FROM NATURAL-LANGUAGE DESCRIPTIONS TO OCEAN REPRESENTATIONS FOR DIALOGUE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural-language descriptions provide an intuitive way to specify conversational agent behavior, but they do not directly support systematic control. We introduce a two-stage framework that maps free-form character descriptions to continuous Five-Factor Model (OCEAN) representations and uses them to guide personality expression in dialogue. NaturalOCEAN combines retrieval-augmented language-model inference with post-hoc calibration to estimate Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Retrieval generally reduces uncalibrated prediction error, while calibration substantially improves alignment with reference OCEAN scores. OCEANDialogue learns five trait-specific LoRA components whose contributions are scaled by the corresponding OCEAN values. OCEANDialogue produces stronger and more monotonic graded control than numerical zero-shot prompting for Openness, Conscientiousness, and Extraversion. Cross-trait analysis indicates that the learned controls are only partially disentangled, while human and end-to-end evaluations show that intended personality differences are perceptible and that salient characteristics from the original descriptions remain recognizable in generated dialogue. The results support continuous OCEAN representations as an interpretable interface between natural-language character specification and controllable personality expression in dialogue.

1 INTRODUCTION

Personality describes characteristic patterns in an individual's thoughts and behaviors (Baumert et al., 2017). Linguistic behavior, facial expressions, body motion, and communication style provide cues from which personality impressions can be formed (Tskhay & Rule, 2014; Kodavati et al., 2025; Durupinar et al., 2016; Mehta et al., 2020). Such cues are important for believable conversational agents, where personality expression can shape user perception and interaction (Sonlu et al., 2025; Derner et al., 2024).

Although natural-language character descriptions provide an intuitive way to specify intended behavior, they do not directly provide a representation that can be systematically manipulated. The Five-Factor Model, also known as the Big Five, provides such a representation along five bipolar, continuous dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN) (Goldberg, 1990; McCrae & John, 1992). This motivates two questions: (1) to what extent can OCEAN scores be estimated from free-form character descriptions, and (2) can these representations be used to induce intended personality differences in generated dialogue?

Personality inference from text is challenging because behavioral evidence is context-dependent and large consistently annotated datasets are difficult to obtain (Fang et al., 2023; Zhu et al., 2024). Building on evidence that language models can infer personality traits from text (Derner et al., 2024; Hu et al., 2024), we introduce NaturalOCEAN, a retrieval-augmented method that estimates OCEAN values from free-form character descriptions and calibrates them to psychometric reference scores.

We then introduce OCEANDialogue, which learns five trait-specific Low-Rank Adaptation (LoRA) components whose contributions are scaled by the corresponding OCEAN values. Together, NaturalOCEAN and OCEANDialogue form a two-stage pipeline from natural-language character de-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

scriptions to personality-conditioned dialogue. We evaluate OCEAN inference on a development set and a disjoint confirmatory set, evaluate continuous dialogue control through automatic and human judgments, and assess the complete pipeline end to end.

2 Related Work

2.1 Personality Inference from Language

Language use contains psychological signals that support prediction of Big Five personality traits (Park et al., 2015). Recent work shows that LLMs can infer Big Five traits from human-authored text, although predictions can exhibit systematic biases (Derner et al., 2024). PADO uses personality-induced agents representing contrasting trait perspectives and a judge model to predict OCEAN trait levels (Yeo et al., 2025). NaturalOCEAN extends this direction by retrieving personality-relevant character profiles as contextual evidence for OCEAN estimation.

2.2 Personality-Conditioned Generation and Control

LLMs can be prompted to express Big Five personality characteristics in language and decision-making (Wang et al., 2025; Huang et al., 2026). BIG5-CHAT constructs personality-controlled dialogues and uses them to train models to express personality-related behavioral patterns (Li et al., 2025). Personality information has also been identified and manipulated through model representations, including linear activation directions (Allbert et al., 2024; Chen et al., 2025; Bhandari et al., 2026), while PsychAdapter conditions language models on continuous psychometric representations (Vu et al., 2024). OCEANDialogue instead learns five trait-specific LoRA components whose contributions are directly parameterized by OCEAN values, providing personality control within a shared frozen language model.

3 Methodology

Given a free-form character description x, NaturalOCEAN predicts an OCEAN representation as $\hat{\mathbf{p}} = f_{\mathrm{NO}}(x) \in [0,1]^5$. OCEANDialogue then generates a personality-conditioned response from conversational context c as $y = g_{\mathrm{OD}}(c, \hat{\mathbf{p}})$. This section describes these two components.

3.1 NaturalOCEAN: Continuous Personality Inference

3.1.1 Development Dataset

We use the dataset introduced by Frising & Balcells (2025) as our development set. It contains 406 fictional characters selected from the OpenPsychometrics character catalog spanning 38 media franchises, with each character paired with a textual personality description and continuous Big Five trait scores. We use the descriptions as model inputs and the trait scores as reference labels.

Frising and Balcells prompted Llama 3.3 70B to answer the 50-item International Personality Item Pool (IPIP) Big Five Factor Markers questionnaire (Goldberg, 1992) from the perspective of each character, providing a five-point Likert response and a brief explanation for each item. The explanations form the textual descriptions, while the Likert responses were aggregated to obtain trait-level scores. Different from our notation, the Frising & Balcells dataset reports Emotional Stability instead of Neuroticism. To match our OCEAN representation, we rescale Emotional Stability to $[0,1]$ and convert it to Neuroticism as $N = 1 - \mathrm{rescaled}(ES)$.

3.1.2 Personality Retrieval Corpus

We augment the LLM with a personality retrieval corpus derived from the OpenPsychometrics SWCPQ dataset (Open Psychometrics, 2026). After preprocessing and excluding characters belonging to the 38 franchises in the development set, the corpus contains 1,602 character profiles. For each character, the retrieval document contains the character name, universe, and favored adjectives from the 12 bipolar items showing the largest absolute deviations from the neutral score. Before

evaluating each split, we remove the corresponding target characters from the retrieval corpus by name to prevent exact-character retrieval.

3.1.3 QUERY CONSTRUCTION AND RETRIEVAL

We consider two strategies for constructing the retrieval query: the original raw query and an IPIP-style rewrite produced by a Llama model. The raw-query strategy directly uses the original description for retrieval. For the IPIP-style strategy, the description is rewritten using a mechanism inspired by standard HyDE (Gao et al., 2023) into concise behavioral statements that more closely match the personality-related language in the retrieval corpus. The rewritten description is used only for retrieval, while the original character description is still used for OCEAN prediction. We evaluate dense, lexical, and full (dense and lexical combined) retrieval strategies (details are provided in Appendix).

3.1.4 OCEAN PREDICTION

We evaluate four reader configurations that vary by reader capability and the presence of retrieval: `Llama` and `Llama-RAG` use Llama-3.1-8B-Instruct without and with retrieval, respectively, while `Sonnet` and `Sonnet-RAG` use Claude Sonnet without and with retrieval. Llama is served locally using vLLM with 4-bit bitsandbytes quantization, while Claude Sonnet 4.5 is accessed through the Anthropic API. Both readers use greedy decoding ($T = 0$) with a maximum generation length of 256 tokens. All configurations use a fixed system prompt requesting continuous OCEAN scores in $[0, 1]^5$ in JSON format. The non-retrieval configurations receive only the target character description, while retrieval-augmented configurations additionally receive the retrieved character profiles as contextual information. The reader is instructed not to assume that characteristics of the retrieved characters apply directly to the target character.

3.2 OCEANDIALOGUE: PERSONALITY-CONTROLLED DIALOGUE GENERATION

3.2.1 PERSONALITY-CONDITIONED DIALOGUE CONSTRUCTION

To train OCEANDialogue, we construct a personality-conditioned dialogue dataset from SODA (Kim et al., 2023). Each scenario provides a narrative and preceding dialogue turns as a fixed conversational context, with the next response generated for ***PersonX***, one of the two participant variables in SODA. We pair each scenario with five distinct OCEAN profiles, producing alternative responses to the same conversational context under different personality conditions. We sample 40% of profiles by independently drawing all traits from $[0.1, 0.9]$, 20% from the near-neutral range $[0.4, 0.6]$, 30% as single-trait contrasts with one trait set to 0.1 or 0.9 and the others to 0.5, and 10% as analogous two-trait contrasts.

For each profile, Llama-3.1-8B-Instruct generates a one- to four-sentence response conditioned on the conversational context, OCEAN values, and a behavioral specification derived from the 30 IPIP-NEO-300 facets (Goldberg, 1999). Using 8,500 scenarios with five profiles per scenario yields 42,500 personality-conditioned training examples. We reserve a separate set of 300 scenarios for evaluation, reusing each fixed context across personality conditions so that generated responses can be compared as the requested OCEAN values vary.

3.2.2 TRAIT-SPECIFIC CONTINUOUS CONTROL

OCEANDialogue uses Llama-3.1-8B-Instruct as a frozen backbone. Personality control is implemented using five independently parameterized LoRA components, one for each OCEAN dimension. For every transformer decoder layer, separate low-rank components (A_t, B_t) are added to the query and value projections for $t \in \{O, C, E, A, N\}$. An OCEAN value $p_t \in [0, 1]$ is transformed into a signed coefficient $c_t = 2p_t - 1$, where $c_t \in [-1, 1]$. The magnitude and sign of c_t determine the contribution of the corresponding trait-specific component. A value of $p_t = 0.5$ produces zero contribution, values above 0.5 apply the learned component in the positive direction, and values below 0.5 apply it in the negative direction. The adapted linear transformation is $W\mathbf{h} + \sum_{t\in\{O,C,E,A,N\}}(2p_t - 1)\frac{\alpha}{r}B_tA_t\mathbf{h}$, where W is the frozen base-model weight matrix, A_t and B_t are the trait-specific low-rank matrices, r is the LoRA rank, and α is the LoRA scaling pa-

rameter. All five contributions are applied simultaneously, allowing a complete OCEAN profile to condition the same model generation.

During inference, the conversational context is supplied as text while the personality representation is supplied through the learned components. Thus, the OCEAN values themselves are not included in the dialogue prompt.

3.2.3 COUNTERFACTUAL TRAINING OBJECTIVE

We train the LoRA parameters on the 42,500 personality-conditioned dialogue examples while keeping the Llama backbone frozen. The generated response is used as the target with token-level cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$. For single-trait contrastive examples, we additionally construct a counterfactual profile by reflecting the manipulated trait to the opposite pole as $p_t^{\mathrm{cf}} = 1 - p_t$, while keeping the other traits unchanged. For each single-trait contrastive example, we evaluate the target response under both the original and counterfactual OCEAN profiles. We compute its mean token log-likelihood as $s_\theta(y \mid c, p) = \frac{1}{T}\sum_{k=1}^{T} \log P_\theta(y_k \mid y_{<k}, c, p)$, and define the difference between the original and counterfactual conditions as $\Delta = s_\theta(y \mid c, p) - s_\theta(y \mid c, p^{\mathrm{cf}})$. We optimize the counterfactual loss $\mathcal{L}_{\mathrm{CF}} = \mathrm{softplus}(-\Delta)$, encouraging the target response to have higher likelihood under its original personality profile than under the counterfactual profile. The final training objective combines the cross-entropy and counterfactual losses as $\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + 0.5\,\mathcal{L}_{\mathrm{CF}}$.

4 EXPERIMENTAL EVALUATION

We evaluate NaturalOCEAN and OCEANDialogue individually and additionally evaluate the complete two-stage pipeline end to end.

4.1 NATURALOCEAN EVALUATION

For exploratory evaluation, we conduct all experiments involving model choice, prompt design, query augmentation, and retrieval configuration on the development set introduced by Frising & Balcells, and select the final configuration based on development-set performance. For confirmatory evaluation, we construct a disjoint holdout set of 400 characters from the OpenPsychometrics SWCPQ dataset. Unlike the development set, whose reference scores are LLM-generated following Frising & Balcells, the holdout reference scores are derived from crowdsourced SWCPQ personality ratings.

To construct the holdout OCEAN scores, we use the 100 SWCPQ items corresponding to the IPIP personality markers, correct item polarity, rescale the ratings to $[0, 1]$, and average the items associated with each OCEAN trait. We generate character descriptions independently by prompting Claude Sonnet 4.5 with only the character name and franchise and asking it to produce an approximately 100-word third-person personality description, without access to the SWCPQ ratings or resulting OCEAN scores. To prevent direct retrieval of evaluated characters, we remove characters in each active evaluation split from the retrieval corpus by name. We evaluate each OCEAN trait using mean squared error (MSE) and the coefficient of determination (R^2).

We assess paired differences using 10,000 bootstrap resamples and apply Holm–Bonferroni correction across traits. Results with adjusted $p < 0.05$ are considered statistically significant.

4.1.1 RETRIEVAL AND QUERY-AUGMENTATION ABLATION

We evaluate a zero-shot baseline and eight retrieval and query-augmentation configurations using Llama-3.1-8B on the development set (Table 1). Dense retrieval, BM25 retrieval, and dense retrieval with IPIP-style query rewriting significantly reduce prediction error relative to zero-shot inference across all five traits.

Differences among retrieval variants are comparatively small. Relative to the full configuration, only BM25-only retrieval yields significantly higher error for Openness; removing individual retrieval components or using the raw query does not produce significant differences after Holm–Bonferroni correction. The full configuration achieves the lowest mean MSE and is used in subsequent experiments.

Table 1: Retrieval and query-augmentation ablations on the development set ($n = 406$) using Llama-3.1-8B. Values are uncalibrated trait-wise MSE. Dense + BM25 denotes RRF fusion.

Retrieval configuration	Query	O	C	E	A	N	Mean
Zero-shot (`Llama`)	–	0.0713	0.0429	0.0339	0.0544	0.0374	0.0480
Dense	Raw	0.0549	0.0325	0.0276	0.0421	0.0283	0.0371
BM25	Raw	0.0562	0.0353	0.0266	0.0449	0.0281	0.0382
Dense	IPIP rewrite	0.0525	0.0336	0.0276	0.0433	0.0277	0.0369
Dense + Reranker	IPIP rewrite	0.0529	0.0345	0.0279	0.0441	0.0278	0.0374
BM25 + Reranker	IPIP rewrite	0.0536	0.0359	0.0269	0.0442	0.0259	0.0373
Dense + BM25	IPIP rewrite	0.0516	0.0360	0.0259	0.0456	0.0283	0.0375
Dense + BM25 + Reranker	Raw	0.0534	0.0348	0.0284	0.0445	0.0270	0.0376
Dense + BM25 + Reranker	IPIP rewrite	0.0527	0.0335	0.0276	0.0431	0.0273	0.0368

4.1.2 Prediction Diagnostics, Calibration, and Confirmatory Evaluation

We first examine uncalibrated R^2 on the development set (Table 2). With the exception of Extraversion for `Sonnet-RAG`, R^2 is negative across all configurations and traits. Retrieval improves R^2 for all five traits with Llama and for four of five traits with Sonnet, with corresponding significant reductions in prediction error for all five Llama traits and four Sonnet traits.

Despite the poor raw R^2, predicted and reference scores have positive same-trait Pearson correlations across all traits, ranging from 0.33 to 0.69 (Figure 1). The predictions are also systematically more dispersed than the reference scores: the predicted-to-reference standard deviation ratio exceeds 1 for every configuration and trait, with a mean of 1.355. Together, these results suggest that differences in prediction scale contribute to the poor raw R^2.

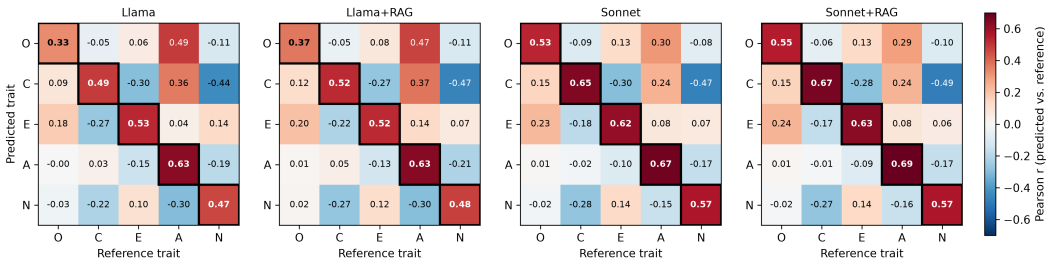


Figure 1: Predicted-versus-reference cross-trait Pearson correlations on the development set across all four configurations. Diagonal cells show same-trait correlations; off-diagonal cells show cross-trait associations.

Motivated by this scale mismatch, we apply trait- and configuration-specific affine calibration, $y_{\text{cal}} = ay_{\text{raw}} + b$, estimated by ordinary least squares. On the development set, five-fold cross-validation significantly improves R^2 for every trait for both `Llama-RAG` and `Sonnet-RAG` (all $p = 0.0010$), yielding positive R^2 across all traits.

We next evaluate the models on the disjoint confirmatory holdout (Table 2). For confirmatory calibration, we retain the development-set slope a and apply label-free batch centering, $y_{\text{cal}} = \bar{y}_{\text{batch}} + a(y_{\text{raw}} - \bar{y}_{\text{batch}})$, where $\bar{y}_{\text{batch}}$ is computed separately for each trait and configuration from the raw confirmatory predictions. No holdout reference scores are used for calibration.

Before calibration, retrieval significantly reduces prediction error for all five traits with Sonnet and for Openness, Agreeableness, and Neuroticism with Llama. Calibration significantly improves R^2 for Openness, Conscientiousness, Extraversion, and Neuroticism with `Llama-RAG`, and for Openness, Conscientiousness, and Neuroticism with `Sonnet-RAG`. After calibration, `Sonnet-RAG` retains a significant advantage over `Sonnet` for Openness, Conscientiousness, Extraversion, and Agreeableness; no significant differences remain between `Llama-RAG` and `Llama`.

Table 2: R^2 on the development and confirmatory holdout sets. Development calibration is five-fold cross-validated; holdout calibration uses the development-set calibration slopes with label-free batch centering. Cal. column indicates whether the model was calibrated or not.

		O		C		E		A		N	
Model	Cal.	Dev	Hold	Dev	Hold	Dev	Hold	Dev	Hold	Dev	Hold
`Llama`	No	−3.398	−0.501	−0.361	0.166	−0.267	0.182	−0.601	0.487	−0.530	−0.742
	Yes	0.098	0.178	0.218	0.362	0.271	0.395	0.391	0.595	0.201	0.193
`Llama-RAG`	No	−2.248	−0.165	−0.062	0.251	−0.030	0.240	−0.268	0.570	−0.115	−0.337
	Yes	0.126	0.156	0.256	0.366	0.264	0.398	0.396	0.571	0.212	0.180
`Sonnet`	No	−1.944	−0.214	−0.241	0.174	−0.086	0.574	−0.675	0.721	−0.532	−0.390
	Yes	0.273	0.247	0.412	0.471	0.375	0.632	0.450	0.724	0.306	0.324
`Sonnet-RAG`	No	−1.968	−0.151	−0.080	0.309	0.029	0.608	−0.493	0.758	−0.420	−0.257
	Yes	0.292	0.275	0.437	0.534	0.397	0.647	0.470	0.743	0.307	0.337

4.2 OCEANDIALOGUE EVALUATION

4.2.1 AUTOMATED EVALUATION

We evaluate continuous personality control across 300 conversational scenarios using Claude Sonnet 4.5 as the evaluator. For each OCEAN dimension, we generate responses at six trait levels, $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, using both zero-shot prompting and OCEANDialogue. We provide the conversational context, the intended trait, and the generated response in the prompt. The generated responses from the two strategies are evaluated by Sonnet to determine whether the intended change in the OCEAN trait is expressed in the response.

Table 3 reports mean perceived-trait scores across the six requested levels. OCEANDialogue shows a clear graded increase for Openness, Conscientiousness, and Extraversion. Agreeableness increases through the middle of the range but declines at the highest levels, while Neuroticism shows weaker variation across the requested values.

Table 3: Mean perceived-trait scores across 300 scenarios at six requested OCEAN levels. ZS denotes zero-shot prompting and OD denotes OCEANDialogue.

	0.0		0.2		0.4		0.6		0.8		1.0	
Trait	ZS	OD	ZS	OD	ZS	OD	ZS	OD	ZS	OD	ZS	OD
O	21.28	21.83	23.03	24.36	23.54	29.25	24.48	33.47	25.63	37.35	33.06	40.20
C	14.26	12.09	14.71	14.10	15.42	16.55	16.12	22.23	17.65	23.60	27.27	29.00
E	18.01	16.85	21.23	19.53	20.39	21.43	22.81	25.73	26.82	38.50	31.59	53.49
A	35.30	37.43	38.03	45.15	39.28	51.15	39.16	55.81	44.37	53.99	53.60	52.20
N	18.29	19.76	18.76	18.72	18.43	17.05	18.94	19.70	22.12	21.81	26.98	23.93

To quantify graded control, we compute a six-level slope for each scenario and trait. A positive slope indicates that perceived expression of the trait increases with the requested OCEAN value. We compare the 300 paired scenario-level slopes between OCEANDialogue and zero-shot prompting using paired t-tests, with Holm–Bonferroni correction across the five traits.

Table 4: Scenario-level continuous-control slopes. p-values are Holm-adjusted.

Trait	Zero-shot	OCEANDialogue	Difference	95% CI	Adj. p	d_z
O	9.659	19.291	+9.632	[7.217, 12.047]	< 0.001	0.453
C	10.654	16.963	+6.309	[4.125, 8.493]	< 0.001	0.328
E	12.442	34.913	+22.471	[19.234, 25.708]	< 0.001	0.789
A	15.770	15.000	−0.770	[−3.922, 2.383]	0.631	−0.028
N	7.718	4.681	−3.037	[−5.606, −0.470]	0.0412	−0.134

OCEANDialogue produces significantly steeper control slopes for Openness, Conscientiousness, and Extraversion, with the largest difference observed for Extraversion. Agreeableness shows no significant difference between methods, while Neuroticism has a significantly shallower slope under OCEANDialogue. Thus, OCEANDialogue strengthens graded control for three of the five OCEAN dimensions, with substantial variation across traits.

Spearman analysis. We assess whether perceived trait expression changes monotonically with the requested OCEAN value by computing Spearman's rank correlation between the six requested trait levels and their corresponding evaluator scores per scenario. We compare the resulting scenario-level correlations between OCEANDialogue and zero-shot prompting using paired permutation tests, with Holm–Bonferroni correction across the five traits. Scenarios with constant evaluator scores across all six levels, for which Spearman's ρ is undefined, are excluded pairwise.

OCEANDialogue significantly increases monotonic control for Openness ($\Delta\rho = 0.163$), Conscientiousness (0.205), and Extraversion (0.243; all adjusted $p < 0.001$). Agreeableness shows no significant difference ($\Delta\rho = -0.066$, $p = 0.073$), while Neuroticism is significantly less monotonic under OCEANDialogue ($\Delta\rho = -0.120$, $p = 0.010$). These results indicate that the stronger control observed for Openness, Conscientiousness, and Extraversion reflects not only larger changes in perceived trait expression, but also more consistent ordering across intermediate OCEAN values.

Cross-scenario consistency. We also examine whether positive six-level slopes persist when evidence is combined across scenarios. For groups of $N \in \{1, 2, 3, 5, 10, 20\}$ matched scenarios, we average the scenario-level slopes and define recovery as a positive mean slope. For each N, we sample 10,000 matched groups and compare methods using paired permutation tests with Holm–Bonferroni correction across traits. OCEANDialogue yields higher recovery than zero-shot prompting for Openness, Conscientiousness, and Extraversion across all group sizes, approaching 100% as N increases. Agreeableness shows no significant difference in recovery between OCEANDialogue and zero-shot prompting, while Neuroticism remains lower for OCEANDialogue.

4.2.2 Cross-Trait Evaluation

We next examine whether continuous personality control is specific to the manipulated OCEAN dimension or also changes the perceived expression of other traits. For each manipulated trait, we score the generated responses on all five OCEAN dimensions and compute the six-level control slope for each manipulated–perceived trait pair. This produces a 5×5 matrix for both zero-shot prompting and OCEANDialogue. Figure 2 reports the difference in slopes, $\Delta\beta = \beta_{\text{OCEANDialogue}} - \beta_{\text{zero-shot}}$. Statistical significance is assessed across all 25 comparisons using Holm–Bonferroni correction.

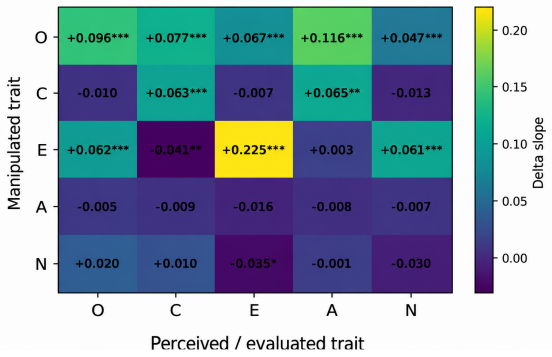


Figure 2: Cross-trait effects of OCEANDialogue relative to numerical zero-shot prompting.

The diagonal results support the continuous-control analysis. OCEANDialogue significantly strengthens control of Openness, Conscientiousness, and Extraversion. Agreeableness remains non-significant, while the Neuroticism effect does not remain significant after correction across all 25 comparisons. The off-diagonal results show that the learned controls are not fully trait-specific. Manipulating Openness produces significant positive changes in all four other perceived traits. Extraversion also produces several cross-trait effects, increasing perceived Openness and Neuroticism

while decreasing perceived Conscientiousness. Manipulating Conscientiousness significantly increases perceived Agreeableness. Agreeableness produces no significant off-diagonal effects, while Neuroticism produces one significant decrease in perceived Extraversion.

4.2.3 HUMAN EVALUATION

We conduct a human evaluation of the high–low personality contrasts. For each trial, the target trait is set to 0 or 1, while the remaining four dimensions are fixed at 0.5. In each trial, participants are shown a conversational context together with responses generated under the corresponding high- and low-trait conditions. The responses are presented in randomized A/B order, and participants select which response better expresses the target personality characteristic without seeing the underlying OCEAN values or condition labels. The evaluation covers all five OCEAN dimensions and both generation strategies in equal proportion.

We recruit 24 participants through convenience sampling (aged 22–45; 9 women, 15 men); four fail quality-control trials, leaving 20 participants and 400 experimental judgments. An a priori simulation-based analysis indicates at least 0.90 power under the considered heterogeneity assumptions.

Across 400 experimental judgments, participants select the intended high-trait response 228 times, the low-trait response 118 times, and report a tie 54 times. Among the 346 non-tie judgments, the intended personality direction is preferred in 65.9% of cases (95% CI: [60.4%, 71.3%]). When separated by generation strategy, intended-direction preference is 62.6% for zero-shot prompting and 69.1% for OCEANDialogue. However, a paired comparison of participant-level preference rates does not yield a statistically significant difference between the two generation strategies ($t(19) = 0.94$, $p = 0.357$).

4.3 FULL PIPELINE EVALUATION

We evaluate the complete NaturalOCEAN–OCEANDialogue pipeline to test whether personality characteristics specified in natural-language character descriptions remain recognizable in the generated dialogue. For this, we sample 100 characters from the confirmatory holdout set. For this evaluation, NaturalOCEAN uses the best-performing calibrated Sonnet-RAG configuration with IPIP-style HyDE query rewriting and the full retrieval pipeline. For each character, NaturalOCEAN infers an OCEAN representation from the original description, which OCEANDialogue then uses to generate responses in five different conversational scenarios. The automatic judge (Claude Sonnet 4.5) receives the original character description, the five conversational contexts, and the corresponding generated responses, and rates how well the described personality is expressed across the responses on a 0–4 scale. 0 indicates that the described characteristics are not evident; 1 indicates weak or isolated evidence; 2 indicates that some characteristics are consistently evident but substantial aspects are missing; 3 indicates that the main personality characteristics are clearly and consistently expressed; and 4 indicates strong and consistent expression of multiple distinctive characteristics.

For 51% of characters, the main personality characteristics are clearly and consistently expressed (score 3), while the remaining 49% show some consistent characteristics but omit substantial aspects (score 2). No character receives a score of 0 or 1, indicating absent or only weak personality evidence, and none receives the highest score of 4.

5 DISCUSSION

The strong effect of calibration suggests that NaturalOCEAN captures useful relative trait information even when its raw numerical scale differs from the reference distribution. This distinction is important for applications where the representation is used comparatively or as an intermediate control signal rather than interpreted as an absolute psychometric measurement.

OCEANDialogue's automatic evaluation results show that an interpretable psychometric representation can be translated into graded changes in generated language, but not equally across all five dimensions. For Openness, Conscientiousness, and Extraversion, increasing the requested OCEAN value produces substantially stronger increases in perceived trait expression than numerical zero-shot prompting. At the same time, the lack of improvement for Agreeableness and the weaker

relationship for Neuroticism show that assigning each dimension its own learned parameter-space direction does not by itself guarantee equally effective control across traits. The Spearman analysis further shows that these gains reflect more consistent ordering across intermediate OCEAN values, rather than only larger changes in perceived trait expression.

The cross-scenario analysis further shows that the learned control for Openness, Conscientiousness, and Extraversion is not confined to a small subset of conversational contexts. When scenario-level continuous-control slopes are averaged across multiple contexts, positive slopes for these traits become increasingly consistent, approaching complete recovery at larger aggregation sizes. This pattern is consistent with the statistical intuition of the aggregation principle (Epstein, 1979), which states that context-specific variation can obscure a systematic effect at the level of individual observations, while averaging across observations reveals whether that effect persists across contexts.

The weaker and less consistent results for Agreeableness and Neuroticism may reflect differences in how readily these traits become observable in language. Prior work on personality perception has found lower interjudge agreement for Agreeableness and particularly weak accuracy for Neuroticism in brief observations, while computational studies of written language report among the smallest trait–language associations for these two dimensions (John & Robins, 1993; Tskhay & Rule, 2014; Moreno et al., 2021). For Neuroticism, prior work has found that trait judgments are relatively inaccurate at zero acquaintance, but become more accurate when behavior is observed in trait-relevant, socially stressful situations (Hirschmüller et al., 2015). More generally, the behavioral expression of Big Five traits depends on situational characteristics, and the relevant characteristics differ across traits (Fleeson, 2007). Our results may therefore reflect variation in how strongly different conversational scenarios afford observable expressions of each trait.

Trait interaction results indicate that OCEANDialogue's learned trait directions are disentangled in their linguistic expression only partially. Dimensions like Openness and Extraversion alter cues that the evaluator associates with multiple OCEAN traits, whereas Agreeableness exhibits substantially more localized effects. The observed interactions may reflect shared linguistic characteristics of different personality dimensions, but our evaluation does not establish whether they correspond to underlying psychological relationships among the traits.

Human judgments also show that the intended personality differences are perceptible, although the 65.9% preference rate indicates that these cues are not uniformly salient at the single-response level.

The pipeline evaluation results provide evidence that personality information can propagate from a natural-language description, through the inferred OCEAN representation, to generated dialogue. However, the absence of highest score cases suggests that some finer-grained or character-specific personality information is not preserved through the complete pipeline.

Several limitations constrain the conclusions of this study. First, the full-pipeline evaluation is performed by an automated judge and therefore does not establish human perception. The use of the same model family for description generation, OCEAN inference, and automatic judging may also introduce model-specific evaluation bias. Second, the NaturalOCEAN evaluations rely on heterogeneous reference construction for the development and confirmatory scores. Although this diversity provides a stronger test of generalization across evaluation settings, it also limits direct comparison of absolute performance between the datasets. Finally, personality control is not equally effective across traits. As future work, we will investigate dynamic, per-trait adjustments to the LoRA scaling parameters to stabilize control for traits like Neuroticism and Agreeableness.

6 Conclusion

We introduce a two-stage framework mapping free-form character descriptions to continuous OCEAN representations for personality-conditioned dialogue. NaturalOCEAN shows that retrieval and calibration improve alignment with reference personality scores, while OCEANDialogue provides stronger graded control for Openness, Conscientiousness, and Extraversion. Cross-trait analysis reveals partial coupling between dimensions, and human and end-to-end evaluations show that intended personality differences and salient character traits remain recognizable in generated dialogue. These results support continuous OCEAN representations as an interpretable interface for controllable dialogue, while showing that control quality and trait specificity vary across dimensions.

AI USE STATEMENT

Generative AI tools were used throughout this work both as components of the research methodology and as research assistants. As described in the paper, language models were used for OCEAN inference, retrieval-assisted prediction, generation of personality-conditioned dialogue training data, generation of character descriptions for the confirmatory evaluation, and automatic evaluation of generated dialogue.

Generative AI tools were also used during the research process to support brainstorming, methodological development, implementation and debugging, statistical analysis, interpretation of results, literature search and summarization, and drafting and editing of the manuscript. The authors reviewed and verified AI-assisted analyses, code, citations, methodological decisions, and manuscript text, and take responsibility for the final content of the work.

ETHICS STATEMENT

NaturalOCEAN estimates OCEAN traits from textual descriptions and is evaluated primarily on fictional characters, while OCEANDialogue uses OCEAN representations to control personality expression in generated dialogue. Applying these methods to real individuals could raise concerns related to privacy, profiling, stereotyping, and the interpretation or manipulation of inferred psychological characteristics. OCEAN predictions should not be treated as clinical assessments or used as the sole basis for consequential decisions about individuals, and personality-conditioned generation should not be used to impersonate, manipulate, or misrepresent real people. The methods in this work are intended for character modeling, controllable dialogue generation, and personality-conditioned conversational agents.

The human evaluation involves adult participants who judge personality differences between generated dialogue responses; it does not involve inferring or manipulating the participants' own personalities. Participant judgments are reported only in aggregate, and demographic information is used only to describe the study sample.

REPRODUCIBILITY STATEMENT

We support the reproducibility of our work by providing information about the data construction and evaluation process, model and retrieval configurations, calibration procedure, OCEANDialogue architecture and training objective, and statistical evaluation methodology in the paper. Additional architecture diagrams, implementation details, and training configurations are provided in the appendix.

REFERENCES

Rumi Allbert, James K Wiles, and Vlad Grankovsky. Identifying and manipulating personality traits in llms through activation engineering. *arXiv preprint arXiv:2412.10427*, 2024.

Anna Baumert, Manfred Schmitt, Marco Perugini, Wendy Johnson, Gabriela Blum, Peter Borkenau, Giulio Costantini, Jaap JA Denissen, William Fleeson, Ben Grafton, et al. Integrating personality structure, personality process, and personality development. *European Journal of Personality*, 31 (5):503–528, 2017.

Pranav Bhandari, Nicolas Fay, Sanjeevan Selvaganapathy, Amitava Datta, Usman Naseem, and Mehwish Nasim. Activation-space personality steering: Hybrid layer selection for stable trait control in llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6388–6403, 2026.

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

Erik Derner, Dalibor Kučera, Nuria Oliver, and Jan Zahálka. Can chatgpt read who you are? *Computers in human behavior: Artificial humans*, 2(2):100088, 2024.

Funda Durupinar, Mubbasir Kapadia, Susan Deutsch, Michael Neff, and Norman I Badler. Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis. *ACM Transactions on Graphics (TOG)*, 36(1):1–16, 2016.

Seymour Epstein. The stability of behavior: I. on predicting most of the people much of the time. *Journal of personality and social psychology*, 37(7):1097, 1979.

Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad, and Daniel Oberski. On text-based personality computing: Challenges and future directions. In *Findings of the association for computational linguistics: ACL 2023*, pp. 10861–10879, 2023.

William Fleeson. Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of personality*, 75(4):825–862, 2007.

Michel Frising and Daniel Balcells. Linear personality probing and steering in llms: A big five study. *arXiv preprint arXiv:2512.17639*, 2025.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777, 2023.

Lewis R Goldberg. An alternative description of personality: The big-five factor structure. *Journal of personality and social psychology*, 59(6):1216–1229, 1990.

Lewis R. Goldberg. The development of markers for the big-five factor structure. *Psychological Assessment*, 4(1):26–42, 1992.

Lewis R Goldberg. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7:7–28, 1999.

Sarah Hirschmüller, Boris Egloff, Stefan C Schmukle, Steffen Nestler, and Mitja D Back. Accurate judgments of neuroticism at zero acquaintance: A question of relevance. *Journal of personality*, 83(2):221–228, 2015.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18234–18242, 2024.

Muhua Huang, Xijuan Zhang, Christopher Soto, and James Evans. Designing AI-agents with personalities: A psychometric approach. *Personality Science*, 7:1–26, 2026. doi: 10.1177/27000710251406471.

Oliver P John and Richard W Robins. Determinants of interjudge agreement on personality traits: The big five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of personality*, 61(4):521–551, 1993.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. SODA: Million-scale dialogue distillation with social commonsense contextualization. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12930–12949, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.799. URL https://aclanthology.org/2023.emnlp-main.799/.

Srikar Kodavati, Wilhen Alberto Hui Mei, Anish Kanade, and Funda Durupinar. A personality-labeled semantic dataset from facial expressions, gaze, and head movement cues. *Computers & Graphics*, 133:104455, 2025. doi: 10.1016/j.cag.2025.104455.

Wenkai Li, Jiarui Liu, Andy T Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. Big5-chat: Shaping llm personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20434–20471, 2025.

Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE international conference on data mining (ICDM)*, pp. 1184–1189. IEEE, 2020.

José David Moreno, Jose A Martinez-Huertas, Ricardo Olmos, Guillermo Jorge-Botana, and Juan Botella. Can personality traits be measured analyzing written language? a meta-analytic study on computational methods. *Personality and individual differences*, 177:110818, 2021.

Open Psychometrics. Swcpq-features-aggregated-dataset. Data from the Statistical "Which Character" Personality Quiz, Jan 2026. URL `https://openpsychometrics.org/tests/characters/data/`. January 2026 release; accessed August 27, 2026; CC BY-NC-SA 4.0.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934—-952, 2015.

Sinan Sonlu, Bennie Bendiksen, Funda Durupinar, and Uĝur Güdükbay. Effects of embodiment and personality in LLM-based conversational agents. In *2025 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, IEEEVR '25, pp. 718–728, Saint-Malo, France, March 2025. doi: 10.1109/VR59515.2025.00094.

Konstantin O Tskhay and Nicholas O Rule. Perceptions of personality in text-based media and osn: A meta-analysis. *Journal of Research in Personality*, 49:25–30, 2014.

Huy Vu, Huy Anh Nguyen, Adithya V Ganesan, Swanie Juhng, Oscar NE Kjell, Joao Sedoc, Margaret L Kern, Ryan L Boyd, Lyle Ungar, H Andrew Schwartz, et al. Psychadapter: Adapting llm transformers to reflect traits, personality and mental health. *arXiv preprint arXiv:2412.16882*, 2024.

Yilei Wang, Jiabao Zhao, Deniz S Ones, Liang He, and Xin Xu. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1):519, 2025.

Haein Yeo, Taehyeong Noh, Seungwan Jin, and Kyungsik Han. PADO: Personality-induced multiagents for detecting OCEAN in human-generated texts. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pp. 5719–5736, Abu Dhabi, UAE, 2025. Association for Computational Linguistics.

Yangfu Zhu, Yue Xia, Meiling Li, Tingting Zhang, and Bin Wu. Data augmented graph neural networks for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 664–672, 2024.

A Appendix

A.1 Architecture Details

The two-stage pipeline implements NaturalOCEAN, which maps a natural character description to a continuous OCEAN profile, and OCEANDialogue, which uses this profile for continuous personality control during dialogue generation.

NaturalOCEAN. NaturalOCEAN rewrites the character description into an IPIP-style query using a HyDE-style Llama rewrite. We use dense and BM25 retrieval based on a personality corpus. The retrieval results from both methods are combined using reciprocal-rank fusion and reranked using a cross-encoder. The retrieved details and the original character description are then provided to the reader to predict the five continuous OCEAN values in $[0, 1]$. Figure A1 shows the complete NaturalOCEAN architecture.

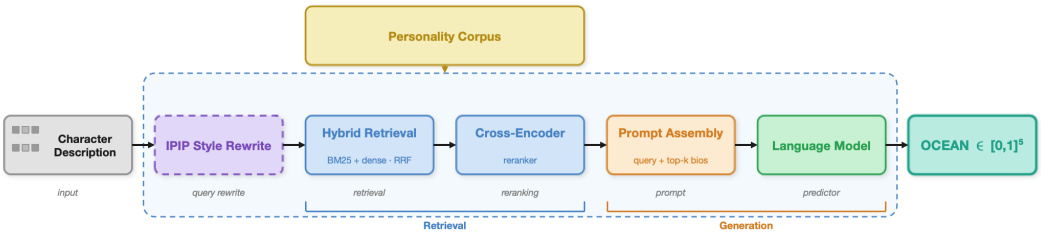


Figure A1: Overview of the NaturalOCEAN architecture. A character description is rewritten into an IPIP-style query and used for hybrid retrieval from the personality corpus. Retrieved profiles are reranked and provided to the language model to predict a continuous OCEAN representation in $[0, 1]^5$.

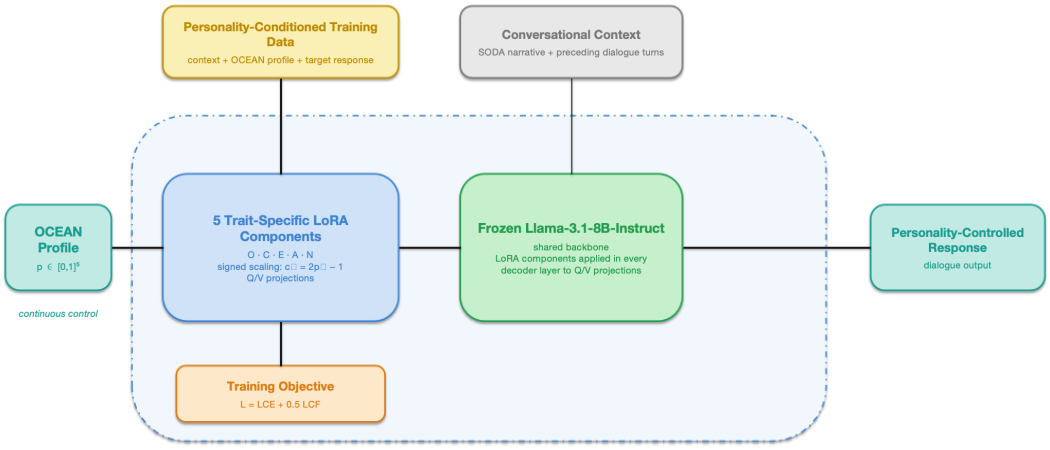


Figure A2: Overview of the OCEANDialogue architecture. The five trait-specific LoRA components correspond to O, C, E, A, and N. Each component is controlled using the signed coefficient $c_t = 2p_t - 1$ and applied to the query and value projections of the frozen Llama-3.1-8B-Instruct backbone. The conversational context is provided as text, and the five trait components are applied simultaneously to generate the personality-controlled response.

OCEANDialogue. OCEANDialogue employs five trait-specific LoRA components over a frozen Llama-3.1-8B-Instruct backbone. Each OCEAN value p_t is mapped to a signed coefficient $c_t =$

$2p_t - 1$, which maps the range from $[0, 1]$ to $[-1, 1]$. The coefficient scales the corresponding trait-specific LoRA component, and these components are applied simultaneously to the projections in each transformer layer. The conversational context is provided as a text prompt, and the model generates the response. Figure A2 shows the OCEANDialogue architecture.

A.2 NaturalOCEAN Implementation Details

NaturalOCEAN implements mean-pooled, L2-normalized `all-mpnet-base-v2` [1] representations for dense retrieval, with a maximum input length of 128 tokens. BM25 retrieval uses $k_1 = 1.5$ and $b = 0.75$. The HyDE-style IPIP rewrite is generated using Llama-3.1-8B-Instruct. Dense and BM25 results are combined using reciprocal-rank fusion with $k = 60$.

The top 50 combined results are reranked using `ms-marco-MiniLM-L-6-v2` [2], with a maximum query-document length of 256 tokens. The top eight reranked profiles and the original character description are provided to the reader to predict the continuous OCEAN values.

Calibration is performed using a separate affine mapping for each OCEAN trait with five-fold cross-validation on the development set. For holdout evaluation, we use the development-set estimated slope and recenter the predictions using the unlabeled holdout prediction mean, without using holdout reference labels.

A.3 OCEANDialogue Training Details

OCEANDialogue uses a frozen Llama-3.1-8B-Instruct backbone with trait-specific LoRA components applied to the projection layers. The main model uses LoRA rank $r = 8$ and $\alpha = 8$, with 17,039,360 trainable parameters across 64 projection layers.

Training is performed on 42,500 responses obtained from 8,500 training scenarios. We train using the AdamW optimizer with a learning rate of 10^{-4}, weight decay of 0.01, micro-batch size of 1, and 32 gradient-accumulation steps, giving an effective batch size of 32. We use a maximum sequence length of 400 tokens, gradient clipping at 1.0, and seed 42.

The training objective combines the response-generation loss with the counterfactual objective using $\lambda_{\mathrm{CF}} = 0.5$. Validation is performed every 200 optimization steps, and the checkpoint with the lowest combined validation loss is retained.

A.4 LoRA Hyperparameter Ablation

We evaluate how changes in LoRA rank r and scaling parameter α affect OCEANDialogue. We compare five configurations: $(r, \alpha) = (4, 4), (8, 4), (8, 8), (8, 16)$, and $(16, 16)$, using the same training setup. For this post-hoc ablation, we use a subset of 50 scenarios from the 300-scenario evaluation set. The main $(r, \alpha) = (8, 8)$ configuration was fixed before this analysis and was not selected based on these results.

Table 5 shows that although the $(8, 16)$ and $(16, 16)$ configurations converge faster initially, all configurations eventually converge to similar validation losses. The $(8, 16)$ and $(16, 16)$ configurations obtain the lowest combined loss of 1.2757, compared with 1.2770 for the final $(8, 8)$ configuration. Hence, increasing the rank and scaling results in only small differences in the final validation loss.

Table 5: LoRA training ablation. Combined validation loss is reported at selected optimization steps.

r	α	Params	Step 400	Step 1200	Step 2000	Step 2600
4	4	8.52M	1.3678	1.3043	1.2897	1.2806
8	4	17.04M	1.3671	1.3055	1.2911	1.2808
8	**8**	**17.04M**	1.3471	1.2945	1.2840	1.2770
8	16	17.04M	1.3323	1.2804	1.2787	**1.2757**
16	16	34.08M	**1.3273**	**1.2793**	1.2792	**1.2757**

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[2] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

We further evaluate in Table 6 the mean scenario-level Spearman correlation across the six control levels and evaluator scores. No single configuration performs best across all traits, while the final $(8, 8)$ configuration obtains the highest macro-average correlation ($\rho = 0.422$). The configurations with higher rank and scaling and the lowest validation loss do not show the strongest overall control, suggesting that small differences in validation loss do not directly translate to continuous-control performance.

Table 6: LoRA continuous-control ablation on the same 50 scenarios. Values are mean scenario-level Spearman correlations between the six requested control levels and evaluator scores. Macro ρ is the unweighted mean across the five traits.

r	α	O	C	E	A	N	Macro ρ
4	4	0.470	0.415	**0.721**	0.309	0.070	0.397
8	4	0.351	0.525	0.634	0.394	**0.105**	0.402
8	**8**	**0.494**	0.574	0.686	0.299	0.056	**0.422**
8	16	0.399	**0.594**	0.595	0.297	0.060	0.389
16	16	0.463	0.521	0.542	**0.403**	0.022	0.390