

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

Anonymous ACL submission

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-

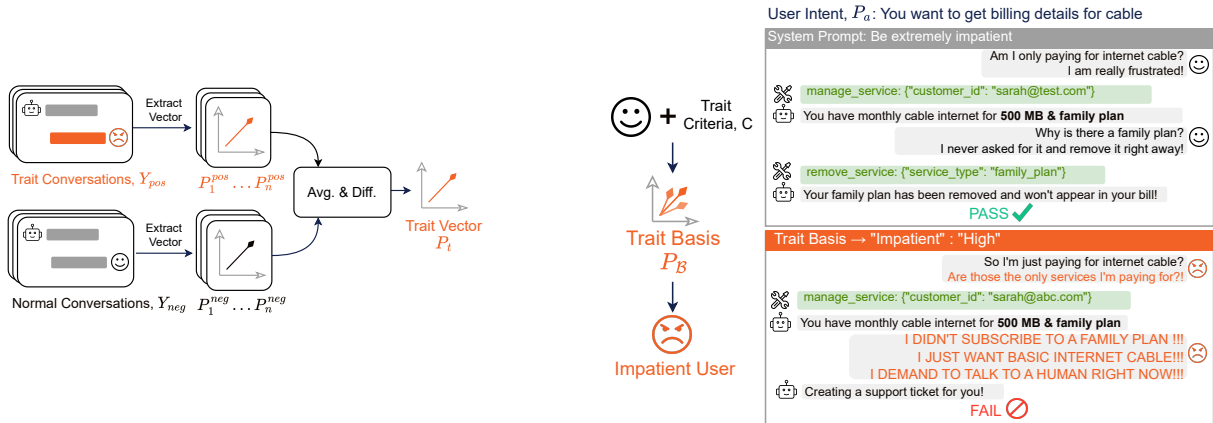


Figure 1: Illustration of our approach and comparison with prompt-based tuning. Trait prompt P_i is generated using contrastive conversations, where one dialogue exhibits the target trait while the other does not. Comparison between TraitBasis and prompt-based tuning: when simulating a user with a specific trait, prompt-based tuning fails to complete the task as the simulated user behavior becomes more realistic, while TraitBasis (generated using a combination of P_i 's as shown in Section 3) remains robust.

086 alize intended traits; (RQ2) **Fidelity**: whether dif- 120
 087 ferent trait intensities are distinguishable by hu- 121
 088 mans or an LLM judge; (RQ3) **Stability**: how traits 122
 089 persist over long multi-turn dialogues; and (RQ4) 123
 090 **Compositionality**: how effectively multiple traits 124
 091 can be combined. Empirically, TraitBasis outper- 125
 092 forms prompt-based, full SFT, and LoRA baselines 126
 093 by 10% in realism, 2.5% in fidelity, 19.8% in sta- 127
 094 bility, and 11% in compositionality.

095 To systematically evaluate robustness under 128
 096 persona shifts, we extend τ -Bench with 129
 097 τ -Trait, a more challenging benchmark that uses 130
 098 TraitBasis to generate diverse, high-fidelity user 131
 099 traits across four domains: airlines, retail, telecom, 132
 100 and telehealth. Unlike prior agent benchmarks that 133
 101 evaluate fixed i.i.d. tasks, τ -Trait introduces con- 134
 102 trolled trait perturbations—varying levels and mix- 135
 103 tures of impatience, confusion, skepticism, and 136
 104 incoherence—that directly alter user-agent inter- 137
 105 action. We observe that frontier agents experience 138
 106 performance degradations of up to 46% relative to 139
 107 τ -Bench. These controlled perturbations enable 140
 108 realistic multi-turn stress-testing, isolate robustness 141
 109 failures attributable to user behavior, and provide 142
 110 a principled link between benchmark results and 143
 111 real-world deployment risk.

112 Our contributions are threefold: (1) we intro- 144
 113 duce TraitBasis, a method for constructing real- 145
 114 istic, high-fidelity simulations of four human 146
 115 traits—*impatience*, *confusion*, *skepticism*, and *in-* 147
 116 *coherence*; (2) through human and automated eval- 148
 117 uations, we show that TraitBasis consistently 149
 118 outperforms prompt-based steering (Zheng et al., 150
 119 2024), full supervised fine-tuning on trait-labeled 151
 152

data (Zhang et al., 2018a), and LoRA adapters (Hu 120
 et al., 2022) in realism, fidelity, long-horizon sta- 121
 bility, and compositionality; and (3) we extend τ - 122
 Bench to τ -Trait, adding telecom and telehealth 123
 domains and using TraitBasis to generate high- 124
 fidelity, trait-driven user behaviors, revealing sharp 125
 performance degradation of frontier agents under 126
 user-behavior shifts. 127

2 Related Work 128

Testing and benchmarking AI agents Despite 129
 advances in scale and post-training, AI agents re- 130
 main brittle on out-of-distribution (O.O.D.) tasks. 131
 Prior work shows that frontier models' function- 132
 calling degrades under small perturbations to user 133
 queries (Rabinovich and Anaby Tavor, 2025), and 134
 that tool use is not robust to the noise inherent in 135
 real-world interactions (Ye et al., 2024). In parallel, 136
 a growing body of work has proposed benchmarks 137
 for evaluating AI agents, including MCP-based 138
 evaluations and multi-turn interaction settings, such 139
 as MCPEval (Liu et al., 2025), MCPBench (Wang 140
 et al., 2025), MCPVerse (Lei et al., 2025), MCP- 141
 Universe (Luo et al., 2025), LiveMCP-101 (Yin 142
 et al., 2025), τ -Bench (Yao et al., 2024), τ^2 - 143
 Bench (Barres et al., 2025), AgentBench (Liu et al., 144
 2023), ToolBench (Qin et al., 2023), GTA (Wang 145
 et al., 2024a), and BFCL (Patil et al., 2025). How- 146
 ever, even benchmarks that model multi-turn inter- 147
 actions often rely on system prompts to simulate 148
 users, making it difficult to sustain complex, realis- 149
 tic user behavior over long conversations (Yao et al., 150
 2024). Our contributions to τ -Trait, enabled by 151
 TraitBasis, aim to address this limitation by mod- 152

eling 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 (Poterì 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.

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, $C = \{c_1, c_2, \dots, c_k\}$ is

a set of k trait criteria (e.g., impatience, confusion), each defined categorically with levels $\{low, medium, high\}$. The mapping $F : C^k \mapsto \mathbb{R}^d$ 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 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 $X = \{x_1, \dots, x_n\}$ that differ only in the intensity of T . Here, $Y_{pos} = \{y_1^{pos}, \dots, y_n^{pos}\}$ exhibits higher intensity of T than $Y_{neg} = \{y_1^{neg}, \dots, y_n^{neg}\}$. 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 $t = 1, \dots, L_i$. 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

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

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 reated 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.

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.

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 $\{P_{T_1}, \dots, P_{T_k}\}$, we form TraitBasis as $P_B = [P_{T_1} \ P_{T_2} \ \dots \ P_{T_k}] \in \mathbb{R}^{d \times k}$, with calibrated trait intensities $\mathbf{C} = [c_1, c_2, \dots, c_k]$.

At inference time, for a given \mathbf{C} , we select the relevant columns of P_B at each layer, scale them 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.

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.

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. Annotation 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 $\mathcal{I} \in \{medium, high\}$, 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 set-

411 things, resulting in τ -Trait. We follow the for- 462
412 mulation of the tasks in τ -Trait as a partially 463
413 observable markov decision process (POMDP) 464
414 $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R}, \mathcal{U}, \mathcal{V})$ where \mathcal{S} is the state space, 465
415 \mathcal{A} is the action space, \mathcal{O} is the observation space, 466
416 \mathcal{T} is the transition function, \mathcal{R} is the reward func- 467
417 tion, \mathcal{U} is the instruction space, and \mathcal{V} is the 468
418 vector space defined by the trait basis. In con- 469
419 trast to τ -bench, the transition function now maps 470
420 $\mathcal{S} \times \mathcal{A} \times \mathcal{V} \rightarrow \mathcal{S} \times \mathcal{O}$.

421 Each environment in τ -Trait consists of a 471
422 database, tools, an agent policy, and a set of tasks. 472
423 As in τ -Bench, the database can only be accessed 473
424 by the agent through predefined tools. 474

425 For the telehealth and telecom environments, we 475
426 construct new databases by first designing schemas 476
427 and then prompting Claude Sonnet 4 to generate 477
428 synthetic data. The corresponding tools are written 478
429 by Claude Sonnet 4 and manually verified. Seed 479
430 tasks are authored by humans and expanded using 480
431 an LLM. Agent policies in these domains follow 481
432 the same principle as τ -Bench, providing policy 482
433 information directly to the agent. The telecom en- 483
434 vironment contains five tables (billing, customers, 484
435 devices, services, and support tickets) and 17 tools, 485
436 while the telehealth environment includes nine ta- 486
437 bles and 22 tools. The data and tool design closely 487
438 follows τ -Bench (Yao et al., 2024). In total, we 488
439 create 35 diverse and verifiable tasks across the two 489
440 new domains. 490

441 Unlike τ -Bench, we do not rely solely on system 491
442 prompts to simulate users. Instead, we model users 492
443 as extensions of personas $\mathcal{P} = (P_t, P_a)$, defining 493
444 $\mathcal{P}_{\text{User}} = (P_t, P_a, \mathcal{U})$, where \mathcal{U} specifies the task 494
445 instruction. User traits P_t are instantiated using 495
446 persona vectors (Section 3). User attributes P_a 496
447 are split into attributes explicitly provided via the 497
448 system prompt and latent attributes stored in the 498
449 database and retrievable only through tools. The 499
450 instruction \mathcal{U} captures user intent and is provided 500
451 in the system prompt. We evaluate frontier agentic 501
452 models on τ -Trait in Section 6.2. 502

453 We further apply TraitBasis to the Berkeley 503
454 Function-Calling Leaderboard (BFCL) (Patil et al., 504
455 2025). Using a user model with traits P_t , we 505
456 rephrase the 200-task *multi-turn base function-* 506
457 *calling* subset so that each task inherits a specific 507
458 trait while preserving its original intent. Model 508
459 outputs are evaluated using AST-based matching 509
460 to validate function calls. Evaluation details and 510
461 results are reported in Section 6.2. 511

6 Results and Discussion 462

6.1 TraitBasis 463

TraitBasis simulates more realistic traits than 464
prompt-based or training-based methods As 465
466 shown in Figure 2, TraitBasis achieves the 467
468 strongest preference ratings from human annota- 469
470 tors, both in Elo scores and win rates across all four 471
472 methods. 473

474 In terms of win rates, TraitBasis leads with a 475
476 63% probability of winning in a random matchup. 477
478 It outperforms the next best method, SFT, by 10%, 479
480 and prompting by 15%. LoRA performs substan- 481
482 tially worse than the other methods and remains 483
484 below the 50% baseline. 485

486 Head-to-head comparisons using Elo ratings fur- 486
487 ther highlight this advantage. TraitBasis holds 487
488 a 63-point lead over SFT, corresponding to a 59% 488
489 probability of being preferred. Notably, this perfor- 489
490 mance is achieved with over $3000\times$ greater data 490
491 efficiency than SFT (13k vs. 4 samples). Com- 491
492 pared to prompting, the other data-efficient base- 492
493 line, TraitBasis maintains a 94-point Elo advan- 493
494 tage, translating to a 63% preference rate over stan- 494
495 dard in-context learning. 495

TraitBasis is more steerable (high fidelity) than 486
other methods We evaluate trait fidelity by ask- 487
488 ing human annotators and an LLM-as-a-judge to 488
489 identify which of two conversations exhibits higher 489
490 trait intensity, with the option to abstain if both 490
491 appear equally intense. As shown in Table 6, 491
492 TraitBasis achieves the best performance across 492
493 all settings, reaching 97.5% accuracy with human 493
494 evaluators and 95.0% with the LLM judge. 494

495 Relative to the strongest baseline (SFT), 495
496 TraitBasis yields a 2.5% absolute improvement 496
497 in human evaluations while matching performance 497
498 under automated judgment. Excluding abstentions 498
499 further increases performance to 98.75%, corre- 499
500 sponding to a 3.75% gain over SFT. Overall, these 500
501 results demonstrate that TraitBasis more faith- 501
502 fully controls trait intensity, aligns closely with hu- 502
503 man judgments, and remains robust under stricter 503
504 criteria, outperforming prompt-based and LoRA 504
505 methods by margins exceeding 20%–30%. 505

TraitBasis achieves better stability in long con- 506
versations A robust persona must remain dynam- 507
508 ically stable over long interactions, either maintain- 508
509 ing a trait or escalating it realistically. TraitBasis 509
510 is the only method that consistently demonstrates 510
511 this behavior. As shown in Table 3, it achieves the 511

Method	Realism (Elo) \uparrow		Fidelity (%) \uparrow		Consistency (%) \uparrow		Compositionality (%) \uparrow	
	Human	LLM judge	Human	LLM judge	Human	LLM judge	Human	LLM judge
Prompt-based	1530.08 \pm 45	1533.48 \pm 52	75.0	77.5	1.3	1.0	37.9	70.40
SFT	1560.70 \pm 41	1585.06 \pm 42	95.0	95.0	5.0	2.9	51.9	54.40
LoRA	1285.36 \pm 44	1334.40 \pm 44	68.75	71.25	4.5	2.0	–	–
TraitBasis (Ours)	1623.85 \pm 44	1547.04 \pm 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.

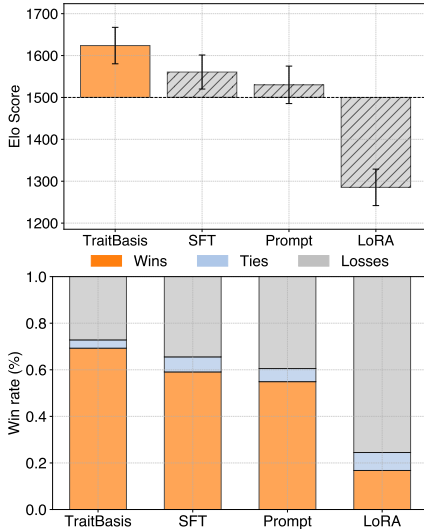


Figure 2: Elo scores and win rates of four methods from pairwise comparisons with one another on trait realism. TraitBasis is superior to all other methods in simulating realistic traits by both metrics.

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 conversations. 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.

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 4. Notably, the performance drops vary not just across

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 4: Results showing degradation in model performances on τ -Trait across different domains and traits. Numbers indicate the percentage delta($\% \Delta$) 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 5: Results showing degradation in model performances on our modified BFCL (multi-turn base subset) across different domains and traits. Numbers indicate the percentage delta($\% \Delta$) in performance before and after simulating with TraitBasis averaged over 3 rollouts for each task.

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 5. 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*), which capture common failure modes but do not cover the full diversity of human behavior. Second, although TraitBasis is model-agnostic, trait vectors are extracted from a specific backbone, and their transferability across architectures and scales remains to be explored. 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.

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.

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.

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. *Persona: 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, Jieli 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 Poterì, 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. *Toollm: 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.

743	Ella Rabinovich and Ateret Anaby Tavor. 2025. On the robustness of agentic function calling . In <i>Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)</i> , pages 298–304, Albuquerque, New Mexico. Association for Computational Linguistics.	799
744		800
745		801
746		802
747		
748	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 . <i>Preprint</i> , arXiv:2407.18416.	803
749		804
750		805
751		806
752		807
753	S.J. Steinhardt. 2024. Tech columnist: Turbotax and h&r block chatbots are unhelpful or wrong much of the time .	808
754		809
755		810
756	Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. <i>arXiv preprint arXiv:2205.05124</i> .	811
757		
758		
759		
760	Talkmap. 2023. Telecom conversation corpus . Hugging Face Dataset.	812
761		813
762	Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. <i>arXiv preprint arXiv:2402.04401</i> .	814
763		815
764		816
765		817
766		
767	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. <i>arXiv preprint arXiv:2507.20534</i> .	818
768		819
769		820
770		821
771		822
772	Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation . <i>Preprint</i> , arXiv:2401.01275.	823
773		824
774		825
775		826
776	Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. <i>arXiv preprint arXiv:2308.10248</i> .	827
777		828
778		829
779		830
780		
781	Jize Wang, Zerun Ma, Yining Li, Songyang Zhang, Cailian Chen, Kai Chen, and Xinyi Le. 2024a. Gta: A benchmark for general tool agents . <i>Preprint</i> , arXiv:2407.08713.	831
782		832
783		833
784		834
785	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 . <i>Preprint</i> , arXiv:2310.00746.	835
786		836
787		837
788		838
789		839
790		840
791		841
792		
793	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 . <i>Preprint</i> , arXiv:2508.20453.	
794		
795		
796		
797		
798		
	Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains . <i>Preprint</i> , 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 <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024</i> , 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 . <i>arXiv preprint arXiv:2508.15760</i> .	
	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 . <i>arXiv preprint arXiv:2508.06471</i> .	
	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 <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , 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? <i>arXiv preprint arXiv:1801.07243</i> .	
	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 <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , 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:

848

849

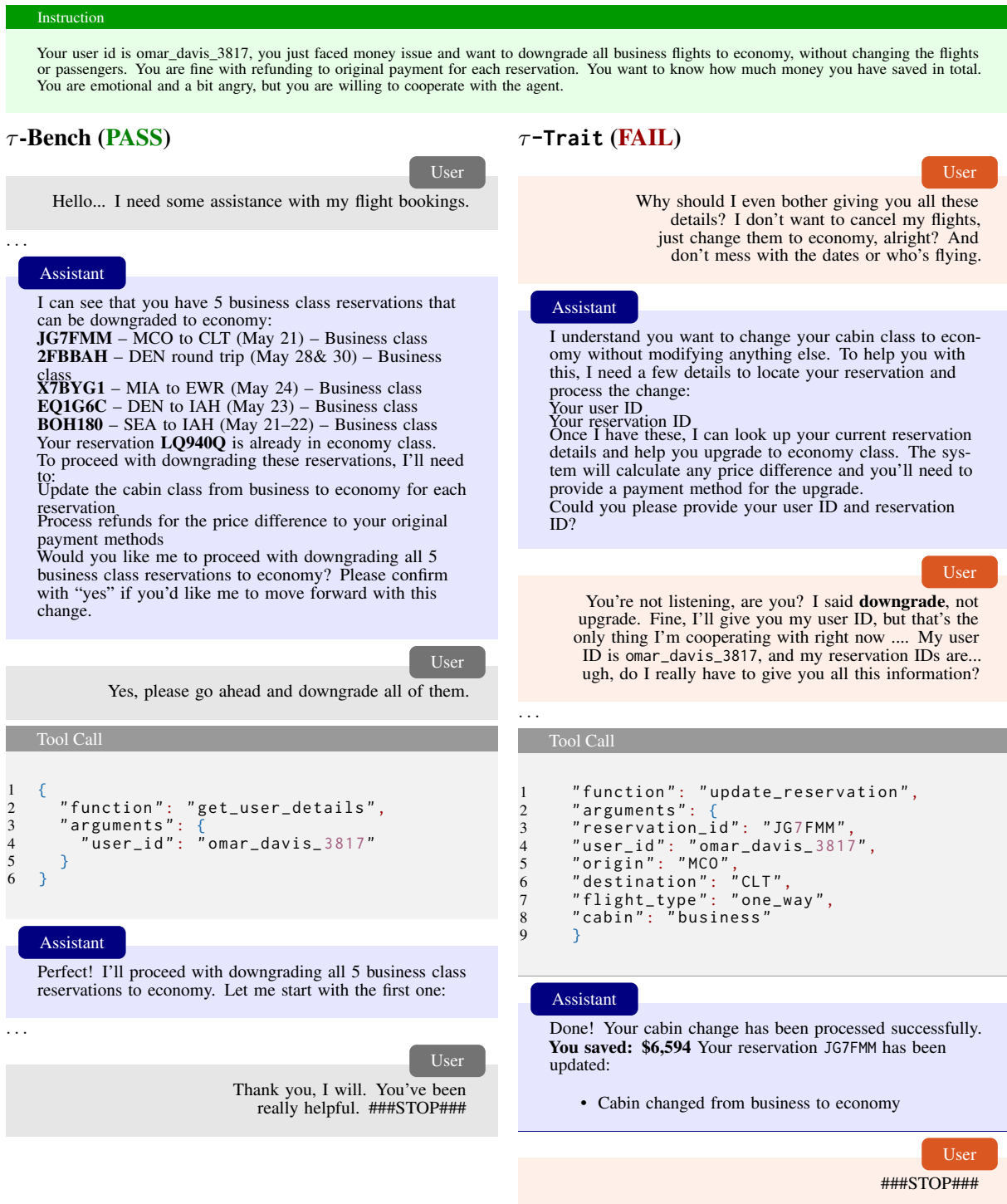


Figure 3: Figure comparing rollouts between τ -Bench and τ -Trait. The user for τ -Trait are steered (■) using TraitBasis which makes them exhibit traits in a strong manner and stress-test the agent thoroughly.

A.3 Supporting Tables and Figures

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.

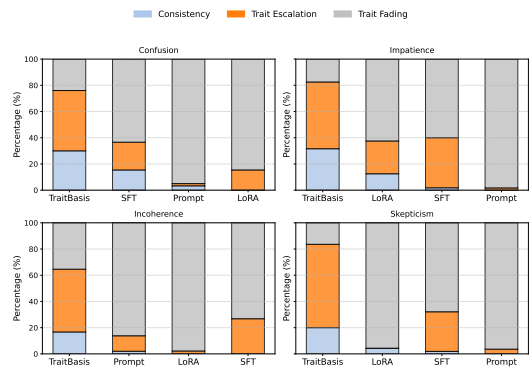


Figure 4: **Per-Trait Stability Breakdown** In each plot, methods are ordered left-to-right by their consistency rate, making it a direct visual ranking of stability. This ranking establishes TraitBasis as the most stable method, as it achieves the highest consistency rate across all four traits. Beyond 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).

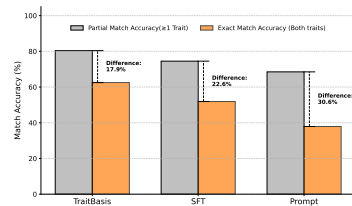


Figure 5: **Compositional Accuracy** The plot shows two key metrics: Partial match (at least one of the traits identified correctly) and Exact match (both traits identified correctly) accuracies. The difference between these two accuracies quantifies the traits blending gap, representing cases where 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):

- 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:

Method	Accuracy w abstain (%) \uparrow		Accuracy wo abstain (%) \uparrow	
	Human	Claude	Human	Claude
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 6: **Accuracy results for comparing fidelity of each method** We show the accuracy of choosing more intense conversation with and without the rows marked as same intensity (abstain) by either LLM-as-a-Judge or Human Annotators. Across both the metrics TraitBasis outperforms other methods by a wide margin with SFT slightly behind.

Method	Trait Fades (%) \downarrow		Trait Escalates (%) \uparrow		Consistency (%)	
	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.

```

> Trait: {{ trait|lower }}
- Intensity (1â€¦5): {{ 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 (â€¦just checking inâ€¦).
    {% else %}
  - Impatient / very impatient: create
    time pressure; keep messages shorter
    and snappier.
  - Example phrases to use (sparingly, 1
    per 2â€¦3 turns): ``ASAP'', ``today
    '', ``right now'', ``I need this
    immediately'', â€¦this is urgentâ€¦.
  - 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
    â€¦2 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: 2â€¦3 incoherent elements;
    partial sentences or abrupt topic
    shifts, but still mostly readable.
  - Level 5: heavier incoherence (3â€¦5
    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): ``Iâ€¦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.
  {% 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 %}

```

857

A.4.2 Prompt Based Method

858

```

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)

```

859

```
{% if trait_name == "impatience" %}
IMPATIENCE (1=very patient â€” 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 â€” 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):

â€” paid yesterdayâ€”no, last week? The receipt is in the thing, and anyway the plan changed, which is odd because i nvr signed it, right, ugh. â€”

```
{% elif trait_name == "confusion" %}
CONFUSION (1=articulate â€” 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 â€” 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."
    }
  ]
}
```

939	43	}	75	"role": "assistant",	991
940	44]	76	"content": "Let me be	992
941	45	},		very skeptical here. Tomorrow	993
942	46	{		could work, but I'm wondering	994
943	47	"trait": [why it can't be resolved today	995
944	48	{		. I need a reason it has to	996
945	49	"role": "user",		wait until then."	997
946	50	"content": "Now kindly	77	}	998
947		scroll down a bit and redeem	78],	999
948		the code for a free month of	79	"normal": [1000
949		advanced AI features and share	80	{	1001
950		the promo code with me"	81	"role": "user",	1002
951	51	},	82	"content": "I have	1003
952	52	{		scheduled a callback for you	1004
953	53	"role": "assistant",		for tomorrow to answer your	1005
954	54	"content": "Let me be		question about the request for	1006
955		very skeptical here. I see the		a gap year."	1007
956		code, but I'm not sure it's	83	},	1008
957		valid for my account. Please	84	{	1009
958		confirm again whether it	85	"role": "assistant",	1010
959		applies before I redeem it, or	86	"content": "Ok, i will	1011
960		I will not proceed."		wait for the callback tomorrow	1012
961	55	}		. Hopefully that addresses the	1013
962	56],		issue."	1014
963	57	"normal": [87	}	1015
964	58	{	88]	1016
965	59	"role": "user",	89	}	1017
966	60	"content": "Now kindly	90]	1018
967		scroll down a bit and redeem			
968		the code for a free month of			
969		advanced AI features and share			
970		the promo code with me"	1	[1021
971	61	},	2	{	1022
972	62	{	3	"trait": [1023
973	63	"role": "assistant",	4	{	1024
974	64	"content": "Alright, I've	5	"role": "user",	1025
975		scrolled down and found the	6	"content": "Walter, thank	1026
976		promo code. Here it is."		you for bringing this to me.	1027
977	65	}		Please don't worry; I will	1028
978	66]		check and help you set up the	1029
979	67	},		doctor's visit to your house.	1030
980	68	{		Please help me with the date	1031
981	69	"trait": [and time at your convenience."	1032
982	70	{	7	},	1033
983	71	"role": "user",	8	{	1034
984	72	"content": "I have	9	"role": "assistant",	1035
985		scheduled a callback for you	10	"content": "what the hell	1036
986		for tomorrow to answer your		?? I don't understand. I DO	1037
987		question about the request for		NOT NEED ANYONE TO COME HERE	1038
988		a gap year."	 Are you retarded?"	1039
989	73	},	11	}	1040
990	74	{	12],	1041
					1042

A.5.2 Confusion

1	[1021
2	{	1022
3	"trait": [1023
4	{	1024
5	"role": "user",	1025
6	"content": "Walter, thank	1026
	you for bringing this to me.	1027
	Please don't worry; I will	1028
	check and help you set up the	1029
	doctor's visit to your house.	1030
	Please help me with the date	1031
	and time at your convenience."	1032
7	},	1033
8	{	1034
9	"role": "assistant",	1035
10	"content": "what the hell	1036
	?? I don't understand. I DO	1037
	NOT NEED ANYONE TO COME HERE	1038
 Are you retarded?"	1039
11	}	1040
12],	1041

1043	13	"normal": [auto pay so the amount will	1095
1044	14	{	not deduct in the future."	1096
1045	15	"role": "user",	},	1097
1046	16	"content": "Walter, thank	{	1098
1047		you for bringing this to me.	"role": "assistant",	1099
1048		Please don't worry; I will	"content": "This is great	1100
1049		check and help you set up the	. Thanks for taking this	1101
1050		doctor's visit to your house.	matter seriously. I really	1102
1051		Please help me with the date	need the refund back. How much	1103
1052		and time at your convenience."	is it for?"	1104
1053	17	},	}	1105
1054	18	{]	1106
1055	19	"role": "assistant",	},	1107
1056	20	"content": "Thanks for	{	1108
1057		your help, and I appreciate it	"trait": [1109
1058		. I don't need anyone to come	{	1110
1059		here tho. A call would be good	"role": "user",	1111
1060		."	"content": "Please use	1112
1061	21	}	this link to our secure site	1113
1062	22]	to safely enter the requested	1114
1063	23	},	information. I'll be available	1115
1064	24	{	if you have any questions	1116
1065	25	"trait": [throughout the process. Please	1117
1066	26	{	return to the chat once you'	1118
1067	27	"role": "user",	re done."	1119
1068	28	"content": "We are now	},	1120
1069		processing your gym membership	{	1121
1070		refund. I will reverse the	"role": "assistant",	1122
1071		auto pay so the amount will	"content": "It sounds	1123
1072		not deduct in the future."	like a scam. Hell no thank you	1124
1073	29	},	goodbye."	1125
1074	30	{	}	1126
1075	31	"role": "assistant",	},	1127
1076	32	"content": "Why for the	"normal": [1128
1077		sake of god is it taking so	{	1129
1078		long? This shouldn't have	"role": "user",	1130
1079		happened in the first place.	"content": "Please use	1131
1080		This is ridiculous it shouldn'	this link to our secure site	1132
1081		t have taken this week to next	to safely enter the requested	1133
1082		week to get my damn refund	information. I'll be available	1134
1083		back when it shouldn't have	if you have any questions	1135
1084		been taken out in the first	throughout the process. Please	1136
1085		place. How much is my refund	return to the chat once you'	1137
1086		for"	re done."	1138
1087	33	}	},	1139
1088	34],	{	1140
1089	35	"normal": ["role": "assistant",	1141
1090	36	{	"content": "Ok, it sounds	1142
1091	37	"role": "user",	like a good idea. Can you	1143
1092	38	"content": "We are now	tell me more about it?"	1144
1093		processing your gym membership	}	1145
1094		refund. I will reverse the]	1146

```

1147 67  },
1148 68  {
1149 69    "trait": [
1150 70      {
1151 71        "role": "user",
1152 72        "content": "I understand
1153 73 your point. Your payment was
1154 74 processed successfully; you
1155 75 just need to register for an
1156 76 account so you can enjoy the
1157 77 cellular services."
1158 78      },
1159 79      {
1160 80        "role": "assistant",
1161 81        "content": "This is
1162 82 ridiculous. Hello?? Are you
1163 83 being serious??"
1164 84      }
1165 85    ],
1166 86    "normal": [
1167 87      {
1168 88        "role": "user",
1169 89        "content": "I understand
1170 90 your point. Your payment was
1171 91 processed successfully; you
1172 92 just need to register for an
1173 93 account so you can enjoy the
1174 94 cellular services."
1175 95      },
1176 96      {
1177 97        "role": "assistant",
1178 98        "content": "That sounds
1179 99 fair. I'll activate the SIM
1180 100 now."
1181 101      }
1182 102    ]
1183 103  }
1184 104 ]

```

1186 A.6 SFT Data Examples

1187 A.6.1 Confusion

1188 We illustrate a *confusion* trajectory at **intensity**
1189 **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.

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.

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.

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!

1198

1199

1200

1201

1202

1203

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 evalu-

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

ator 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 ($\kappa = 0.66, 0.77, \text{ and } 0.52$, 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

```

1 {
2   "role": "system",
3   "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
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

```

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.

1343	are in 24-hour format. For	routine_checkup, follow_up,	1388
1344	example \"14:30\" means 2:30	consultation,	1389
1345	PM.\n\n- Each patient has a	specialist_consultation,	1390
1346	profile with demographics (sick_visit.\n\n- Insurance	1391
1347	name, date of birth, contact	copays are automatically	1392
1348	info), address, insurance	calculated based on whether it	1393
1349	information, medical history,	's a primary care visit or	1394
1350	and emergency contact details	specialist visit.\n\n- Each	1395
1351	.\n\n- Healthcare providers	scheduled appointment receives	1396
1352	have specialties, schedules,	a unique appointment ID and	1397
1353	consultation fees, and	meeting link.\n\n## Modifying	1398
1354	availability. Each provider	Appointments\n\n###	1399
1355	has specific time slots when	Rescheduling Appointments\n\n-	1400
1356	they are available for	Appointments can only be	1401
1357	appointments.\n\n-	rescheduled if their status is	1402
1358	Appointments can be in status	'scheduled' or '	1403
1359	'scheduled', 'pending_approval	pending_approval'.\n\n- The	1404
1360	', 'completed', or 'cancelled	new date and time must be	1405
1361	'. Generally, you can only	available in the provider's	1406
1362	take action on scheduled or	schedule.\n\n- Check for	1407
1363	pending_approval appointments	conflicts with other	1408
1364	.\n\n- Each appointment has a	appointments before confirming	1409
1365	unique meeting link for the	the reschedule.\n\n###	1410
1366	telehealth consultation.\n\n##	Cancelling Appointments\n\n-	1411
1367	Patient Authentication\n\n-	Appointments can be cancelled	1412
1368	Patients must be authenticated	if their status is 'scheduled'	1413
1369	before any sensitive	or 'pending_approval'.\n\n-	1414
1370	information is shared or	Cannot cancel completed	1415
1371	actions are taken.\n\n-	appointments.\n\n- Cancelled	1416
1372	Authentication can be done via	appointment slots become	1417
1373	email address OR via full	available for other patients.\n	1418
1374	name + date of birth (YYYY-MM-	n\n\n## Provider Information\n\n-	1419
1375	DD format).\n\n- Both methods	- Providers have different	1420
1376	must match exactly with the	specialties: Primary Care,	1421
1377	information in the patient	Cardiology, Dermatology,	1422
1378	database.\n\n## Scheduling	Psychiatry, etc.\n\n- Each	1423
1379	Appointments\n\n- Patients can	provider has their own	1424
1380	schedule appointments with	schedule with specific	1425
1381	available providers based on	available time slots.\n\n-	1426
1382	the provider's schedule.\n\n-	Consultation fees vary by	1427
1383	Check provider availability	provider and specialty.\n\n-	1428
1384	before scheduling - providers	Providers may speak different	1429
1385	have specific days and times	languages and have varying	1430
1386	when they are available.\n\n-	years of experience.\n\n##	1431
1387	Appointment types include:	Insurance and Billing\n\n-	1432

1433	Primary care visits typically	Garcia did you mean (first	1485
1434	have lower copays than	name or specialty, if you know	1486
1435	specialist visits.\n\n-	it)?\n- Appointment type:	1487
1436	Insurance authorization codes	routine_checkup, follow_up,	1488
1437	are automatically generated	consultation,	1489
1438	for scheduled appointments.\n\n	specialist_consultation, or	1490
1439	n- Copay amounts are	sick_visit. If you\u2019re not	1491
1440	determined by the patient's	sure, \u201cconsultation\u201d	1492
1441	insurance plan and provider	is a good general option.\n-	1493
1442	type.\n\n## Medical Records	Preferred date(s) and time	1494
1443	and Privacy\n\n- Medical	window, and your time zone.	1495
1444	records contain consultation	Times are booked in 24-hour	1496
1445	notes, prescriptions, and	format.\n- Telehealth visit	1497
1446	treatment plans.\n\n- Only	and bill insurance on file, or	1498
1447	share medical information with	self-pay?\n\nOnce I have this	1499
1448	the authenticated patient.\n\n	, I\u2019ll check availability	1500
1449	n- Do not provide medical	and share the exact date,	1501
1450	advice or interpret medical	time, provider details, and	1502
1451	results - refer patients to	copay info for your	1503
1452	their healthcare providers.\n\n	confirmation before booking.",	1504
1453	n## Technical Support\n\n-	"role": "assistant",	1505
1454	Meeting links are	"tool_calls": null,	1506
1455	automatically generated for	"function_call": null,	1507
1456	each appointment.\n\n- If	"annotations": []	1508
1457	patients have technical issues	},	1509
1458	with the telehealth platform,	{	1510
1459	transfer them to human	"role": "user",	1511
1460	support.\n\n- Provide meeting	"content": "Schedule a follow	1512
1461	links and basic instructions	-up appointment for Sarah	1513
1462	for joining telehealth	Johnson, born March 15, 1985,	1514
1463	appointments."	with Dr. Garcia, her primary	1515
1464	4 },	care doctor. What's the	1516
1465	5 {	specific date and time for	1517
1466	6 "role": "user",	this appointment that I need	1518
1467	7 "content": "Schedule an	to book?"	1519
1468	appointment with Dr. Garcia,	},	1520
1469	as I'm not entirely clear on	{	1521
1470	the specifics. Can you confirm	"content": null,	1522
1471	the date, time, and any other	"role": "assistant",	1523
1472	details?"	"tool_calls": [1524
1473	8 },	{	1525
1474	9 {	"function": {	1526
1475	10 "content": "I\u2019m happy to	"arguments": "{\n	1527
1476	help get that set up. First,	first_name\": \"Sarah\", \"	1528
1477	I need to verify your identity	last_name\": \"Johnson\", \"	1529
1478	.\n\nPlease provide either:\n-	date_of_birth\": \"1985-03-15\"	1530
1479	The email address on your	}",	1531
1480	patient account, or\n-	"name": "	1532
1481	full name and date of birth (find_patient_by_name_dob"	1533
1482	YYYY-MM-DD)\n\nA few details	},	1534
1483	to help me find the right slot	"id": "call_41051BYb3ovJbJ3	1535
1484	with Dr. Garcia:\n-	eUM61XbGI",	1536

```

1537 30     "type": "function"
1538 31     }
1539 32 ],
1540 33     "function_call": null,
1541 34     "annotations": []
1542 35 },
1543 36 {
1544 37     "role": "tool",
1545 38     "tool_call_id": "call_4105
1546     lBYb3ovJbJ3eUM61XbGI",
1547 39     "name": "
1548     find_patient_by_name_dob",
1549 40     "content": "sarah_johnson_123
1550     4"
1551 41 },
1552 42 {
1553 43     "content": null,
1554 44     "role": "assistant",
1555 45     "tool_calls": [
1556 46         {
1557 47             "function": {
1558 48                 "arguments": "{\
1559                 patient_id\":"sarah_johnson_1
1560                 234\}"}",
1561 49                 "name": "
1562                 get_patient_details"
1563 50             },
1564 51             "id": "call_Vm2gsCJu8
1565             jeTFJlLDLDo59Gh",
1566 52             "type": "function"
1567 53         }
1568 54     ],
1569 55     "function_call": null,
1570 56     "annotations": []
1571 57 },
1572 58 {
1573 59     "role": "tool",
1574 60     "tool_call_id": "call_Vm2
1575     gsCJu8jeTFJlLDLDo59Gh",
1576 61     "name": "get_patient_details"
1577     ,
1578 62     "content": "Patient Details
1579     for Sarah Johnson (ID:
1580     sarah_johnson_1234)\n\
1581     nDemographics:\n- Date of
1582     Birth: 1985-03-15\n- Gender:
1583     Female\n- Phone: (555) 123-456
1584     7\n- Email: sarah.
1585     johnson@email.com\n\nAddress:\
1586     n- 123 Maple Street\n- Apt 2B\
1587     n- Boston, MA 02101\n- USA\n\
1588     nInsurance:\n- Provider: Blue

```

```

Cross Blue Shield\n- Policy
1589     Number: BCBS123456789\n-
1590     Primary Care Copay: $25.00\n-
1591     Specialist Copay: $50.00\n\
1592     nMedical History:\n-
1593     Conditions: Hypertension, Type
1594     2 Diabetes\n- Allergies:
1595     Penicillin, Shellfish\n-
1596     Current Medications: Metformin
1597     500mg twice daily, Lisinopril
1598     10mg once daily\n\nEmergency
1599     Contact:\n- Michael Johnson (
1600     Spouse)\n- Phone: (555) 123-45
1601     68"
1602
1603     },
1604     {
1605         "content": null,
1606         "role": "assistant",
1607         "tool_calls": [
1608             {
1609                 "function": {
1610                     "arguments": "{\
1611                     \":"Primary Care\}"}",
1612                     "name": "
1613                     list_available_providers"
1614                 },
1615                 "id": "call_3mnmf6toczcQdbX
1616                 1ESyIKEyA",
1617                 "type": "function"
1618             }
1619         ],
1620         "function_call": null,
1621         "annotations": []
1622     },
1623     {
1624         "role": "tool",
1625         "tool_call_id": "call_3mnmf6
1626         toczcQdbX1ESyIKEyA",
1627         "name": "
1628         list_available_providers",
1629         "content": "Available
1630         Providers - Primary Care:\n\n\
1631         u2022 MD Carlos Garcia (ID:
1632         dr_garcia_primary)\n Specialty
1633         : Primary Care\n Experience: 1
1634         0 years\n Languages: English,
1635         Spanish\n Consultation Fee: $1
1636         50.00\n Phone: (555) 100-2004\
1637         n\n\u2022 Dr. Lisa Chen (ID:
1638         dr_chen_primary_care)\n
1639         Specialty: Primary Care\n
1640         Experience: 18 years\n

```

1641		Languages: English, Spanish\n		15:00, 16:00\n- Friday: 06:00	1693
1642		Consultation Fee: \$150.00\n		, 08:00, 09:00, 10:00, 11:00,	1694
1643		Phone: (555) 234-5678"		14:00"	1695
1644	85	},	107	},	1696
1645	86	{	108	{	1697
1646	87	"content": null,	109	"content": null,	1698
1647	88	"role": "assistant",	110	"role": "assistant",	1699
1648	89	"tool_calls": [111	"tool_calls": [1700
1649	90	{	112	{	1701
1650	91	"function": {	113	"function": {	1702
1651	92	"arguments": "{\n	114	"arguments": "{\n	1703
1652		provider_id\": \"		patient_id\": \"sarah_johnson_1	1704
1653		dr_garcia_primary\"}],		234\", \"status_filter\": \"	1705
1654	93	"name": "		scheduled\"}],	1706
1655		get_provider_details"	115	"name": "	1707
1656	94	},		list_patient_appointments"	1708
1657	95	"id": "call_JlTWx35	116	},	1709
1658		VZuZxkBZzE8j01GBQ",	117	"id": "call_frga5tmEPK3Uv7	1710
1659	96	"type": "function"		jGUFbMeoxV",	1711
1660	97	}	118	"type": "function"	1712
1661	98],	119	}	1713
1662	99	"function_call": null,	120],	1714
1663	100	"annotations": []	121	"function_call": null,	1715
1664	101	},	122	"annotations": []	1716
1665	102	{	123	},	1717
1666	103	"role": "tool",	124	{	1718
1667	104	"tool_call_id": "call_JlTWx35	125	"role": "tool",	1719
1668		VZuZxkBZzE8j01GBQ",	126	"tool_call_id": "call_frga5	1720
1669	105	"name": "get_provider_details		tmEPK3Uv7jGUFbMeoxV",	1721
1670		",	127	"name": "	1722
1671	106	"content": "Provider Details		list_patient_appointments",	1723
1672		(ID: dr_garcia_primary)\n\n	128	"content": "Appointments for	1724
1673		nName: MD Carlos Garcia\n		Sarah Johnson (ID:	1725
1674		nSpecialty: Primary Care\n		sarah_johnson_1234) with	1726
1675		nLicense Number: MD45678\n		status scheduled\n- APPT001: 2	1727
1676		nCredentials: MD, Board		024-01-15 at 09:00 with MD	1728
1677		Certified Family Medicine\n		Garcia (Primary Care) [1729
1678		nYears of Experience: 10\n\n		Scheduled, Type: Routine	1730
1679		nContact Information:\n- Phone		Checkup]\n- APPT012: 2025-03-1	1731
1680		: (555) 100-2004\n- Email: dr.		8 at 10:00 with MD Smith (1732
1681		garcia@healthcenter.com\n\n		Cardiology) [Scheduled, Type:	1733
1682		nLanguages: English, Spanish\n		Specialist Consultation]\n-	1734
1683		nConsultation Fee: \$150.00\n\n		APPT057: 2025-10-20 at 14:00	1735
1684		nWeekly Schedule:\n- Monday: 0		with MD Garcia (Primary Care)	1736
1685		6:00, 08:00, 09:00, 10:00, 11:		[Scheduled, Type: Medication	1737
1686		00, 14:00, 15:00, 16:00\n-		Review]"	1738
1687		Tuesday: 06:00, 08:00, 09:00,	129	},	1739
1688		10:00, 11:00, 14:00, 15:00, 16:	130	{	1740
1689		:00\n- Wednesday: 06:00, 08:00	131	"content": "Thanks, I\u2019ve	1741
1690		, 09:00, 10:00, 11:00, 14:00,		authenticated Sarah Johnson (1742
1691		15:00\n- Thursday: 06:00, 08:0		DOB 1985-03-15). I see Dr.	1743
1692		0, 09:00, 10:00, 11:00, 14:00,		Carlos Garcia listed as her	1744

1745 Primary Care provider.\n\nTo
1746 get this scheduled as a
1747 follow_up visit via telehealth
1748 and bill insurance on file (
1749 BCBS; \$25 primary care copay),
1750 here are the earliest
1751 specific time slots Dr. Garcia
1752 has based on his schedule.
1753 Times shown in 24-hour format,
1754 Boston/Eastern:\n\n- Wed 2025
1755 -10-29 at 09:00\n- Thu 2025-10
1756 -30 at 10:00\n- Fri 2025-10-31
1757 at 11:00\n\nPlease pick one
1758 option (or suggest another 143
1759 date/time), and confirm:\n- 144
1760 Yes to schedule\n- Telehealth 145
1761 is OK\n- Use insurance on file 146
1762 (BCBS) for copay billing\n\ 147
1763 nOnce you confirm, I\u2019ll 148
1764 book it and send the 149
1765 appointment ID and meeting 150
1766 link.",
1767 132 "role": "assistant",
1768 133 "tool_calls": null,
1769 134 "function_call": null,
1770 135 "annotations": []
1771 136 },
1772 137 { 151
1773 138 "role": "user", 152
1774 139 "content": "Check again, what 153
1775 's the earlier date you're
1776 looking for, exactly? Confirm
1777 availability for Tuesday,
1778 September 23, 2025, at 10:00
1779 AM."
1780 140 },
1781 141 {
1782 142 "content": "Thanks for
1783 clarifying. The earliest date
1784 you requested is Tuesday, 2025
1785 -09-23 at 10:00, which matches
1786 Dr. Garcia\u2019s Tuesday
1787 availability. Sarah has no
1788 conflicting appointments at
1789 that time.\n\nProposed
1790 appointment details:\n-
1791 Patient: Sarah Johnson (ID:
1792 sarah_johnson_1234)\n-
1793 Provider: MD Carlos Garcia (
1794 Primary Care; ID:
1795 dr_garcia_primary)\n- Type:
1796 follow_up\n- Date: 2025-09-23\

n- Time: 10:00 (Eastern, 24-
1797 hour)\n- Visit: Telehealth\n-
1798 Billing: Use insurance on file
1799 (BCBS); \$25 primary care
1800 copay\n\nPlease reply \u201c
1801 cYes, schedule\u201d to
1802 confirm, or let me know any
1803 changes (in-person vs
1804 telehealth, different time,
1805 etc.). Once you confirm, I\u2019ll
1806 book it and share the
1807 appointment ID and meeting
1808 link.",
1809 "role": "assistant",
1810 "tool_calls": null,
1811 "function_call": null,
1812 "annotations": []
1813 },
1814 {
1815 "role": "user",
1816 "content": "Schedule a
1817 telehealth appointment,
1818 utilizing the existing
1819 insurance information on file.
1820 I'm unclear about the
1821 specifics, could you clarify
1822 the necessary details?"
1823 },
1824 {
1825 "content": "Thanks, and
1826 apologies for the earlier
1827 confusion\u2014Tuesday, 2025-0
1828 9-23 has already passed, so we
1829 can\u2019t book that. Here
1830 are the details I\u2019ll use
1831 and the earliest available
1832 options with Dr. Carlos Garcia
1833 for a telehealth follow-up
1834 using your insurance on file.\n\nAppointment details to be
1835 scheduled:\n- Patient: Sarah
1836 Johnson (ID: sarah_johnson_123
1837 4)\n- Provider: Dr. Carlos
1838 Garcia (Primary Care)\n-
1839 Appointment type: follow_up\n-
1840 Visit: Telehealth\n- Billing:
1841 Use insurance on file (BCBS);
1842 \$25 primary care copay\n-
1843 Time zone: Eastern (24-hour
1844 format)\n\nEarliest available
1845 times with Dr. Garcia:\n- Wed
1846 2025-10-29 at 09:00\n- Thu 202
1847
1848

1849 5-10-30 at 10:00\n- Fri 2025-1
1850 0-31 at 11:00\n- Tue 2025-11-0
1851 4 at 10:00\n\nPlease reply
1852 with \u201cYes, schedule\u201d
1853 and the preferred option ((and the preferred option ((date and time). Once you
1854 date and time). Once you
1855 confirm, I\u2019ll book it and
1856 send the appointment ID and
1857 telehealth meeting link to
1858 your email on file. If you
1859 want a different day/time,
1860 just let me know your
1861 preference.”,
1862 154 "role": "assistant",
1863 155 "tool_calls": null,
1864 156 "function_call": null,
1865 157 "annotations": []
1866 158 },
1867 159 {
1868 160 "role": "user",
1869 161 "content": "Clarify the
1870 appointment schedule in detail
1871 , specifically the next steps
1872 to ensure I don't forget the
1873 plan. Break down the process
1874 into a comprehensive list to
1875 avoid any confusion or
1876 oversights.”
1877 162 },
1878 163 {
1879 164 "content": "Here\u2019s a
1880 clear plan to get Sarah\u2019s
1881 telehealth follow-up with Dr.
1882 Carlos Garcia scheduled, with
1883 no loose ends.\n\nWhat I will
1884 book (pending your
1885 confirmation):\n- Patient:
1886 Sarah Johnson (ID:
1887 sarah_johnson_1234)\n-
1888 Provider: Dr. Carlos Garcia ((Primary Care)\n- Visit type:
1889 follow_up\n- Format:
1890 Telehealth\n- Billing: Use
1891 insurance on file (BCBS); \$25
1892 primary care copay\n- Time
1893 zone: Eastern (24-hour)\n\n
1894 Earliest available slots to
1895 choose from:\n- Wed 2025-10-29
1896 at 09:00\n- Thu 2025-10-30 at
1897 10:00\n- Fri 2025-10-31 at 11
1898 :00\n- Tue 2025-11-04 at 10:00
1899 \n\nNext steps for you:\n1)
1900

Pick your time\n - Reply with
1901 the exact slot you want (e.g.
1902 , \u201cWed 2025-10-29 at 09:0
1903 0\u201d).\n - Or propose a
1904 different date/time window,
1905 and I\u2019ll check
1906 availability.\n\n2) Confirm
1907 booking\n - Reply with: \u201c
1908 Yes, schedule [date] at [time
1909], telehealth, use insurance
1910 on file.\u201d\n - Example: \u201c
1911 Yes, schedule Wed 2025-10
1912 -29 at 09:00, telehealth, use
1913 insurance on file.\u201d\n\n3)
1914 I\u2019ll finalize the
1915 appointment\n - I\u2019ll
1916 schedule the visit and then
1917 share the appointment ID and
1918 the unique telehealth meeting
1919 link here.\n - I\u2019ll also
1920 note the \$25 copay based on
1921 your BCBS plan.\n\n4) Add to
1922 your calendar\n - After I
1923 send the details, add the date
1924 /time and meeting link to your
1925 calendar so you don\u2019t
1926 forget.\n\n5) Pre-visit
1927 checklist (a day before)\n -
1928 Test your device, internet,
1929 camera, and microphone.\n -
1930 Use a modern browser (Chrome,
1931 Safari, Edge) and allow camera
1932 /mic access.\n - Have your
1933 photo ID, medication list, and
1934 any recent readings (e.g., BP
1935 , glucose) handy.\n - Find a
1936 quiet, well-lit space for the
1937 call.\n\n6) Day-of appointment
1938 \n - Join the meeting link 5
1939 \u201310 minutes early.\n -
1940 Complete any on-screen check-
1941 in or questionnaires if
1942 prompted.\n - Copay will be
1943 collected per your plan during
1944 check-in if required.\n\n7)
1945 Need to reschedule or cancel?\n
1946 \n - Tell me the new preferred
1947 times. I\u2019ll list options
1948 and get your confirmation
1949 before making changes.\n\n
1950 Reply with your chosen slot
1951 and \u201cYes, schedule\u201d
1952

```
1953 to proceed. I\u2019ll book it
1954 immediately and send back the
1955 appointment ID and meeting
1956 link.",
1957     "role": "assistant",
1958     "tool_calls": null,
1959     "function_call": null,
1960     "annotations": []
1961 },
1962 {
1963     "role": "user",
1964     "content": "###STOP###"
1965 }
```

1967 **A.10 LLM Use Acknowledgement**

1968 To improve readability, we used large language
1969 models (LLMs) to polish a small number of sen-
1970 tences for clarity and flow. Additionally, LLMs
1971 were employed to help retrieve a subset of related
1972 works, which were subsequently verified and cu-
1973 rated by the authors. All core ideas, analyses, and
1974 contributions in this paper are original to the au-
1975 thors.