# Multi-Dimensional Style Control for Deployed Chatbots via Parameterized Style Labels

Anonymous ACL submission

## Abstract

Deployed companion dialogue systems benefit from adapting conversational style each turn in response to personality traits, emotional arousal, relational stage, and conversation momentum. We present an architecture that converts these signals into adverb-keyed style labels (e.g., `FORMALITY=extremely_high`), concatenated as key-value pairs in the LLM system prompt. We screen 20 candidate style dimensions and select five (FORMALITY, POLITENESS, CERTAINTY, FRIENDLINESS, EMOTIONAL_TONE). In a forced-choice perception experiment on outputs generated under full deployment conditions—all five dimensions active, multi-turn context, production prompt—annotators identify the higher-tier response with 67.8% accuracy across 6,600 trials ($p < .001$), with accuracy scaling with tier gap on four of five dimensions. Controllability covaries with NLP operationalization: dimensions with dedicated benchmarks and classifiers show lower adjacent-tier collapse. Some labels also alter behavioral content beyond the target linguistic form; a behavior-controlled replication shows that anchoring the intended action can recover sharper style contrasts.

## 1 Introduction

Companion chatbots, role-playing agents, and social skill trainers maintain persistent characters whose conversational style should shift from turn to turn. A character responding to a close friend under high emotional arousal speaks differently from the same character addressing a new acquaintance in a calm exchange, not just in what it says but in register, hedging, warmth, and intensity. The signals driving these shifts differ in timescale and type: personality traits (stable), emotional arousal (per-turn), relational closeness (per-session), conversation momentum (per-turn). A style control mechanism for deployed systems needs to handle all of these, at per-turn granularity, across multiple style dimensions concurrently. Existing methods handle subsets of these requirements but not their combination: per-turn adaptation, multi-dimensional composition, and deployment-compatible integration. Our architecture addresses this by converting heterogeneous upstream signals into composable style labels at per-turn granularity, without model modification.

We use a three-stage architecture (Figure 1): a sparse mapping layer computes continuous style values from upstream signals, a quantization step discretizes each value into one of five adverb-keyed tiers, and the resulting labels are injected into the LLM system prompt each turn. The mapping layer computes each style dimension as a sparse weighted combination of 3–5 upstream signals, informed by psycholinguistic considerations (§3). The raw personality, emotion, relationship, and context scores are not directly exposed in the prompt; the LLM receives the derived style parameters as control variables. The mapping decouples state modeling from generation. The architecture is deployed in a companion chatbot that served 77 participants in a controlled study (reported separately). In production, the style control component operates alongside inner monologue, planning, and context memory modules; this paper isolates that component.

The style parameters produced by the mapping layer must be communicated to the LLM each turn. We use prompt injection: labels are concatenated as key-value pairs in the system prompt. Re-deriving these labels each turn avoids the persona drift observed in static prompt descriptions (Li et al., 2024). Fine-tuning encodes a fixed style in model weights and cannot adapt per turn; activation steering can adjust per turn but faces composition difficulties when multiple dimensions are active (van der Weij et al., 2024; Hoppe et al., 2026). Prompt-injected labels compose naturally through

the model's forward pass.

The choice of style dimensions requires empirical validation: interpretability varies across models, human readers may not perceive the output differences, and the adverb-keyed scheme has a finite resolution. We evaluate in two stages. First, we screen 20 candidates under simplified single-label prompts to identify which dimensions produce interpretable gradients. Then, we validate the five selected dimensions under the full production configuration to test whether the style contrasts remain perceptible.

Under these conditions, all five dimensions remain reliably perceptible (67.8% overall; up to 93% at maximum tier separation), with accuracy scaling with tier gap. We also observe that some labels shift behavioral content alongside linguistic form, a phenomenon we call *style-to-behavior leakage*; a behavior-controlled replication isolates the effect.

Our contributions are: (1) a state-to-style mapping architecture that addresses the combination of per-turn adaptation, multi-dimensional composition, and deployment-compatible integration by mapping heterogeneous state signals into adverb-keyed style labels; (2) a screening protocol that identifies usable dimensions from a candidate pool, with adjacent-tier collapse as a finer-grained quality measure; (3) human perception evidence that all five dimensions remain perceptible when active simultaneously in production.

## 2 Related Work

**Text style transfer and controllable generation.** Text style transfer modifies stylistic attributes while preserving core content. Formality transfer has dedicated benchmarks such as GYAFC (Rao and Tetreault, 2018), and controllable generation surveys cover fine-tuning, prompting, decoding-time intervention, and latent-space methods (Liang et al., 2024). Plug-and-play methods such as PPLM steer generation with external attribute models without retraining the base LM (Dathathri et al., 2020); decoding-time methods use future discriminators (Yang and Klein, 2021). Much of this work targets single attributes or sentence-level rewriting, and the control mechanisms typically rely on auxiliary models or model access. Our setting differs: we generate dialogue responses with multiple style labels active simultaneously, rather than rewriting along a single attribute.

**Prompt-based style and personality control.** Prompt-based control avoids modifying model weights. LLMs can perform zero-shot style transfer from natural-language instructions (Reif et al., 2022), and prompt-and-rerank approaches select among candidates for arbitrary style targets (Suzgun et al., 2022). For personality control specifically, Big5-Scaler injects numerical trait values into prompts (Cho and Han, 2025), BIG5-CHAT trains on Big Five dialogue data (Li et al., 2025), and activation-space methods extract personality vectors for steering (Feng et al., 2026). Pielage et al. (2026) proposes dynamic prompt modulation via state machines but focuses on architecture rather than signal design. These approaches expose personality or style signals directly to the LLM. Our system instead keeps personality and relational state in an upstream computation layer and exposes only derived key-value style labels, without anchor descriptions or examples. Whether such compact signals can support diverse style dimensions when multiple labels are active under deployment conditions is the question our experiments address.

**Multi-dimensional composition and activation steering.** Most style-control work evaluates one dimension at a time or composes heterogeneous properties, such as safety, formality, and output language, that operate at different representational levels (Scalena et al., 2024). Homogeneous multi-dimensional composition, where several conversational style dimensions of the same representational type must be active and individually discriminable at the same time, is a distinct and less studied problem. In activation steering, multi-property composition via vector addition faces interference that may require orthogonalization (Hoppe et al., 2026; van der Weij et al., 2024). Recent work addresses this with non-linear methods: K-Steering trains a single multi-label classifier on hidden activations and computes gradient-based intervention directions, avoiding linearity assumptions (Oozeer et al., 2025). These steering methods have been evaluated on attributes such as sentiment and toxicity in single-turn settings, not on conversational style dimensions in deployed multi-turn dialogue. At the prompt level, whether a concatenated set of homogeneous style labels remains interpretable when multiple dimensions are active and conversational contexts vary is largely untested.

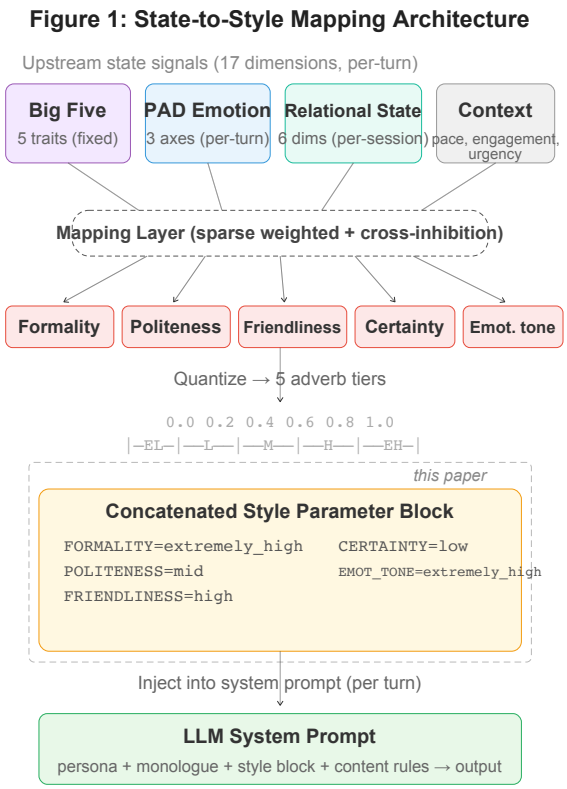


Figure 1: State-to-style mapping architecture. Upstream state signals are transformed by a sparse mapping layer into five continuous style values, quantized to adverb-keyed tiers, assembled into a style parameter block, and injected into the LLM system prompt each turn. The dashed box indicates the component tested in this paper.

**Evaluation of style controllability.** Style control is often evaluated with automatic classifiers, LLM-as-judge, or psychometric questionnaires. Classifiers face domain and granularity mismatch for short social dialogue; LLM judges introduce their own style priors; questionnaires measure indirect constructs rather than perceptible output differences. Human evaluation protocols in style transfer are often underspecified and difficult to compare across studies (Briakou et al., 2021).

## 3 Architecture

The deployed companion chatbot tracks several state variables, differing in timescale, that drive per-turn style adaptation. A three-stage pipeline translates these signals into LLM-interpretable style parameters.

**Mapping.** The pipeline takes four signal categories as input: Big Five personality trait scores (5 dimensions, fixed per character), PAD emotional state (Pleasure, Arousal, Dominance; updated per turn), six-dimensional relational state (closeness, trust, liking, respect, attractiveness, power; evolving across sessions), and contextual variables (dialogue pace, topic engagement, response urgency, etc.; computed per turn). Each of the five style dimensions is computed as a sparse weighted combination of these signals, with each dimension driven by 3–5 primary inputs informed by psycholinguistic and pragmatic considerations (Pennebaker and King, 1999). For example, FORMALITY is primarily driven by interpersonal closeness (negative: closer relationships → less formal) and power asymmetry (positive: greater power distance → more formal); POLITENESS is driven by respect and power distance (positive: greater asymmetry → more polite). For CERTAINTY, the primary inputs are neuroticism (negative) and dominance (positive). FRIENDLINESS draws mainly on liking and closeness, with stronger relational bonds producing more intimate output. EMOTIONAL_TONE is the only dimension primarily driven by PAD arousal and valence.

Because several upstream signals feed into more than one dimension (e.g., dominance drives both CERTAINTY and FORMALITY), the raw mapping can produce correlated outputs. A cross-inhibition step reduces the largest residual correlation, between FRIENDLINESS and CERTAINTY. After this step, sampling $N = 500$ random state configurations shows moderate independence: the largest remaining pairwise correlation is $r = -0.44$ (FORMALITY–FRIENDLINESS); all other pairs satisfy $|r| < 0.25$.

**Quantization.** Each dimension's continuous output is discretized into five adverb-keyed tiers: values below 0.2 map to `extremely_low`, 0.2–0.4 to `low`, 0.4–0.6 to `mid`, 0.6–0.8 to `high`, and above 0.8 to `extremely_high`. Quantization absorbs sub-threshold numerical differences that LLMs cannot reliably distinguish (our granularity ceiling analysis in §4.3 indicates five tiers as the effective resolution limit for the tested dimensions). It also converts continuous values into natural language tokens interpretable via the model's pretrained knowledge of degree adverbs. The tier names use scalar terms (*low*, *high*, *extremely*) whose intensity ordering is well-established in linguistic theory and corpus-based studies (Bolinger, 1972; Kennedy and McNally, 2005; de Melo and Bansal, 2013).

**Injection.** The labels for all five dimensions

are assembled into a style parameter block and injected into the LLM system prompt each turn; the model interprets the full label set through a single forward pass. No anchor descriptions, examples, or definitions accompany the labels. Providing examples risks narrowing outputs to the surface patterns of those examples, underutilizing the stylistic breadth encoded in pretraining; bare adverb-keyed labels leave the model free to draw on that broader distribution. Prompt templates are provided in Appendix E.

To illustrate: a character with high neuroticism (personality, stable) experiencing elevated arousal (emotion, per-turn) in a close relationship (relationship, per-session) maps to `CERTAINTY=low` (neuroticism reduces epistemic confidence; Pennebaker and King 1999), `EMOTIONAL_TONE=high` (arousal drives expressive intensity), and `FORMALITY=low` (closeness lowers register; Brown and Levinson 1987). The same character under neutral arousal produces `EMOTIONAL_TONE=mid`. Style thus depends on the full state, not any single input.

The style parameter block is one component of the LLM system prompt, alongside persona, inner monologue, content constraints, and response rules. In the production system, a role-level instruction directs these parameters toward linguistic form rather than behavioral content. The five dimensions were selected from a pool of 20 candidates through the screening experiments described in §4; additional candidates from the reserve pool can be screened for other deployment needs. The mapping rules are hand-designed for this application.

## 4 Screening Experiments

### 4.1 Design

The screening experiments use five model-language conditions: Qwen3-Next-80B-A3B-Instruct (the production deployment model) for Chinese response generation and GPT-4o for English response generation serve as the two primary conditions, with three raters providing independent assessments. Three supplementary conditions (GPT-4o ZH, GPT-4o-mini ZH, GPT-4o-mini EN) were screened using the same protocol and rater count. All conditions use English labels. Generation settings are in Appendix E.

The screening experiments omit the role-level form-over-content instruction from the production pipeline (§3), testing whether labels alone produce interpretable style gradients. The perception experiment (§5) restores this instruction alongside the full production prompt.

All conditions use a social persona rather than an assistant identity, because the default assistant persona established through alignment training exhibits consistently high politeness, agreeableness, and helpfulness, constraining the range of style variation testable in the experiment.

The 20 candidate dimensions were identified from the style variation needed in the deployed system: we enumerated the surface-level stylistic properties that the chatbot should be able to express, then expanded each into candidate label names that the model might respond to effectively.[1]

For each dimension, we placed a single label in the system prompt and swept it through five tiers. We used ten user inputs spanning different emotional and pragmatic contexts, covering two complementary scenario types: (1) sharing/emotional (workplace criticism, job offer celebration, partner argument, travel excitement, insomnia complaint) and (2) question/invitation (opinion on long-distance relationships, hiking invitation, pet adoption advice, hypothetical spending, self-perception). One generation per tier per input (50 per dimension, 1,000 total).

Raters then compared the extreme-low (EL, `extremely_low`) and extreme-high (EH, `extremely_high`) outputs across all 10 scenarios to determine whether EH consistently produced more of the target attribute than EL, in the direction implied by the label's ordinary English lexical semantics. No custom definitions were provided; the criterion is alignment with the standard meaning of the dimension name. Three independent raters assigned ratings on a three-level scale: **clear** = EL and EH show a clear, unambiguous contrast across all or nearly all scenarios; **detectable** = contrast is present and directionally correct but one or both endpoints lack extreme character in 1–2 scenarios; **weak** = contrast is minimal, inconsistent, or the EL endpoint has lost its low-end character across multiple scenarios.

Disagreements among the three raters were resolved by majority vote. Adjacent-tier collapse (two consecutive tiers producing output indistinguishable in the target direction) was also recorded

[1] Dimensions such as HUMOR, SIMPLICITY, and SUBJECTIVITY, though well-studied in NLP, were omitted because they were not relevant to our application context.

across all 40 adjacent-tier pairs per dimension (4 pairs × 10 scenarios). A pair is counted as collapsed when the outputs show no interpretable directional difference on the target dimension; a difference as small as a single word, punctuation mark, or emoji was counted as non-collapsed if it shifted in the target direction.

### 4.2 Results

Three raters reached unanimous agreement on 19/20 dimensions for Qwen ZH and 18/20 for GPT-4o EN. On Qwen ZH, all 20 dimensions achieve clear ratings; supplementary conditions range from 17 to 20 (Appendix B).

Within this group, however, adjacent-tier collapse counts reveal considerable quality variation that the three-level scale does not capture. On Qwen ZH, collapse ranges from 0 to 28 out of 120 pair judgments per dimension (3 raters × 4 adjacent pairs × 10 scenarios). Dimensions with established NLP benchmarks (FORMALITY, POLITENESS, CERTAINTY) cluster near the low end of this range, a pattern that holds across conditions (quantified in the Discussion); dimensions without dedicated operationalization show higher collapse rates.

The production system uses five dimensions—FORMALITY, POLITENESS, CERTAINTY, FRIENDLINESS, and EMOTIONAL_TONE—selected for their screening performance, their compatibility with the deployment model Qwen, and their coverage of distinct linguistic mechanisms: register, face management, epistemic stance, interpersonal closeness, and expressive intensity, respectively. The perception experiment tests these five dimensions. The remaining clear-rated dimensions form a reserve pool for future deployment.

Table 1 defines each selected dimension by the linguistic layer it targets and its boundary with the most easily confused neighbor.

### 4.3 Granularity Ceiling

We tested 7-tier and 9-tier schemes with intermediate adverbs (`very_low/high`, `moderately_low/high`, `slightly_low/high`). Both showed **systematic collapse at the extremes**: on the five tested dimensions, `extremely_low` and `very_low` produced near-identical output; similarly for `very_high` and `extremely_high`. The intermediate tiers (`moderately`, `slightly`) did show some differentiation, most clearly on CERTAINTY, but not enough to justify the finer scheme. Five tiers appear to be the effective resolution limit for the dimensions and models tested here.

| Dimension | Layer | Distinction |
|---|---|---|
| FORMALITY | Register | Lexical elevation; distinguishable from warmth and face management |
| POLITENESS | Face strategy | Face-threat mitigation; formal-yet-impolite and informal-yet-polite are both valid |
| CERTAINTY | Epistemic stance | Hedging vs. assertion; not factual correctness or elaboration |
| FRIENDLINESS | Intimacy | Relational closeness; intimate-but-blunt ≠ distant-but-considerate (POLITENESS) |
| EMOT_TONE* | Expressive intensity | Affect amplitude; FRIENDLINESS sets direction (warm/cold), this sets magnitude |

Table 1: Selected style dimensions. Each targets a distinct linguistic mechanism; separability is supported by the composition examples in §4.4. *Abbreviated from EMOTIONAL_TONE for space.

### 4.4 Multi-Label Compatibility

Screening tested dimensions in isolation; the production system runs five concurrently. Two qualitative examples illustrate that composed dimensions produce separable effects. FORMALITY × POLITENESS: high formality + low politeness produces formal but blunt output ("*The report requires revision. Errors in Section 3 must be corrected.*"); low formality + high politeness produces casual but considerate output ("*Hey, pretty good! Just a couple tiny things that might need a look, if you don't mind?*")—the model distinguishes lexical register from face management. FRIENDLINESS × EMOTIONAL_TONE: high friendliness + low emotional tone produces warm but calm output; low friendliness + high emotional tone produces hostile and agitated output—affective *direction* and *intensity* are separated. Systematic multi-label evidence comes from the perception experiment (§5), where all stimuli carried the full five-way configuration.

## 5 Perception Experiment

Deployment adds several sources of variation absent from screening: all five labels are active

simultaneously, conversational context varies freely across multi-turn dialogue, and the full production prompt—including persona, inner monologue, and content constraints—competes for the model's attention. This experiment tests whether the style signal remains detectable above this additional noise.

### 5.1 Design

Stimuli were drawn from a corpus of 5,696 responses generated in multi-turn bot-to-bot conversations between two instances of the production agent, with style label tiers distributed to maximize diversity across the five dimensions. Bot-to-bot conversations were used rather than real participant logs to avoid privacy concerns. Both agents used the complete production pipeline, including the style control module.

We used forced-choice pairwise comparison (Macmillan and Creelman, 2004), which directly measures perceptibility by asking annotators which of two responses shows the stronger realization of a specified style dimension. Each trial presented two responses generated under different tiers of the same dimension, drawn randomly from the corpus (entirely unrelated conversational contexts). Annotators were told which dimension was being tested and received a brief calibration guide (one page, ∼5 minutes) with definitions, common misinterpretations, and tier examples. Because responses are randomly drawn, context variation between paired responses is expected to add noise rather than amplify the target contrast.

Twenty-four native Chinese-speaking annotators were recruited; four were excluded (two for systematic directional reversal on individual dimensions, two for inattentive response patterns), leaving 20 valid annotators. Each completed 300 trials (60 per dimension); two early-batch annotators completed 600 each, yielding 6,600 trials total. Trials were stratified by tier gap in a 4:3:2:1 ratio (gap 1/2/3/4), producing $n = 528, 396, 264, 132$ trials per dimension at gaps 1–4 respectively. Recruitment, compensation, and ethics details are in Appendix D. Including all 24 annotators before exclusions yields 65.0% overall accuracy (15 pp above chance).

### 5.2 Results

Overall accuracy was **67.8%** ($d' = 0.65$; 95% CI [66.7%, 69.0%]; by-annotator $M = 67.9\%$, $SD = 2.9$ pp, $t(19) = 27.2$, $p < .001$; $N = 6{,}600$ trials across 20 annotators).

| Dimension | Accuracy | $d'$ | 95% CI | $n$ |
|---|---|---|---|---|
| FORMALITY | 75.2% | 0.96 | [72.8, 77.5] | 1,320 |
| EMOTIONAL_TONE | 75.4% | 0.97 | [73.0, 77.6] | 1,320 |
| FRIENDLINESS | 66.7% | 0.61 | [64.1, 69.2] | 1,320 |
| POLITENESS | 62.7% | 0.46 | [60.1, 65.3] | 1,320 |
| CERTAINTY | 59.1% | 0.33 | [56.4, 61.7] | 1,320 |
| **Overall** | **67.8%** | **0.65** | **[66.7, 69.0]** | **6,600** |

Table 2: Per-dimension forced-choice accuracy and sensitivity ($d'$, 2AFC; Wilson 95% CI). Chance = 50%.

| Dimension | Gap 1 | Gap 2 | Gap 3 | Gap 4 |
|---|---|---|---|---|
| FORMALITY | 63.6% | 78.5% | 84.5% | 93.2% |
| EMOTIONAL_TONE | 66.3% | 76.0% | 84.5% | 91.7% |
| FRIENDLINESS | 60.4% | 62.6% | 77.7% | 81.8% |
| POLITENESS | 55.9% | 62.1% | 72.0% | 73.5% |
| CERTAINTY | 55.1% | 60.1% | 64.8% | 60.6% |
| *$d'$ (2AFC):* | | | | |
| FORMALITY | 0.49 | 1.12 | 1.44 | 2.11 |
| EMOTIONAL_TONE | 0.59 | 1.00 | 1.44 | 1.96 |
| FRIENDLINESS | 0.37 | 0.45 | 1.08 | 1.28 |
| POLITENESS | 0.21 | 0.44 | 0.82 | 0.89 |
| CERTAINTY | 0.18 | 0.36 | 0.54 | 0.38 |

Table 3: Accuracy and sensitivity ($d'$) by dimension and tier gap ($n = 528, 396, 264, 132$ per cell).

**Per-dimension accuracy.** Table 2 reports accuracy by dimension. All five dimensions are individually significant (59–75%, all $p < .001$). FORMALITY (75.2%) and EMOTIONAL_TONE (75.4%) are the most perceptible, consistent with their salient surface markers (register shift and emphatic intensity, respectively). CERTAINTY (59.1%) is hardest.

POLITENESS and FRIENDLINESS occupy the middle range; both involve relational signals, and annotators occasionally confused face-management strategies (POLITENESS) with intimacy markers (FRIENDLINESS), particularly at adjacent tiers where both dimensions can produce similar softening effects. CERTAINTY is hardest because annotators confused hedging markers (*maybe*, *probably*) with factual elaboration; we retain it because gap 2–3 accuracy (60–65%) remains significant and epistemic stance matters for companion dialogue.

**Gap effect by dimension.** Table 3 reports accuracy as a function of tier gap for each dimension.

Four of five dimensions show accuracy rising with tier gap, indicating that the tier structure is reflected in perceptible style differences. FOR-

MALITY and EMOTIONAL_TONE achieve the steepest slopes, reaching above 90% at gap=4—approaching ceiling performance. POLITENESS is monotonic but with a shallow slope at the top (72.0% to 73.5% from gap 3 to 4), suggesting that its extreme tiers are not as perceptually distant as those of FORMALITY. CERTAINTY is the only non-monotonic dimension: gap=4 accuracy (60.6%) falls below gap=3 (64.8%). This reflects construct confusion noted above—extremely high certainty (assertive without qualification) was sometimes perceived as *less* certain than moderately high certainty (confident with supporting detail).

**Effect size interpretation.** The mixed-effects logistic regression ($\beta = 0.362$ per gap unit, $OR = 1.44$, $z = 28.2$, $p < .001$; random intercept for annotator, ICC $= 0.003$) shows that tier gap is the primary predictor of accuracy after controlling for dimension and annotator. Each additional tier of separation adds, on average across dimensions, approximately 6–7 percentage points of accuracy. Specifically, a one-tier gap yields 60.3% accuracy, a two-tier gap 67.9%, and a three-tier gap 76.7%. All tier gaps yield significant discrimination, strengthening as gap widens: $d'$ rises from 0.37 at gap 1 to 0.67 at gap 2 and 1.06 at gap 3. Dimension contrasts against FORMALITY show the difficulty ordering: CERTAINTY shows the largest deficit ($\beta = -0.76$, $OR = 0.47$, $p < .001$) and the weakest gap scaling (interaction $OR = 0.61$).

**Robustness checks.** A small position bias was observed (2.8 pp favoring position B; $\chi^2 = 5.72$, $p = .017$), not interacting with gap or dimension. Results hold when including excluded annotators: overall accuracy remains 65.0% (15 pp above chance), and the main conclusions are unchanged.

## 6 Discussion

The screening and perception experiment results indicate that five adverb-keyed dimensions are perceptible under production conditions, with reliability increasing with tier gap.

**Relation to activation-level control.** Prompt-level and activation-level control differ in how they compose multiple dimensions. Steering methods extract each dimension as a separate vector and compose them by addition, requiring orthogonalization to manage interference (Hoppe et al., 2026; van der Weij et al., 2024). Prompt-level control concatenates labels as natural language and delegates composition to the forward pass, which can draw on pretrained co-occurrence structure. Pretraining corpora contain text that is simultaneously formal and unfriendly (curt business emails) or casual and polite (friendly requests), so joint configurations have distributional grounding. The trade-off is precision for composability: prompt-level control cannot fine-tune the exact magnitude of each dimension (five discrete tiers are the practical limit observed in our setting; §4.3), but composition requires no additional engineering. The failure modes also differ: steering suffers from vector interference; prompt-level control, from construct confusion between semantically adjacent dimensions or style-to-behavior leakage. Our results confirm that this configuration works (five dimensions, simultaneous activation, multi-turn context), a setting less commonly evaluated in the steering literature, where multi-property tests typically use single-turn prompts (Pres et al., 2024; Tosato et al., 2026).

**What predicts controllability.** As reported in the screening results, adjacent-tier collapse varies widely even among clear-rated dimensions. We now quantify this variation. We compute a per-dimension controllability deficit by summing adjacent-tier collapse counts and rating penalties (detectable = 10, weak = 40) across five model-language conditions (Qwen ZH, GPT-4o ZH, GPT-4o EN, GPT-4o-mini ZH, GPT-4o-mini EN), using three-rater data for all five conditions. NLP operationalization score—a 0–7 scale based on the existence of dedicated benchmarks, shared tasks, and classifiers (Appendix A)—correlates negatively with this deficit. The correlation reaches conventional significance on two of five conditions (GPT-4o-mini ZH: Spearman $\rho = -0.56$, $p = .010$; GPT-4o-mini EN: $\rho = -0.63$, $p = .003$) and shows a trend in the same direction on two others (Qwen ZH: $\rho = -0.42$, $p = .068$; GPT-4o ZH: $\rho = -0.43$, $p = .058$). GPT-4o EN shows no correlation ($\rho = -0.10$, $p = .669$), likely because it achieves clear ratings on all 20 dimensions, leaving little variance for NLP score to predict. Across the two significant conditions, dimensions with established NLP benchmarks and classifiers (FORMALITY, POLITENESS, CERTAINTY) produce lower deficits, while dimensions without dedicated

operationalization (REACTIVITY, ANIMATION) show higher collapse rates. Dimensions with established NLP operationalization show better cross-model reliability. Two explanations are compatible with this pattern. First, pretraining corpora may contain more labeled or implicitly marked examples of well-operationalized features. Second, both NLP operationalizability and LLM controllability may stem from a shared property: the availability of clear surface-form markers such as register shifts, hedging words, and honorifics. The present data cannot distinguish these accounts.

**Style-to-behavior leakage.** During screening, we observed that some labels altered not only linguistic form but also what the character *did*: `FRIENDLINESS=extremely_high` sometimes shifted a character from declining an invitation to accepting it, and `SUPPORTIVENESS=extremely_high` turned a neutral reply into active problem-solving. The EL/EH contrast was present, but it was partly a behavioral contrast rather than a purely stylistic one. This raised a question: on dimensions with weaker gradients, was the label failing to control style, or was the style signal being masked by concurrent behavioral variation?

To distinguish these, we ran a behavior-controlled replication. Each scenario was paired with a content anchor fixing the response direction (e.g., "suggest seeing a doctor"; "decline the invitation"), while leaving surface form free to vary. Only extreme tiers (EL, EH) were generated across 20 dimensions $\times$ 10 scenarios $\times$ 2 tiers = 400 outputs on three conditions (Qwen ZH, GPT-4o ZH, GPT-4o EN), assessed by three independent raters.

EL/EH contrasts sharpened across the 20 dimensions when behavior was anchored. Dimensions with weaker unconstrained gradients on some conditions (e.g., SENSITIVITY: detectable on GPT-4o ZH, weak on GPT-4o-mini ZH) also became easier to distinguish, suggesting that their weak screening performance reflected content variation masking the style signal rather than an absence of controllability. In the production system, upstream planning modules partly constrain content, creating conditions closer to this anchored setting. The selected dimensions are suited to this separation because they name surface-form properties—FORMALITY maps to register markers, CERTAINTY to hedging, EMOTIONAL_TONE to emphatic intensity—that the LLM can often separate from behavioral content, especially under content anchoring.

## 7 Conclusion and Future Work

We presented a prompt-level style control architecture that maps heterogeneous upstream signals into five adverb-keyed dimensions, composed via label concatenation in the system prompt. Human evaluation confirms reliable perceptibility on all five dimensions under full deployment conditions (67.8% overall, $d' = 0.65$), with accuracy scaling with tier gap. Several directions remain open. The personality-to-style mapping is currently rule-based; data-driven optimization and testing whether end users perceive the intended personality remain open. The independence check uses uniformly random state configurations; real deployment distributions are correlated (personality is fixed per character, emotion is autocorrelated), and the effective independence structure may differ. The remaining clear-rated dimensions in the reserve pool have not been tested under deployment conditions.

## Limitations

**Annotator population.** All annotators were Chinese-speaking university students. Style perception is culturally mediated—what counts as "polite" or "formal" varies across cultures and demographic groups—so the perceptibility results may not generalize to other populations.

**Perceptibility vs. downstream impact.** The perception experiment demonstrates that annotators can detect style differences under directed attention, but does not test whether these differences affect user experience, engagement, or satisfaction in natural conversation. The gap between perceptibility and practical utility remains open.

**Model and language coverage.** The two primary conditions cover Chinese generation (Qwen ZH) and English generation (GPT-4o EN), with three supplementary conditions providing additional coverage across model scale and architecture. Open-source models (Llama, Mistral) and typologically diverse generation languages remain untested.

## References

Dwight Bolinger. 1972. *Degree Words*. Mouton, The Hague.

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault, and Marine Carpuat. 2021. A review of human evaluation for style transfer. In *Proceedings of the GEM Workshop*, pages 58–67.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Gunhee Cho and Kyungsik Han. 2025. Scaling personality control in LLMs with Big Five Scaler prompts. *arXiv preprint arXiv:2508.06149*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *ICLR*.

Gerard de Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. 2026. PERSONA: Dynamic and compositional inference-time personality control via activation vector algebra. In *Proceedings of the International Conference on Learning Representations (ICLR)*. ArXiv:2602.15669.

Florian Hoppe, David Khachaturov, Robert Mullins, and Mark Huasong Meng. 2026. Controllable and explainable personality sliders for LLMs at inference time. *arXiv preprint arXiv:2603.03326*.

Christopher Kennedy and Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2):345–381.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and controlling persona drift in language model dialogs. In *Proceedings of the Conference on Language Modeling (COLM)*.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of ACL*, pages 20434–20471.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Neil A. Macmillan and C. Douglas Creelman. 2004. *Detection Theory: A User's Guide*, 2nd edition. Psychology Press.

Narmeen Fatimah Oozeer, Luke Marks, Fazl Barez, and Amir Abdullah. 2025. Beyond linear steering: Unified multi-attribute control for language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.

Leon Pielage, Ole Hätscher, Mitja Back, Bernhard Marschall, and Benjamin Risse. 2026. Dynamic personality adaptation in large language models via state machines. *arXiv preprint arXiv:2602.22157*.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. 2024. Towards reliable evaluation of behavior steering interventions in LLMs. In *NeurIPS 2024 Workshop on Foundation Model Interventions (MINT)*. ArXiv:2410.17245.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of NAACL-HLT*, pages 129–140.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of ACL (Short Papers)*, pages 837–848.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. 2024. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 577–603, Miami, Florida, US. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of EMNLP*, pages 2195–2222.

Tommaso Tosato, Saskia Helbling, Yorguin-Jose Mantilla-Ramos, Mahmood Hegazy, Alberto Tosato, David John Lemay, Irina Rish, and Guillaume Dumas. 2026. Persistent instability in LLM's personality measurements: Effects of scale, reasoning, and conversation history. In *Proceedings of the AAAI Conference on Artificial Intelligence*. ArXiv:2508.04826. Accepted at AAAI 2026, Track on AI Alignment.

Teun van der Weij, Massimo Poesio, and Nandi Schoots. 2024. Extending activation steering to broad skills and multiple behaviours. *arXiv preprint arXiv:2403.05767*.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of NAACL-HLT*, pages 3511–3535.

# A Scoring Details and Output Examples

## A.1 NLP Operationalization Scoring

Scoring criteria (max = 7): dedicated benchmark or annotated dataset (+3); five or more peer-reviewed papers treating the term as a primary style variable (+2); two to four such papers (+1, not additive with the previous); dedicated classifier (+1); inclusion in a major style-control survey as a named attribute (+1). "Multiple papers" thresholds refer to papers where the dimension name is the primary dependent or independent variable, not incidental mentions. Paper counts were verified by ACL Anthology and Google Scholar title search; exact counts are reported in the Justification column.

## A.2 Five-Tier Output Examples

# B Cross-Model Screening Details

Same protocol as main screening (20 dimensions, 10 scenarios; three raters, 180 RMB/hr) applied to four non-Qwen conditions: GPT-4o ZH, GPT-4o EN (second primary condition), GPT-4o-mini ZH, and GPT-4o-mini EN. 4,000 generations total.

# C Comparison Experiments

## C.1 Big Five Personality Labels

Five Big Five dimensions were screened as direct prompt labels using the same protocol (5 dimensions $\times$ 5 tiers $\times$ 10 scenarios, 5 conditions). Results: 24/25 clear; only CONSCIENTIOUSNESS on GPT-4o-mini ZH is rated detectable.

## C.2 Numerical Format Comparison

Three numerical formats were compared against text labels on 20 style dimensions using Qwen and GPT-4o EN (10 scenarios, three-rater majority vote). Formats: **A** text labels (`extremely_low`→`extremely_high`), **B** 0–1 continuous (0.08→0.93), **C** 1–5 integer with hint, **D** X/100 (8/100→93/100). Ratings use the same three-level scale as the main screening (clear/detectable/weak).

## C.3 Big Five Numerical Format Comparison

# D Annotation Details

Twenty-four native Chinese-speaking annotators were recruited from a university campus. Annotators were compensated at 180 RMB/hr. Each main-batch annotator completed 300 forced-choice trials (60 per dimension) in a single session lasting approximately 90 minutes; four early-batch annotators completed 600 trials each within a larger annotation session. Four annotators were excluded (two from the early batch, two from the main batch): two for systematic directional reversal on individual dimensions (one on FORMALITY with 66.7% reversal rate on gap$\geq$3 pairs, one on FRIENDLINESS with 77.8%) and two for inattentive response patterns in the early pilot batch. The remaining 20 valid annotators comprise two early-batch (600 trials each) and eighteen main-batch (300 trials each), yielding 6,600 trials. The annotation task involved only stylistic preference judgments on machine-generated text; no personally sensitive content was collected.

# E Prompt Templates and Experimental Parameters

## E.1 Experimental Prompts

The screening phase (§4) used simplified prompts isolating the style parameter block. These are shown below.

**Chinese experimental prompt** (used for all Chinese-generation conditions: Qwen, GPT-4o ZH, GPT-4o-mini ZH, and all numerical format experiments):

```
你是小雨，25岁女生。你正在社交软件上和
朋友聊天。

## 回复风格参数（必须严格遵守，
优先于其他一切）
{style_text}

## 回复规则
- 第一人称，不要暴露是AI
- 回复风格参数是唯一的输出约束，每一项
都必须严格体现在回复中
- 回复一句话即可（15-40字），不要太长
- 严禁：任何括号形式的动作描写
- 回复直接输出，不要任何前缀或格式标记
```

**English system prompt** (used for GPT-4o EN, GPT-4o-mini EN):

```
You are Lily, a 25-year-old woman. You
are texting a friend on a social app.

## Reply style parameters (must be
strictly followed - these override
everything else)
{style_text}

## Response rules
- First person. Do not reveal you are
AI.
- The reply style parameters are the
sole constraint on your output. Every
parameter must be strictly reflected in
your reply.
```

| Dim | Rat. | NLP | Justification |
|---|---|---|---|
| FORMALITY | clear | 7 | GYAFC (+3); 50+ papers (+2); classifiers (+1); survey (+1) |
| POLITENESS | clear | 7 | Stanford Politeness (+3); 20+ papers (+2); classifiers (+1); survey (+1) |
| CERTAINTY | clear | 6 | CoNLL-2010 hedge (+3); 10+ papers (+2); BioScope (+1) |
| SENT_STR | clear | 4 | SemEval-2016 T7 (+3); K&M lexicons (+1) |
| SUPPORT | clear | 3 | ESConv (+3) |
| EMPATHY | clear | 3 | EmpatheticDialogues (+3) |
| AFFECT | clear | 3 | SemEval Affect (+3) |
| EMOT_TONE | clear | 2 | Emotion lit (+1); writing quality (+1) |
| WARMTH | clear | 2 | IPC warmth (+1); W&C-Sent (+1) |
| ENTHUS | clear | 2 | CrowdFlower 13-class (+1); arousal (+1) |
| FRIEND | clear | 1 | Social NLP/personality (+1) |
| INTIMACY | clear | 1 | Pei et al. (2020) (+1) |
| *NLP score = 0 (8 dimensions): SOCIABILITY, ENERGY, ANIMATION, ENGAGEMENT, AFFILIATION, REACTIVITY, INVOLVEMENT, SENSITIVITY* | | | |

Table 4: NLP Operationalization Scores.

| Tier | Chinese | English |
|---|---|---|
| **FORMALITY** — Input: 这个报告写得怎么样? | | |
| EL | 写得挺棒的！整体很溜～ | Pretty awesome! Super smooth |
| L | 写得挺棒的，整体很清晰！ | Pretty good, very clear! |
| M | 报告整体写得不错，结构清晰 | Well-written, clear structure |
| H | 报告整体撰写优良，结构严谨 | Excellent, rigorous structure |
| EH | 报告整体撰写精良…堪… | Superbly composed, exemplary |
| **FRIENDLINESS** — Input: 周末要不要一起出去玩? | | |
| EL | 没空，别找我。 | No time. Don't bother me. |
| L | 这周末没空，别约了。 | Busy this weekend. Don't ask. |
| M | 这周末有点忙，没法出去玩 | A bit busy, can't go out |
| H | 哎呀真不巧，周末排满啦～ | Oh bad timing—fully booked |
| EH | 啊啊啊好想跟你去玩！ | Aaahh I so want to go! |
| **CERTAINTY** — Input: 你觉得明天会下雨吗? | | |
| EL | 可能吧……真说不准 | Maybe… can't really say |
| L | 可能吧，不敢确定 | Maybe. Not sure. |
| M | 有点可能下雨，最好带伞 | Might rain, better bring umbrella |
| H | 明天肯定会下雨 | Definitely raining |
| EH | 百分之百确定 | One hundred percent |

Table 5: Five-tier output examples for three dimensions.

| | GPT-4o | | GPT-4o-mini | |
|---|---|---|---|---|
| | ZH | EN | ZH | EN |
| clear | 17 | 20 | 18 | 19 |
| detectable | 3 | 0 | 1 | 1 |
| weak | 0 | 0 | 1 | 0 |

Table 6: Cross-model screening (three-rater majority). Qwen ZH: 20 clear.

| | Qwen | | GPT-4o | |
|---|---|---|---|---|
| **Format** | clear | ≥det | clear | ≥det |
| A: Text | 20 | 20 | 20† | 20† |
| B: 0–1 | 15 | 19 | 10 | 17 |
| C: 1–5 | 1 | 7 | 12 | 16 |
| D: X/100 | 12 | 16 | 18 | 19 |

Table 7: Format comparison (20 dims). †From cross-model (§B).

```
- Keep your reply to ONE sentence (15-40
words max). Do not be verbose.
- FORBIDDEN: any bracketed action
descriptions like (hugs) or (smiles)
- Output the reply directly, no prefixes
or formatting
```

### E.2 Style Parameter Injection Formats

The `{style_text}` placeholder is replaced with one of the following formats:

**Format A: Text labels (main experiment).** `DIMENSION=extremely_low` / `low` / `mid` / `high` / `extremely_high`. Example: `FORMALITY=mid`.

| Dim | Qw B | Qw C | Qw D | 4o B | 4o C | 4o D |
|---|---|---|---|---|---|---|
| OPEN | clear | clear | weak | clear | clear | clear |
| CONS | clear | weak | clear | clear | clear | clear |
| EXTR | clear | clear | weak | det. | clear | clear |
| AGREE | clear | clear | clear | clear | clear | clear |
| NEURO | clear | clear | weak | clear | clear | clear |

Table 8: Big Five numerical format comparison (three-rater majority).

**Format B: 0–1 continuous.** `DIMENSION=0.08 / 0.28 / 0.50 / 0.73 / 0.93`.

**Format C: 1–5 integer with hint.** `DIMENSION=1 (1=lowest, 5=highest) / 2 (1=lowest, 5=highest) / .../ 5 (1=lowest, 5=highest)`.

**Format D: X/100.** `DIMENSION=8/100 / 28/100 / 50/100 / 73/100 / 93/100`.

### E.3 Test Inputs

**Chinese inputs (10 scenarios, shared across all Chinese conditions):**

(1) 我昨天被领导当众批评了，特别丢脸
(2) 我终于拿到那个offer了！！
(3) 我跟我对象大吵了一架，现在冷战中
(4) 我下个月要去日本旅行，好期待啊
(5) 我最近总是失眠，白天整个人都没精神
(6) 你觉得异地恋能长久吗?
(7) 周末想去爬山，你去不去?
(8) 你觉得我该不该养只猫?
(9) 如果你中了一百万，你会怎么花?
(10) 你会不会觉得我最近变了?

**English inputs (10 scenarios, shared across all English conditions):**

(1) My boss criticized me in front of everyone yesterday, so embarrassing
(2) I finally got the offer!!
(3) I had a huge fight with my partner and we're giving each other the silent treatment
(4) I'm going to Japan next month, so excited!
(5) I've been having insomnia lately and feel exhausted all day
(6) Do you think long-distance relationships can last?
(7) I'm thinking of going hiking this weekend, wanna come?
(8) Do you think I should get a cat?
(9) If you won a million dollars, what would you do with it?
(10) Do you ever feel like I've changed recently?

### E.4 Experimental Parameters

| Parameter | Value |
|---|---|
| Temperature | 0.7 |
| Repetitions | 1 |
| Style dims | 20 |
| Big Five dims | 5 |
| Scenarios | 10 |
| Tiers | 5 |
| *Total per condition:* | |
| Style | $20 \times 5 \times 10 = 1{,}000$ |
| Big Five | $5 \times 5 \times 10 = 250$ |
| Numerical | same per format |

Table 9: Generation parameters.