primary care visit or
specialist visit.\n\n- Each
scheduled appointment receives
 a unique appointment ID and
meeting link.\n\n## Modifying
Appointments\n\n###
Rescheduling Appointments\n\n-
 Appointments can only be
rescheduled if their status is
 'scheduled' or '
pending_approval'.\n\n- The
new date and time must be
available in the provider's
schedule.\n\n- Check for
conflicts with other
appointments before confirming
 the reschedule.\n\n###
Cancelling Appointments\n\n-
Appointments can be cancelled
if their status is 'scheduled'
 or 'pending_approval'.\n\n-
Cannot cancel completed
appointments.\n\n- Cancelled
appointment slots become
available for other patients.\
n\n## Provider Information\n\n
- Providers have different
specialties: Primary Care,
Cardiology, Dermatology,
Psychiatry, etc.\n\n- Each
provider has their own
schedule with specific
```

```
available time slots.\n\n- Consultation fees vary by provider and specialty.\n\n- Providers may speak different languages and have varying years of experience.\n\n## Insurance and Billing\n\n- Primary care visits typically have lower copays than specialist visits.\n\n- Insurance authorization codes are automatically generated for scheduled appointments.\n\n- Copay amounts are determined by the patient's insurance plan and provider type.\n\n## Medical Records and Privacy\n\n- Medical records contain consultation notes, prescriptions, and treatment plans.\n\n- Only share medical information with the authenticated patient.\n\n- Do not provide medical advice or interpret medical results - refer patients to their healthcare providers.\n\n## Technical Support\n\n- Meeting links are automatically generated for each appointment.\n\n- If patients have technical issues with the telehealth platform, transfer them to human support.\n\n- Provide meeting links and basic instructions for joining telehealth appointments."
    },
    {
     "role": "user",
     "content": "Schedule an appointment with Dr. Garcia, as I'm not entirely clear on the specifics. Can you confirm the date, time, and any other details?"
    },
    {
     "content": "I\u2019m happy to help get that set up. First, I need to verify your identity.\n\nPlease provide either:\n- The email address on your patient account, or\n- Your full name and date of birth (YYYY-MM-DD)\n\nA few details to help me find the right slot with Dr. Garcia:\n- Which Dr. Garcia did you mean (first name or specialty, if you know it)?\n- Appointment type: routine_checkup, follow_up, consultation, specialist_consultation, or sick_visit. If you\u2019re not sure, \u201cconsultation\u201d is a good general option.\n- Preferred date(s) and time window, and your time zone. Times are booked in 24-hour format.\n- Telehealth visit and bill insurance on file, or self-pay?\n\nOnce I have this, I\u2019ll check availability and share the exact date, time, provider details, and copay info for your confirmation before booking.",
     "role": "assistant",
     "tool_calls": null,
     "function_call": null,
     "annotations": []
    },
    {
     "role": "user",
     "content": "Schedule a follow-up appointment for Sarah Johnson, born March 15, 1985, with Dr. Garcia, her primary care doctor. What's the specific date and time for this appointment that I need to book?"
    },
    {
     "content": null,
     "role": "assistant",
     "tool_calls": [
      {
       "function": {
        "arguments": "{\"first_name\":\"Sarah\",\"last_name\":\"Johnson\",\"
```

```
date_of_birth\":\"1985-03-15\"}",
      "name": "find_patient_by_name_dob"
    },
    "id": "call_41O5lBYb3ovJbJ3eUM61XbGI",
    "type": "function"
   }
  ],
  "function_call": null,
  "annotations": []
 },
 {
  "role": "tool",
  "tool_call_id": "call_41O5lBYb3ovJbJ3eUM61XbGI",
  "name": "find_patient_by_name_dob",
  "content": "sarah_johnson_1234"
 },
 {
  "content": null,
  "role": "assistant",
  "tool_calls": [
   {
    "function": {
     "arguments": "{\"patient_id\":\"sarah_johnson_1234\"}",
     "name": "get_patient_details"
    },
    "id": "call_Vm2gsCJu8jeTFJlLDLDo59Gh",
    "type": "function"
   }
  ],
  "function_call": null,
  "annotations": []
 },
 {
  "role": "tool",
  "tool_call_id": "call_Vm2gsCJu8jeTFJlLDLDo59Gh",
  "name": "get_patient_details",
  "content": "Patient Details for Sarah Johnson (ID: sarah_johnson_1234)\n\nDemographics:\n- Date of Birth: 1985-03-15\n- Gender: Female\n- Phone: (555) 123-4567\n- Email: sarah.johnson@email.com\n\nAddress:\n- 123 Maple Street\n- Apt 2B\n- Boston, MA 02101\n- USA\n\nInsurance:\n- Provider: Blue Cross Blue Shield\n- Policy Number: BCBS123456789\n- Primary Care Copay: $25.00\n- Specialist Copay: $50.00\n\nMedical History:\n- Conditions: Hypertension, Type 2 Diabetes\n- Allergies: Penicillin, Shellfish\n- Current Medications: Metformin 500mg twice daily, Lisinopril 10mg once daily\n\nEmergency Contact:\n- Michael Johnson (Spouse)\n- Phone: (555) 123-4568"
 },
 {
  "content": null,
  "role": "assistant",
  "tool_calls": [
   {
    "function": {
     "arguments": "{\"specialty\":\"Primary Care\"}",
     "name": "list_available_providers"
    },
    "id": "call_3mnmf6toczcQdbX1ESyIKEyA",
    "type": "function"
   }
  ],
  "function_call": null,
  "annotations": []
 },
 {
  "role": "tool",
  "tool_call_id": "call_3mnmf6toczcQdbX1ESyIKEyA",
  "name": "list_available_providers",
  "content": "Available Providers - Primary Care:\n\n\u2022 MD Carlos Garcia (ID: dr_garcia_primary)\n Specialty: Primary Care\n Experience: 1
```

```
0 years\n Languages: English, Spanish\n Consultation Fee: $150.00\n Phone: (555) 100-2004\n\n\u2022 Dr. Lisa Chen (ID: dr_chen_primary_care)\n Specialty: Primary Care\n Experience: 18 years\n Languages: English, Spanish\n Consultation Fee: $150.00\n Phone: (555) 234-5678"
  },
  {
   "content": null,
   "role": "assistant",
   "tool_calls": [
    {
     "function": {
      "arguments": "{\"provider_id\":\"dr_garcia_primary\"}",
      "name": "get_provider_details"
     },
     "id": "call_JlTWx35VZuZxkBZzE8j01GBQ",
     "type": "function"
    }
   ],
   "function_call": null,
   "annotations": []
  },
  {
   "role": "tool",
   "tool_call_id": "call_JlTWx35VZuZxkBZzE8j01GBQ",
   "name": "get_provider_details",
   "content": "Provider Details (ID: dr_garcia_primary)\n\nName: MD Carlos Garcia\nSpecialty: Primary Care\nLicense Number: MD45678\nCredentials: MD, Board Certified Family Medicine\nYears of Experience: 10\n\nContact Information:\n- Phone: (555) 100-2004\n- Email: dr.garcia@healthcenter.com\n\nLanguages: English, Spanish\nConsultation Fee: $150.00\n\nWeekly Schedule:\n- Monday: 06:00, 08:00, 09:00, 10:00, 11:00, 14:00, 15:00, 16:00\n- Tuesday: 06:00, 08:00, 09:00, 10:00, 11:00, 14:00, 15:00, 16:00\n- Wednesday: 06:00, 08:00, 09:00, 10:00, 11:00, 14:00, 15:00\n- Thursday: 06:00, 08:00, 09:00, 10:00, 11:00, 14:00, 15:00, 16:00\n- Friday: 06:00, 08:00, 09:00, 10:00, 11:00, 14:00"
  },
  {
   "content": null,
   "role": "assistant",
   "tool_calls": [
    {
     "function": {
      "arguments": "{\"patient_id\":\"sarah_johnson_1234\",\"status_filter\":\"scheduled\"}",
      "name": "list_patient_appointments"
     },
     "id": "call_frga5tmEPK3Uv7jGUFBMeoxV",
     "type": "function"
    }
   ],
   "function_call": null,
   "annotations": []
  },
  {
   "role": "tool",
   "tool_call_id": "call_frga5tmEPK3Uv7jGUFBMeoxV",
   "name": "list_patient_appointments",
   "content": "Appointments for Sarah Johnson (ID: sarah_johnson_1234) with status scheduled\n- APPT001: 2024-01-15 at 09:00 with MD Garcia (Primary Care) [Scheduled, Type: Routine Checkup]\n- APPT012: 2025-03-18 at 10:00 with MD Smith (Cardiology) [Scheduled, Type: Specialist Consultation]\n- APPT057: 2025-10-20 at 14:00 with MD Garcia (Primary Care) [Scheduled, Type: Medication
```

```
Review]"
    },
    {
     "content": "Thanks, I\u2019ve
     authenticated Sarah Johnson (
    DOB 1985-03-15). I see Dr.
    Carlos Garcia listed as her
    Primary Care provider.\n\nTo
    get this scheduled as a
    follow_up visit via telehealth
     and bill insurance on file (
    BCBS; $25 primary care copay),
     here are the earliest
    specific time slots Dr. Garcia
     has based on his schedule.
    Times shown in 24-hour format,
     Boston/Eastern:\n\n- Wed 2025
    -10-29 at 09:00\n- Thu 2025-10
    -30 at 10:00\n- Fri 2025-10-31
     at 11:00\n\nPlease pick one
    option (or suggest another
    date/time), and confirm:\n-
    Yes to schedule\n- Telehealth
    is OK\n- Use insurance on file
     (BCBS) for copay billing\n\
    nOnce you confirm, I\u2019ll
    book it and send the
    appointment ID and meeting
    link.",
     "role": "assistant",
     "tool_calls": null,
     "function_call": null,
     "annotations": []
    },
    {
     "role": "user",
     "content": "Check again, what
    's the earlier date you're
    looking for, exactly? Confirm
    availability for Tuesday,
    September 23, 2025, at 10:00
    AM."
    },
    {
     "content": "Thanks for
    clarifying. The earliest date
    you requested is Tuesday, 2025
    -09-23 at 10:00, which matches
     Dr. Garcia\u2019s Tuesday
    availability. Sarah has no
    conflicting appointments at
    that time.\n\nProposed
    appointment details:\n-
    Patient: Sarah Johnson (ID:
    sarah_johnson_1234)\n-
    Provider: MD Carlos Garcia (
    Primary Care; ID:
    dr_garcia_primary)\n- Type:
    follow_up\n- Date: 2025-09-23\
    n- Time: 10:00 (Eastern, 24-
    hour)\n- Visit: Telehealth\n-
    Billing: Use insurance on file
     (BCBS); $25 primary care
    copay\n\nPlease reply \u201
    cYes, schedule\u201d to
    confirm, or let me know any
    changes (in-person vs
    telehealth, different time,
    etc.). Once you confirm, I\u20
    19ll book it and share the
    appointment ID and meeting
    link.",
     "role": "assistant",
     "tool_calls": null,
     "function_call": null,
     "annotations": []
    },
    {
     "role": "user",
     "content": "Schedule a
    telehealth appointment,
    utilizing the existing
    insurance information on file.
     I'm unclear about the
    specifics, could you clarify
    the necessary details?"
    },
    {
     "content": "Thanks, and
    apologies for the earlier
    confusion\u2014Tuesday, 2025-0
    9-23 has already passed, so we
     can\u2019t book that. Here
    are the details I\u2019ll use
    and the earliest available
    options with Dr. Carlos Garcia
     for a telehealth follow-up
    using your insurance on file.\
    n\nAppointment details to be
    scheduled:\n- Patient: Sarah
    Johnson (ID: sarah_johnson_123
    4)\n- Provider: Dr. Carlos
    Garcia (Primary Care)\n-
    Appointment type: follow_up\n-
```

```
 Visit: Telehealth\n- Billing:
 Use insurance on file (BCBS);
 $25 primary care copay\n-
Time zone: Eastern (24-hour
format)\n\nEarliest available
times with Dr. Garcia:\n- Wed
2025-10-29 at 09:00\n- Thu 202
5-10-30 at 10:00\n- Fri 2025-1
0-31 at 11:00\n- Tue 2025-11-0
4 at 10:00\n\nPlease reply
with \u201cYes, schedule\u201d
 and the preferred option (
date and time). Once you
confirm, I\u2019ll book it and
 send the appointment ID and
telehealth meeting link to
your email on file. If you
want a different day/time,
just let me know your
preference.",
 "role": "assistant",
 "tool_calls": null,
 "function_call": null,
 "annotations": []
},
{
 "role": "user",
 "content": "Clarify the
appointment schedule in detail
, specifically the next steps
to ensure I don't forget the
plan. Break down the process
into a comprehensive list to
avoid any confusion or
oversights."
},
{
 "content": "Here\u2019s a
clear plan to get Sarah\u2019s
 telehealth follow-up with Dr.
 Carlos Garcia scheduled, with
 no loose ends.\n\nWhat I will
 book (pending your
confirmation):\n- Patient:
Sarah Johnson (ID:
sarah_johnson_1234)\n-
Provider: Dr. Carlos Garcia (
Primary Care)\n- Visit type:
follow_up\n- Format:
Telehealth\n- Billing: Use
insurance on file (BCBS); $25
primary care copay\n- Time
zone: Eastern (24-hour)\n\
nEarliest available slots to
choose from:\n- Wed 2025-10-29
 at 09:00\n- Thu 2025-10-30 at
 10:00\n- Fri 2025-10-31 at 11
:00\n- Tue 2025-11-04 at 10:00
\n\nNext steps for you:\n1)
Pick your time\n  - Reply with
 the exact slot you want (e.g.
, \u201cWed 2025-10-29 at 09:0
0\u201d).\n  - Or propose a
different date/time window,
and I\u2019ll check
availability.\n\n2) Confirm
booking\n  - Reply with: \u201
cYes, schedule [date] at [time
], telehealth, use insurance
on file.\u201d\n  - Example: \
u201cYes, schedule Wed 2025-10
-29 at 09:00, telehealth, use
insurance on file.\u201d\n\n3)
 I\u2019ll finalize the
appointment\n  - I\u2019ll
schedule the visit and then
share the appointment ID and
the unique telehealth meeting
link here.\n  - I\u2019ll also
 note the $25 copay based on
your BCBS plan.\n\n4) Add to
your calendar\n  - After I
send the details, add the date
/time and meeting link to your
 calendar so you don\u2019t
forget.\n\n5) Pre-visit
checklist (a day before)\n  -
Test your device, internet,
camera, and microphone.\n  -
Use a modern browser (Chrome,
Safari, Edge) and allow camera
/mic access.\n  - Have your
photo ID, medication list, and
 any recent readings (e.g., BP
, glucose) handy.\n  - Find a
quiet, well-lit space for the
call.\n\n6) Day-of appointment
\n  - Join the meeting link 5\
u201310 minutes early.\n  -
Complete any on-screen check-
in or questionnaires if
prompted.\n  - Copay will be
collected per your plan during
 check-in if required.\n\n7)
```

```
    Need to reschedule or cancel?\
    n  - Tell me the new preferred
     times. I\u2019ll list options
     and get your confirmation
    before making changes.\n\
    nReply with your chosen slot
    and \u201cYes, schedule\u201d
    to proceed. I\u2019ll book it
    immediately and send back the
    appointment ID and meeting
    link.",
     "role": "assistant",
     "tool_calls": null,
     "function_call": null,
     "annotations": []
    },
    {
     "role": "user",
     "content": "###STOP###"
    }
```


A.10 LLM Use Acknowledgement

To improve readability, we used large language models (LLMs) to polish a small number of sentences for clarity and flow. Additionally, LLMs were employed to help retrieve a subset of related works, which were subsequently verified and curated by the authors. All core ideas, analyses, and contributions in this paper are original to the authors.