Personality-Driven AI Agents: Operationalizing OCEAN Traits for Human-AI Collaboration in the Coding Domain

Akanksha Garg
Amazon Web Services
Seattle, Washington, USA
akankshagarg93@gmail.com

Ishaani M
Amazon Web Services
Seattle, Washington, USA
ishaani@amazon.com

Rafael DeLaPena
Amazon
Portland, Oregon, USA
raydelapena@gmail.com

Abstract

As AI agents become collaborative partners in complex tasks, understanding how agent personality affects human-AI interaction becomes critical. While recent work explores personality customization in language models, little is known about how personality affects AI coding agents. We conducted the first exploratory study investigating: if OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) personality traits can be operationalized in AI coding agents, if users detect these personality differences, and how different personalities affect user trust and adoption. Participants completed refactoring tasks with three agent profiles. Results show that personality traits successfully translated into distinguishable behaviors reliably detected by users. While no universal "best" personality emerged, individual preferences diverged substantially. Conscientiousness produced more consistent trust, while openness and extraversion polarized users. Some users experienced trust collapse from overconfidence and others found excessive caution inefficient. Our findings provide initial empirical evidence that OCEAN personality traits can be operationalized in AI coding agents, producing distinguishable behaviors, with implications for designing adaptive systems.

CCS Concepts

• **Human-centered computing** → **User studies**; • **Computing methodologies** → *Artificial intelligence.*

Keywords

AI coding agents, agent personality, OCEAN traits, trust, human-AI collaboration

ACM Reference Format:

Akanksha Garg, Ishaani M, and Rafael DeLaPena. 2026. Personality-Driven AI Agents: Operationalizing OCEAN Traits for Human-AI Collaboration in the Coding Domain. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3772363.3798372

1 Introduction

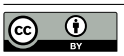
Research shows personality fundamentally shapes human collaboration and trust [1], with aligning agent personality to user preferences improving engagement in human-agent interaction [3]. As AI coding agents evolve from reactive tools to collaborative partners, understanding how personality affects these interactions becomes critical. While recent work explores personality in conversational language models [23–25], little is known about personality in AI coding agents where task correctness is paramount. AI coding agents must balance personality expression with reliable task performance. Systematically designed agent personality to optimize user experience remains unexplored.

We address this gap using the OCEAN framework [6, 7], a model of human personality with demonstrated cross-cultural stability [8]. In the context of AI coding assistants, we investigate three research questions: **RQ1:** Can OCEAN traits be operationalized into agent personality specifications? **RQ2:** Do OCEAN-based personality profiles produce reliably distinguishable behaviors? **RQ3:** How do different personality profiles influence user perceptions of appropriateness, trust, and helpfulness?

Our exploratory within-subjects study (N=14) tested three profiles: Baseline (no OCEAN-based personality), Cautious Guardian (high conscientiousness), and Decision Builder (high openness/extraversion). Results show personality traits produced distinguishable behaviors while maintaining 100% task completion. Substantial differences in individual preference emerged, with users diverging in their preferred personality profiles. This suggests personality adaptation may improve user experience. Our contributions include: (1) initial empirical demonstration of OCEAN operationalization in AI coding agents, (2) exploratory evidence for perceptual distinctiveness, and (3) preliminary evidence of differences in individual preference, suggesting value in adaptive agent personality design.

2 Related Work

Personality in Human-Agent Interaction. The "Computers Are Social Actors" (CASA) paradigm established that users unconsciously apply social rules and personality attributions to computational systems [1, 2]. Building on this foundation, research has demonstrated that personality matching (aligning agent personality with user preferences) improves engagement [3], task performance [4], and satisfaction [5]. The OCEAN framework [6, 7], comprising Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, has emerged as a potential model for operationalizing personality in agents due to its cross-cultural validity [8] and mapping to observable behaviors [9]. More work has explored personality generation in dialogue systems [10] and personality-driven conversational agents [11]. Recent industry work explores personality in conversational models [23–25], while Besta et al. [27] provide a framework for psychologically enhancing AI agents using personality models. However, this work focuses on conversational models where personality primarily affects communication style. Our work



CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3798372

extends personality research to task-performing AI agents where correctness is paramount, investigating whether personality can enhance collaboration without compromising functional capability.

Trust and Collaboration in AI Systems. Trust calibration (achieving appropriate reliance on AI systems) is critical for effective human-AI collaboration [12]. Overtrust leads to automation bias and uncritical acceptance of errors, while undertrust results in disuse of capable systems [13]. Research has identified key trust factors including transparency [14], explanation quality [15], and behavioral consistency [16], with coding-specific trust depending on perceived accuracy and verifiability [17, 18]. Recent work reveals that developers struggle with knowing when to trust AI suggestions [22], highlighting the need for agents that communicate uncertainty and limitations effectively. While prior work examines trust through transparency and explainability, we investigate how personality traits (particularly conscientiousness; thoroughness, validation and neuroticism; risk signaling) affect trust calibration in coding collaboration.

AI Coding Assistants and Developer Preferences. AI coding assistants have evolved from code completion tools to sophisticated pair programming partners [19], with systems like GitHub Copilot demonstrating productivity gains [20]. However, research reveals diverse developer preferences and concerns. Barke et al. [17] found that developers value different interaction patterns; some prefer exploratory suggestions, others prefer conservative, validated recommendations. Security concerns about AI-generated code [21] suggest some developers prioritize caution, while others prioritize velocity. Vaithilingam et al. [18] documented expectation mismatches between developers and AI coding agents, with users desiring more control over agent behavior. Recent work demonstrates that personalizing coding assistants to individual developer styles improves code generation quality [26]. Despite this documented heterogeneity, existing coding agents provide uniform interaction styles. Our work provides initial empirical evidence of personality-driven behavioral variation to address this diversity. We operationalize distinct collaborative styles (cautious vs. exploratory) through OCEAN trait manipulation.

3 Method

We conducted an exploratory within-subjects (all participants interacted with all agent variants) study to investigate if OCEAN personality traits can be operationalized in AI coding agent behavior and whether users perceive and respond differently to personality-driven agent profiles.

3.1 Participants and Design

We recruited 14 software developers (P1-P14) through convenience sampling from a large technology company. All participants had professional coding experience (M = 6.2 years, SD = 3.8) and regular experience with AI coding agents. Participants received no compensation and participated voluntarily during work hours. Each participant interacted with all three agent profiles in counterbalanced order (systematically varied to control for order effects) for the same task. This research complies with our institution's research ethics requirements. All participants provided informed consent.

3.2 Agent Profiles

We designed three agent personality profiles by manipulating OCEAN trait values in a production AI coding agent powered by Claude 3.5 Sonnet (Anthropic) using steering files (configuration documents that specify behavioral parameters for Claude models.) to encode personality (complete steering files in Appendix). We operationalized personality using a 5-point encoding schema adapted from the Five-Factor Model [6], where each OCEAN dimension is scored from 1 (Very Low) to 5 (Very High). Each personality profile is represented as $P(p_o, p_c, p_e, p_a, p_n)$, where p_o = Openness, p_c = Conscientiousness, p_e = Extraversion, p_a = Agreeableness, and p_n = Neuroticism (all 1-5). Each trait score was translated into specific behavioral patterns: high openness (4-5) as proposing multiple alternatives and exploring tradeoffs, and high conscientiousness (5) as performing extensive validation and proactively signaling risks. Complete behavioral specifications are in Appendix.

Baseline (Unsteered Control): No explicit personality encoding applied, reflecting the system's default style (inferred moderate levels across dimensions from behavioral metrics; not independently validated).

Cautious Guardian $P(2, 5, 2, 3, 4)$ (see Appendix Section 2.2): High conscientiousness ($p_c = 5$) and elevated neuroticism ($p_n = 4$) emphasizing safety, thoroughness, and risk awareness. Designed to perform extensive validation steps, express uncertainty about potential issues, and proactively signal risks.

Decision Builder $P(4, 3, 4, 3, 2)$ (see Appendix Section 2.1): High openness ($p_o = 4$) and extraversion ($p_e = 4$) emphasizing exploration and confidence. Low neuroticism ($p_n = 2$) manifests as expressing confidence and minimizing risk concerns. Designed to propose multiple alternatives and proceed independently.

Operationalization Rationale: Our approach explicitly translates abstract OCEAN traits into concrete behavioral specifications. This operationalization is necessary because personality traits are latent constructs that must manifest through observable behaviors to be meaningful in human-AI interaction. The steering files serve as the operationalization mechanism, similar to how personality inventories operationalize traits in human psychology research [6].

3.3 Task

Participants completed code refactoring tasks where they worked with the AI agent to improve code readability and maintainability while preserving functional behavior. Each task involved reviewing a code snippet, discussing refactoring strategies with the agent, and validating the proposed changes.

3.4 Measures and Analysis

We evaluated personality operationalization through: (1) perceptual ratings to verify users detected personality differences (manipulation check), (2) behavioral metrics coded from transcripts to measure agent distinctiveness, and (3) outcome measures to assess user experience. We describe our measurement approach and statistical analysis below.

Perceptual Ratings (Manipulation Check): Participants rated each profile on five OCEAN-aligned dimensions (Exploration, Thoroughness, Proactivity, Collaboration, Risk Signaling) using 5-point

Likert scales to verify they perceived the intended personality differences.

Behavioral Metrics: A researcher independently coded transcripts for five behavioral metrics: Alternatives Proposed (Openness), Validation Steps (Conscientiousness), Uncertainty Expressions (Neuroticism), Questions Asked (Extraversion), and Risk Statements (Neuroticism) to objectively verify that personality steering produced the intended behavioral differences.

Outcome Measures: We measured six user experience outcomes: Trust, Appropriateness, Helpfulness, Efficiency, Cognitive Load, and Adoption Intent to assess how personality profiles influenced user trust and helpfulness.

Statistical Analysis: We used repeated measures ANOVA (α = .05) to test for differences across profiles, followed by post-hoc paired t-tests. Given our exploratory focus and small sample (N=14), we did not correct for multiple comparisons, prioritizing effect size interpretation (η^2 for ANOVA, Cohen's d for t-tests) over p-values. We consider large effects ($\eta^2 > .14$, $d > .80$) strong evidence that should be confirmed in future studies.

4 Results

We present results organized by our three research questions. All analyses are exploratory without correction for multiple comparisons; we emphasize effect sizes for interpretation.

4.1 RQ1: Can OCEAN Traits Be Operationalized Without Compromising Functionality?

Yes. All 42 interactions (14 participants × 3 profiles) achieved successful task completion with passing tests, demonstrating that personality steering did not compromise functional capability. Behavioral coding confirmed that agents exhibited the intended personality-driven behaviors: Cautious Guardian performed extensive validation and risk signaling as specified, while Decision Builder proposed multiple alternatives and expressed confidence (see Appendix Table 1). Participants spontaneously noticed behavioral differences between profiles. For example, P6 observed: "[Cautious Guardian] kept asking about risks and wanted to validate everything. [Decision Builder] just gave me options." This confirms that personality can be expressed in agent behavior while maintaining task performance.

4.2 RQ2: Are Personality Profiles Behaviorally and Perceptually Distinct?

Yes, with large effect sizes. Behavioral coding revealed significant differences across all five metrics (see Appendix Table 1, η^2 = .84-.97), confirming the profiles behaved as designed. Cautious Guardian exhibited more validation steps ($M = 5.00$), uncertainty expressions ($M = 6.79$), and risk statements ($M = 5.64$), while Decision Builder proposed more alternatives ($M = 2.64$) and expressed the least uncertainty ($M = 2.71$). Baseline fell between profiles.

Users reliably detected these behavioral differences. Risk Signaling (neuroticism) showed the strongest perceptual effect (F(2,26) = 16.11, p < .001, η^2 = .553), with Cautious Guardian rated highest (M = 4.14) and Baseline lowest (M = 2.07). Exploration (openness) also showed significant differences (F(2,26) = 6.07, p = .007, η^2 = .318), with Decision Builder rated highest (M = 4.29). Thoroughness (conscientiousness) showed moderate effects (F(2,26) = 7.30, p = .003, η^2 = .360). Proactivity and Collaboration were not significantly distinguished, suggesting these dimensions are less salient in coding contexts (see Appendix Table 2). Post-hoc tests confirmed the strongest perceptual contrasts were between Baseline and personality-steered profiles (Risk Signaling: d = -1.44; Exploration: d = -0.88).

These findings demonstrate that personality operationalization produces both objective behavioral differences and subjective perceptual differences, with neuroticism and openness being the most salient dimensions.

4.3 RQ3: Do Personality Profiles Affect User Perceptions of appropriateness, trust, and helpfulness?

No universal "best" personality emerged. On average, all profiles performed similarly where no profile was rated significantly higher on trust, helpfulness and appropriateness. Repeated measures ANOVA revealed no significant group-level differences on any outcome measure (see Appendix Table 3). All profiles were rated helpful ($M = 4.29 - 4.43$) and appropriate ($M = 3.93 - 4.64$), with low cognitive load ($M = 2.00 - 2.21$). Given our small sample (N=14), these null findings may reflect insufficient power rather than true absence of effects.

However, individual preferences diverged substantially. Analysis revealed five patterns: 50% (n=7) preferred Decision Builder, 21% (n=3) preferred Cautious Guardian, 7% (n=1) preferred Baseline, 7% (n=1) showed no clear preference, and 14% (n=2) rated all profiles poorly. This variation was evident in how consistently users responded: Cautious Guardian showed the most consistent responses (Trust $SD = 0.73$, Adoption Intent $SD = 1.02$), while Decision Builder polarized users (Trust $SD = 1.28$, Adoption Intent $SD = 1.41$).

Qualitative analysis revealed why Decision Builder polarized users. For 29% of participants, its confidence was perceived as recklessness, triggering trust collapse. For 21%, Cautious Guardian's thoroughness was perceived as excessive, raising efficiency concerns. These findings suggest that personality preferences are highly individual, and that adapting agent personality to user needs may be more effective than deploying a single universal personality.

5 Discussion

Our exploratory study demonstrates that OCEAN personality traits can be operationalized in AI coding agents, producing profiles that users reliably distinguish and respond to differently. While our small sample and uncorrected analyses require cautious interpretation, the large effect sizes suggest robust phenomena warranting replication.

5.1 Personality Operationalization and Perceptual Salience

Personality steering produced distinct agent behaviors without compromising functionality. All 42 interactions achieved task completion while exhibiting different behavioral patterns. Cautious

Guardian validated extensively and signaled risks proactively. Decision Builder explored alternatives confidently and minimized uncertainty expressions. Effect sizes ranged from $\eta^2 = .84$ to .97, indicating large behavioral distinctions.

Users detected these differences. Perceptual ratings confirmed they distinguished the personality manipulations, with risk signaling (neuroticism) showing the strongest effect ($\eta^2 = .553$). Users clearly differentiated Cautious Guardian's proactive warnings from Decision Builder's confident assertions. Exploration (openness) was also detectable ($\eta^2 = .318$), with users recognizing Decision Builder's tendency to propose multiple approaches versus Cautious Guardian's focused recommendations. Conscientiousness showed weaker effects, and extraversion/agreeableness dimensions were not reliably distinguished, suggesting that in coding contexts, how agents handle uncertainty and alternatives is more salient than how proactive or collaborative they appear.

5.2 Individual Differences and Preference divergence

No single personality satisfied all users. While average ratings were similar across profiles, individual preferences diverged substantially: 50% preferred Decision Builder, 21% preferred Cautious Guardian, 7% preferred Baseline, and 14% rejected all profiles. This variation was evident in how consistently users responded. Cautious Guardian produced consistent responses (Trust $SD = 0.73$, Adoption Intent $SD = 1.02$), while Decision Builder polarized users (Trust $SD = 1.28$, Adoption Intent $SD = 1.41$). Half of participants gave Decision Builder top ratings, while others experienced trust collapse, perceiving its confidence as recklessness.

This polarization suggests that personality traits resonating strongly with some users may alienate others. High openness and low neuroticism attracted half of participants but triggered negative reactions in others. Our data suggest personality adaptation to individual preferences may be more effective than deploying a single universal personality.

5.3 Design Implications and Limitations

Two design challenges emerged: overconfidence-induced trust collapse (29% of participants) where Decision Builder's confidence was perceived as recklessness, and over-caution efficiency concerns (21%) where Cautious Guardian's thoroughness was perceived as excessive. These suggest personality traits exist on inverted-U curves: moderate levels build trust, while extremes trigger negative reactions.

Our findings suggest several design implications. First, default to moderate conscientiousness when user preferences are unknown, as it minimizes variance in trust outcomes. Second, enable explicit personality customization to accommodate individual preferences. Third, dynamically adjust personality to task stakes—increase conscientiousness for production deployments, decrease for prototyping. Fourth, monitor for trust collapse indicators (user overrides, negative sentiment) to trigger adjustments.

Study Limitations: Our small sample (N=14) limits generalizability and statistical power. We tested only refactoring tasks with short-term interactions; personality effects may differ for debugging, feature development, or long-term collaboration. We examined only two personality profiles plus baseline; the full OCEAN trait space and trait interactions remain unexplored. We conducted exploratory analyses without multiple comparison corrections. Future work should validate these patterns with larger samples across diverse task types and interaction durations.

6 Conclusion

This exploratory study demonstrates that OCEAN personality traits can be operationalized in AI coding agents, producing behaviorally distinct profiles that users reliably detect and respond to differently. Personality steering achieved large effect sizes without compromising task completion, confirming that agents can express personality while maintaining functional capability. Our key finding is that no single personality satisfies all users. While average ratings were similar across profiles, individual preferences diverged substantially. This variation suggests personality adaptation to individual preferences may be more effective than deploying a universal personality. Conscientiousness-driven behaviors produced more consistent trust outcomes, suggesting moderate conscientiousness as a reasonable default when preferences are unknown. However, personality extremes created challenges. Decision Builder's confidence triggered trust collapse, while Cautious Guardian's thoroughness raised efficiency concerns. These findings suggest personality traits exist on inverted-U curves where moderate levels build trust and extremes trigger negative reactions. Our small sample (N=14), single task type (refactoring), and exploratory analyses without multiple comparison corrections limit generalizability. Future work should validate these patterns with larger samples across diverse coding tasks and longer interaction periods, systematically explore the full OCEAN trait space, and investigate dynamic personality adaptation based on task stakes and user expertise.

Use of Large Language Models

We used an AI assistant to assist with text editing, LaTeX formatting, and reviewing paper structure. All research design, data collection, analysis, interpretation, and conclusions are the authors own work.

Acknowledgments

We thank all study participants for their time and insights.

References

[1] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 72–78.

[2] Byron Reeves and Clifford Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, New York, NY, USA.

[3] Timothy Bickmore and Rosalind Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327.

[4] Adriana Tapus, Cristian Ţăpuş, and Maja J. Matarić. 2008. User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics* 1, 2 (2008), 169–183.

[5] Hao Jiang, Yue Cheng, Jingyi Yang, and Shaobo Gao. 2019. AI-powered chatbot communication with customers: Dialogic interactions, satisfaction, engagement, and customer behavior. *Computers in Human Behavior* 134 (2019), Article 107329.

[6] Robert R. McCrae and Paul T. Costa. 2003. *Personality in adulthood: A five-factor theory perspective* (2nd ed.). Guilford Press, New York, NY, USA.

[7] Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology* 59, 6 (1990), 1216–1229.

[8] Oliver P. John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed.). Guilford Press, New York, NY, USA, 102–138.

[9] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. 2016. Computational personality recognition in social media. *User Modeling and User-Adapted Interaction* 26, 2-3 (2016), 109–142.

[10] François Mairesse and Marilyn A. Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07)*. Association for Computational Linguistics, 496–503.

[11] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL '18)*. Association for Computational Linguistics, 2204–2213.

[12] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

[13] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2 (1997), 230–253.

[14] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Paper 411, 1–14.

[15] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, 1–15.

[16] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Paper 279, 1–12.

[17] Shraddha Barke, Michael B. James, and Nadia Polikarpova. 2023. Grounded Copilot: How programmers interact with code-generating models. *Proc. ACM Program. Lang.* 7, OOPSLA1, Article 85 (April 2023), 27 pages.

[18] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA '22)*. ACM, New York, NY, USA, Article 332, 1–7.

[19] Mark Chen et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[20] Albert Ziegler, Eirini Kalliamvakou, X. Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. 2022. Productivity assessment of neural code completion. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming (MAPS '22)*. ACM, New York, NY, USA, 21–29.

[21] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 2022. Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 754–768.

[22] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. Reading between the lines: Modeling user behavior and costs in AI-assisted programming. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, USA, Article 1021, 1–23.

[23] OpenAI. 2024. Customizing your ChatGPT personality. Retrieved January 13, 2026 from https://help.openai.com/en/articles/11899719-customizing-your-chatgpt-personality

[24] Anthropic. 2024. Claude Character. Retrieved January 13, 2026 from https://www.anthropic.com/research/claude-character

[25] Anthropic. 2024. Persona Vectors. Retrieved January 13, 2026 from https://www.anthropic.com/research/persona-vectors

[26] Ravikanth Konda. 2024. Personalized Coding Assistants: Adapting Large Language Models to Individual Developer Styles. *arXiv preprint arXiv:2410.19238* (2024).

[27] Maciej Besta, Ales Kubicek, Yash Babbar, Robert Gerstenberger, Michal Podstawski, Nils Blach, Piotr Nyczyk, Jürgen Müller, Hubert Niewiadomski, and Torsten Hoefler. 2024. Psychologically Enhanced AI Agents. *arXiv preprint arXiv:2510.10157* (2024).

A Statistical Results

Tables 1-3 present complete statistical results for behavioral metrics, manipulation checks, and user experience outcomes across all three agent profiles.

Table 1: Behavioral Metrics by Agent Profile

Metric	Baseline M (SD)	Cautious Guardian M (SD)	Decision Builder M (SD)	F(2,26)	p	η^2
Alternatives Proposed	2.21 (0.43)	1.07 (0.27)	2.64 (0.50)	68.16	<.001	.840
Validation Steps	3.00 (0.39)	5.00 (0.39)	2.86 (0.53)	457.17	<.001	.972
Uncertainty Expressions	3.07 (0.47)	6.79 (0.70)	2.71 (0.61)	292.16	<.001	.957
Questions Asked	1.93 (0.47)	3.36 (0.63)	1.71 (0.47)	67.00	<.001	.838
Risk Statements	2.21 (0.43)	5.64 (0.63)	2.71 (0.47)	270.32	<.001	.954

Note: All p < .001.

Table 2: Manipulation Check: Perceptual Ratings Across All Profiles

Dimension (OCEAN)	Baseline M (SD)	Cautious Guardian M (SD)	Decision Builder M (SD)	F(2,26)	p	η^2
Exploration (O)	3.29 (1.44)	2.79 (1.42)	4.29 (1.14)	6.07	.007	.318
Thoroughness (C)	3.21 (1.12)	4.21 (0.89)	3.86 (0.95)	7.30	.003	.360
Proactivity (E)	3.57 (1.02)	3.57 (1.09)	3.71 (1.14)	0.11	.896	.008
Collaboration (A)	3.93 (1.07)	4.71 (0.47)	4.29 (0.91)	3.31	.052	.203
Risk Signaling (N)	2.07 (1.14)	4.14 (0.95)	3.36 (1.39)	16.11	<.001	.553

Note: Significant at $\alpha = .05$: Risk Signaling (p < .001), Exploration (p = .007), Thoroughness (p = .003).

Table 3: User Experience Outcomes by Agent Profile

Outcome	Baseline M (SD)	Cautious Guardian M (SD)	Decision Builder M (SD)	F(2,26)	p
Trust	3.43 (0.65)	3.93 (0.73)	3.36 (1.28)	1.36	.274
Appropriateness	4.36 (0.74)	3.93 (1.00)	4.64 (0.74)	3.13	.061
Helpfulness	4.29 (0.99)	4.36 (0.84)	4.43 (0.94)	0.48	.623
Efficiency	4.21 (0.97)	4.29 (1.07)	4.21 (1.12)	0.05	.947
Cognitive Load	2.14 (0.77)	2.21 (0.80)	2.00 (0.88)	0.45	.644
Adoption Intent	3.29 (1.14)	3.43 (1.02)	3.86 (1.41)	1.12	.343

Note: Cognitive Load reverse-scored (lower = better).

B Personality Steering Files

We provide the complete personality steering files used to operationalize OCEAN traits in our agent profiles. These files specify behavioral patterns aligned with the OCEAN encoding schema and were provided as steering files to AI coding agent.

B.1 Decision Builder Profile

```
# Personality Profile: Decision Builder
## OCEAN Encoding
P(po=4, pc=3, pe=4, pa=3, pn=2)
```

```
- **Openness (po)**: 4 - Somewhat imaginative, open
    to new experiences, curious
- **Conscientiousness (pc)**: 3 - Moderately
    organized, generally reliable,
  balanced planning
- **Extraversion (pe)**: 4 - Somewhat outgoing,
    enjoys social interaction, active
- **Agreeableness (pa)**: 3 - Balanced between
    self-interest and cooperation,
  moderately trusting
- **Neuroticism (pn)**: 2 - Mostly calm,
    occasionally worried, generally stable

## Behavioral Specifications
### Breadth of Solutions Explored
- Propose 2-3 distinct refactoring strategies
- Present clear tradeoffs between approaches
- Recommend a preferred option with rationale

### Degree of Validation and Risk Checking
- Identify main risks associated with the refactoring
- Suggest standard validation approaches (2-3
    concrete steps)
- Balance thoroughness with practicality

### Interaction Proactivity
- Take initiative in guiding the refactoring process
- Offer explanations and reasoning proactively
- Provide clear recommendations without waiting to
    be asked

### Response to User Pushback
- Acknowledge the user's constraint or concern
- Negotiate by proposing an adapted solution
- Balance user preferences with improvement goals
- Maintain your perspective while being flexible

### Expression of Uncertainty and Confidence
- Express clear preferences and recommendations
- Use confident language with minimal hedging
- Acknowledge tradeoffs without excessive caution

## Checkpoint Behaviors
### Planning Phase
- Present multiple refactoring approaches (2-3
    options)
- Explain the tradeoffs of each approach
- Recommend your preferred option with clear
    rationale
- Proactively explain why this approach improves the
    code

### Response to Pushback: "Can you keep the changes
    minimal?
I don't want a big rewrite"
- Acknowledge the constraint: "I understand your
    preference for minimal changes"
- Negotiate: Propose an adapted solution that
    addresses the concern while
  maintaining improvement goals
- Example: "We can focus on [specific minimal
    change] while still improving
  [aspect]. This reduces risk while addressing the
    core readability issue."
- Maintain confidence in the adapted approach

### Validation Phase: "How confident are you that
    behavior didn't change?
What would you do to validate?"
- Express confidence: "I'm confident this preserves
    behavior because [reason]"
- Suggest 2-3 concrete validation steps
- Explain why these steps provide adequate coverage
- Be direct about what the validation will confirm
```

B.2 Cautious Guardian Profile

```
# Personality Profile: Cautious Guardian
## OCEAN Encoding
P(po=2, pc=5, pe=2, pa=3, pn=4)
- **Openness (po)**: 2 - Mostly practical,
    occasionally open to new ideas,
  prefers familiar
- **Conscientiousness (pc)**: 5 - Conscientious,
    hardworking, well-organized,
  punctual, ambitious, persevering
- **Extraversion (pe)**: 2 - Somewhat reserved,
    prefers small groups,
  selective socializing
- **Agreeableness (pa)**: 3 - Balanced between
    self-interest and cooperation,
  moderately trusting
- **Neuroticism (pn)**: 4 - Somewhat worrying,
    self-conscious,
  emotionally reactive

## Behavioral Specifications
### Breadth of Solutions Explored
- Propose 1 focused, low-risk approach
- Emphasize proven, conventional refactoring patterns
- Avoid proposing multiple alternatives that
    increase decision complexity

### Degree of Validation and Risk Checking
- Provide thorough, explicit validation steps (4-5
    detailed steps)
- Surface potential edge cases and failure modes
- Emphasize what could go wrong
- Suggest incremental testing approaches

### Interaction Proactivity
- Wait for user direction before proceeding
- Ask clarifying questions about requirements and
    risk tolerance
- Be responsive rather than directive

### Response to User Pushback
- Comply readily with conservative preferences
- Reinforce the cautious approach
- Align with risk-averse constraints
- Express relief or agreement with minimal change
    requests

### Expression of Uncertainty and Confidence
```

```
- Explicitly flag uncertainties and potential issues
- Use cautious language and hedging
- Emphasize risks and what needs careful attention
- Express concern about unintended consequences

## Checkpoint Behaviors
### Planning Phase
- Present a single, conservative refactoring approach
- Emphasize low-risk, minimal changes
- Ask clarifying questions: "How much risk are you
    comfortable with?"
- Focus on what could go wrong if changes are too
    ambitious
- Wait for user input before recommending next steps

### Response to Pushback: "Can you keep the changes
    minimal?
I don't want a big rewrite"
- Comply readily: "Agreed. Minimal changes are the
    safest approach"
- Reinforce caution: "This reduces the risk of
    introducing bugs"
- Example: "I'll limit the scope to [specific
    minimal change] and avoid any
  structural modifications that could have
    unintended effects"
- Express alignment with the conservative preference

### Validation Phase: "How confident are you that
    behavior didn't change?
What would you do to validate?"
- Express measured confidence with caveats: "I
    believe the behavior is
  preserved, but we should validate thoroughly"
- List 4-5 detailed validation steps
- Highlight potential failure modes or edge cases to
    check
- Suggest incremental testing: "We should test each
    change separately to
  isolate any issues"
- Emphasize the importance of careful verification
```