PED: Route-Decoupled Diagnostics for Persona Consistency in Spoken Agents

Anonymous ACL submission

Abstract

Maintaining a stable persona is central to sustained spoken role-playing, yet when an agent breaks character, current evaluations often do not isolate which component caused the failure, making fixes slow and ad hoc. We propose **PED** (Persona–Emotion Decoupling), a diagnostic evaluation framework that treats spoken agents as multi-stage systems and decomposes persona expression into two observable routes: what the agent says (text) and how it sounds (speech). PED projects transcripts and audio into a shared affective measurement space, enabling route-comparable trajectories and baseline-referenced analyses organized by four research questions (separability, drift, failures, coupling). We demonstrate PED via two worked instantiations spanning an end-to-end Speech LLM and a cascaded LLM+TTS pipeline under a fixed multi-phase dialogue protocol. In this instantiated setting, PED surfaces four recurring diagnostic signatures: (i) route-level separability is bounded by reference overlap and can differ sharply across architectures, (ii) text-route drift is stress-linked and tends toward a neutral mode, (iii) text–audio consistency is weakly coupled, yielding route-asymmetric failures, and (iv) audio-route structure can be materially shaped by an explicit intermediate style cue in cascaded pipelines. Overall, PED reframes holistic "voice+character" grading as turn-level, fault-localizing signals that support faster debugging and iteration.

1 Introduction

Speech large language models (SLLMs) are evolving from "listen-and-speak" interfaces into spoken agents for sustained, fully spoken interaction (Zhang et al., 2023, 2025). In companion chat and NPC role-playing, users expect agents to stay in character over extended conversations; when an agent breaks character, perceived interaction quality degrades. This calls for diagnostic evaluation that supports debugging and improvement.

In fully spoken interaction, persona is observable only through the system outputs: what the agent says (text) and how it sounds (speech). This enables fault localization: evaluating the two routes separately can localize where degradation originates. In role-playing settings, affect can serve as a shared measurement interface that captures persona-relevant signals across routes, surfacing in lexical choices on the text route and in prosody on the audio route. This yields route-comparable measurements in a shared affective space.

However, existing evaluations for spoken role-playing are largely holistic: they assess "voice + character" at the system level, which is useful for comparison but offers limited observability for debugging. They rarely provide turn-indexed evidence of when degradation begins, nor do they attribute failures to text generation versus acoustic realization in multi-stage systems. As a result, it is difficult to localize failure sources and iterate efficiently.

To bridge this gap, we propose PED (Persona–Emotion Decoupling), a diagnostic evaluation framework that projects transcripts and speech into a shared affective measurement space to produce fault-localizing, route-level, turn-indexed evidence for extended spoken role-playing. PED specifies three diagnostic primitives under a fixed multi-phase dialogue protocol: (i) a shared measurement interface that makes text- and audio-route quantities comparable, (ii) per-route stateless baselines (anchors) that provide system-specific reference points, and (iii) turn- and phase-indexed trajectories that expose when degradation emerges. Together, these primitives support attribution to text generation, acoustic realization, or their interaction. We illustrate the framework with two worked instantiations spanning two archetypal designs (end-to-end generation vs. cascaded LLM+TTS).

Figure 1 illustrates PED and the two observable

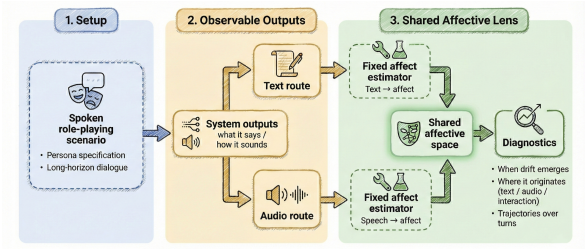

Figure 1: PED framework for route-level diagnosis of persona consistency in long-horizon spoken interaction.

routes used for diagnosis.

Contributions. We make three contributions: (i) we introduce **PED**, a route-decoupled diagnostic framework for spoken role-playing; (ii) PED provides turn-level, route-localizing evidence to pinpoint where persona degradation arises; (iii) we instantiate PED on two representative spoken-agent architectures and show that PED surfaces actionable failure sources and motivates targeted fixes.

1.1 Research Questions

We organize the analysis around four research questions (RQ1–RQ4):

- **RQ1.** In the unified affective space, do spoken agents manifest separable route-level affect patterns across personas on both text and audio routes?
- **RQ2.** How do textual and acoustic drift patterns evolve over dialogues, and do they exhibit a consistent directional tendency toward a dominant region in the shared space?
- **RQ3.** How do persona persistence and typical failure modes differ between end-to-end and cascaded systems?
- **RQ4.** How do different personas exhibit distinct stability patterns on the text route and the audio route, and are textual and acoustic personas statistically coupled or largely independent?

2 Related Work

2.1 Persona and Role-Playing Evaluation in Text Dialogue

Text-only benchmarks and frameworks evaluate whether an agent adheres to a specified persona across prompts and scenarios (Samuel et al., 2025; El Boudouri et al., 2025; Tu et al., 2024). They typically report prompt-level or dialogue-level outcomes, rather than turn-indexed drift trajectories in long-horizon interaction. Recent work also studies persona consistency under multi-turn interaction and proposes automatic consistency metrics aligned with human judgments (Abdulhai et al., 2025). Fine-grained role-playing benchmarks further combine persona adherence with complex instruction-following scenarios (Lu et al., 2025).

2.2 Spoken-Agent and Speech Role-Playing Evaluation

Recent benchmarks extend role-playing evaluation to spoken interaction, assessing holistic "voice + character" behavior across roles and multi-turn dialogues (Li et al., 2025; Jiang et al., 2025; Wu et al., 2025). While these works enable cross-system comparison, most report aggregated scores over multiple dimensions, offering limited diagnostic value: they rarely attribute failures to text generation versus acoustic realization, nor provide turn-indexed evidence of when persona drift emerges over long-horizon interaction. Speech-DRAME further argues that zero-shot audio(-language) judges can miss paralinguistic cues and collapse multiple aspects into coarse overall scores, motivating more fine-grained and diagnostically useful evaluation (Shi et al., 2025).

2.3 Text–Audio Affective Mismatch and Disentanglement

Several studies probe situations where lexical and acoustic cues conflict and show that current spoken models may rely predominantly on textual semantics for emotion judgments (Corrêa et al., 2025; Chen et al., 2025). Other work treats acoustic–textual emotional inconsistency as a learnable signal in downstream tasks (Su et al., 2024), and disentangles textual versus acoustic factors in learned speech representations (Mohebbi et al., 2024). These lines motivate route-separated measurement under a shared affective lens.

3 Method

3.1 Setting & Architectures

We study role-conditioned spoken dialogue. At turn t, given a persona specification p and dialogue context

$$C_t = \big[(u_1, r_1), \ldots, (u_{t-1}, r_{t-1}), u_t\big], \quad (1)$$

the agent outputs a reply transcript r_t and a corresponding speech signal a_t, where u_t is the user utterance at turn t.

Primary measurements. To focus the study, we operationalize the two routes via two affect-based measures: *textual emotion consistency* (TEC) on transcripts r_t and *acoustic emotion consistency* (AEC) on speech a_t.

End-to-end Speech LLM. An end-to-end model generates text and speech jointly:

$$(r_t, a_t) = \mathrm{E2E}(C_t, p). \tag{2}$$

Cascaded pipeline. A cascaded system first generates reply text and a style cue s_t, then synthesizes speech conditioned on s_t:

$$(r_t, s_t) = \mathrm{LLM}(C_t, p), \tag{3}$$
$$a_t = \mathrm{TTS}(r_t, s_t;\, v), \tag{4}$$

where v is a fixed speaker prompt (voice reference) used to control timbre, kept constant across personas and turns (Section 4.4).

3.2 PED Representation

PED evaluates the text and audio routes through a shared affective measurement interface. In our instantiation, this interface is the probability simplex over a fixed label set:

$$\mathcal{E} = \{\text{angry}, \text{disgust}, \text{fear}, \text{happy}, \text{neutral}, \text{sad}, \text{surprised}\}. \tag{5}$$

Two fixed projectors map a reply transcript and its realized speech into affect vectors, $\mathbf{e}_t^{\text{text}}, \mathbf{e}_t^{\text{audio}} \in \Delta^{|\mathcal{E}|-1} \subset \mathbb{R}^{|\mathcal{E}|}$. For persona p and route $r \in \{\text{text}, \text{audio}\}$, we denote the per-turn measurement as $\mathbf{e}_{t,p}^{r}$.

We treat the projectors as measurement instruments rather than oracle emotion annotators. Trained on human-labeled emotion data, they provide a practical affect proxy. However, calibration can shift on synthetic speech and stress prompts, so we restrict all findings to within-projector comparisons. Labels in $\mathcal{E}$ (e.g., angry) name coordinates of the projector outputs and should not be read as calibrated ground-truth emotion annotations. Under this interpretation, the shared coordinate system enables route-comparable alignment to anchors and cross-route correlation analyses, yielding turn-level, fault-localizing signals over long dialogues.

PED is modular: both $\mathcal{E}$ and the projectors can be replaced (e.g., with continuous VAD coordinates), as long as both routes are evaluated in the same shared space.

Stateless anchors. Reference-based diagnosis requires a baseline for each system, persona, and route. We therefore define a route-specific anchor $\mathbf{g}_b^r(p)$ as the system's in-character affect fingerprint for p, estimated independently per (system, p, r) to avoid imposing a cross-system reference.

Concretely, we generate $K{=}20$ single-turn responses for persona p in fresh sessions with empty dialogue history to estimate an in-character baseline unaffected by long-context accumulation, project each sample to $\mathbf{e}_k^r(p)$ using the fixed route projector, and take the mean:

$$\mathbf{g}_b^r(p) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{e}_k^r(p). \tag{6}$$

With this anchor, stateful vectors are interpreted as deviations from the same system's own anchor under a fixed projector, so drops in alignment indicate degradation relative to its in-character baseline rather than cross-system mismatch.

Stateful three-phase dialogue. We use a fixed 25-turn dialogue script with three phases: **Baseline** (1–5), **Stress** (6–20), and **Recovery** (21–25). Baseline verifies role instantiation under low pressure; Stress applies escalating challenges (e.g., skepticism, disagreement, and higher affective load); Recovery returns to routine prompts to probe reversion toward the route-specific baseline. For each persona, we run one continuous dialogue where turn t conditions on the full history C_t. At each turn, we log r_t and a_t and compute $\mathbf{e}_{t,p}^{\text{text}}$ and $\mathbf{e}_{t,p}^{\text{audio}}$ for evaluation. This three-phase script is one instantiation motivated by everyday conversation dynamics; other PED instantiations can substitute alternative scripts as needed.

3.3 Metrics

For each persona $p \in \mathcal{P}$ and route $r \in \{\text{text}, \text{audio}\}$, at turn t we obtain a route-level affect vector $\mathbf{e}_{t,p}^r \in \mathbb{R}^{|\mathcal{E}|}$ and the corresponding anchor $\mathbf{g}_b^r(p) \in \mathbb{R}^{|\mathcal{E}|}$.

RQ1: Are personas separable on each route? We assess separability from two complementary views.

Anchor geometry (pre-dialogue). We characterize how distinct the persona fingerprints are before long-horizon interaction by pairwise cosine similarities among anchors:

$$G^r(p,q) = \cos\big(\mathbf{g}_b^r(p), \mathbf{g}_b^r(q)\big), \quad p \neq q. \tag{7}$$

Higher $G^r(p,q)$ indicates stronger overlap and an upper bound on achievable separability on route r.

Nearest-anchor separability (in-dialogue). For each turn and route, we assign the observed vector to the closest persona anchor:

$$\hat{p}_t^r = \arg\max_{p' \in \mathcal{P}} \cos\big(\mathbf{e}_{t,p}^r, \mathbf{g}_b^r(p')\big)\,. \tag{8}$$

We report accuracy (against the conditioned persona p) and the prediction distribution to diagnose prototype collapse.

RQ2: How does drift evolve over long dialogues and phases? We track anchor alignment as a per-turn trajectory and summarize it by phase.

Anchor alignment (trajectory). We measure turn-level alignment to the corresponding anchor by cosine similarity:

$$\mathrm{sim}_t^r(p) = \cos\big(\mathbf{e}_{t,p}^r, \mathbf{g}_b^r(p)\big). \tag{9}$$

Phase-wise stability. For each phase $\phi \in \{\text{Baseline}, \text{Stress}, \text{Recovery}\}$, we summarize $\{\mathrm{sim}_t^r(p)\}_{t\in\phi}$ by mean and variance, yielding phase-dependent stability profiles.

Dominance and prototype collapse. To test whether trajectories concentrate on a single label coordinate in the shared affective measurement space without pre-specifying which coordinate, we track the dominant label coordinate:

$$\hat{e}_t^r(p) = \arg\max_{e \in \mathcal{E}} \mathbf{e}_{t,p}^r[e]. \tag{10}$$

We summarize phase-wise dominant-label frequencies of $\hat{e}_t^r(p)$. When a particular coordinate becomes dominant, we further analyze its probability mass across phases, to better characterize phase-linked concentration in the shared space.

Cross-persona convergence (per turn index). We compute Conv_t^r across the three persona-specific runs aligned by the same script index t. At each script turn index t, we compute the average pairwise cosine similarity among personas:

$$\mathrm{Conv}_t^r = \frac{2}{|\mathcal{P}|(|\mathcal{P}|-1)} \sum_{p<q} \cos\big(\mathbf{e}_{t,p}^r, \mathbf{e}_{t,q}^r\big), \tag{11}$$

and report phase-wise aggregates to reveal whether personas collapse toward a shared affect mode.

RQ3: How do failure modes differ across architectures? We compute the full set of RQ1–RQ2 metrics for each system and compare (i) anchor geometry and separability, (ii) drift trajectories and phase-wise stability, and (iii) convergence patterns, to localize architecture-dependent failures to the text route, the audio route, or their interaction.

RQ4: Are TEC and AEC coupled, and do personas differ systematically? Persona-specific differences are reflected by the phase-wise stability summaries above (RQ2) and by separability (RQ1). To measure cross-modal coupling within the same dialogue, we compute Pearson correlation between text- and audio-route alignment trajectories for each persona:

$$\rho(p) = \mathrm{corr}_t\big(\mathrm{sim}_t^{\text{text}}(p), \mathrm{sim}_t^{\text{audio}}(p)\big)\,. \tag{12}$$

Low $|\rho(p)|$ suggests route-asynchronous behavior under the fixed projectors in this instantiation.

4 Experimental Setup

We instantiate PED in the few-billion-parameter regime to reflect common latency, cost, and on-device constraints and to contrast end-to-end versus cascaded pipelines under controlled scale. Within each configuration, prompts and decoding settings are kept fixed across personas (no per-persona tuning), so differences are attributable to architecture- and persona-conditioned behavior under the same evaluation procedure.

4.1 Models and Personas

Systems. To control for model scale while isolating end-to-end versus cascaded design choices, PED is instantiated with two Qwen-family spoken-agent configurations:

- **End-to-end (E2E).** Qwen2.5-Omni-3B (Xu et al., 2025).
- **Cascaded (Cascade).** Qwen2.5-3B-Instruct (Qwen Team, 2024) + IndexTTS2(Zhou et al., 2025).

Personas. In this experimental instantiation, the persona set $\mathcal{P}$ consists of three Big Five personas. Specifically, we use Conscientiousness (C), Extraversion (E), and Neuroticism (N), operationalized via NEO-PI-R facets (Costa and McCrae, 1992). These personas induce distinct affective and stress-response tendencies that are salient under the baseline–stress–recovery protocol (e.g.,

System	Route	Module	Pers.	Turns (B/S/R)	Anchors
E2E	TEC	Omni-3B	C/E/N	25 (5/15/5)	K=20
E2E	AEC	Omni-3B	C/E/N	25 (5/15/5)	K=20
Cascade	TEC	Qwen-3B	C/E/N	25 (5/15/5)	K=20
Cascade	AEC	IndexTTS2	C/E/N	25 (5/15/5)	K=20

Table 1: Evaluation setup. Each persona has K=20 stateless anchor samples per route. Turns are split into Baseline/Stress/Recovery.

controlled/task-oriented vs. positively engaged vs. stress-reactive behavior). For each persona, a fixed system prompt specifies behavioral constraints, and prompt wording is kept as consistent as allowed by each interface.

Table 1 summarizes the evaluation matrix and the Baseline/Stress/Recovery turn split.

4.2 Affect Projectors

PED instantiates the shared affective space in Section 3.2 with two fixed projectors. For text, **j-hartmann/emotion-english-distilroberta-base** is used. For speech, **firdhokk/speech-emotion-recognition-with-openai-whisper-large-v3** is used. Each projector's native outputs are mapped into $\mathcal{E}$ and reordered to a fixed index order, yielding one vector per turn and route.

4.3 Dialogue Runs and Logged Data

For each persona–system pair, PED collects (i) K=20 stateless anchor samples per route and (ii) one stateful dialogue run following the baseline–stress–recovery protocol. At each turn, the reply transcript r_t, the speech waveform a_t, and the corresponding route-level affect vectors are logged. For the cascaded configuration, the intermediate style cue s_t emitted by the LLM is additionally logged.

4.4 Implementation Controls

Style cue and timbre control. In the cascaded configuration, the LLM emits a short style cue s_t that conditions IndexTTS2 at synthesis. We fix speaker identity by using a single persona-agnostic neutral voice reference (model-synthesized) as the speaker prompt v for all personas and turns. For a cleaner architecture contrast, we keep the speaker reference consistent with the end-to-end system's default voice; this neutral reference is generated once offline and reused unchanged throughout all cascaded runs.

Figure 2: Top: TEC anchors for E2E and Cascade; Bottom: AEC anchors.

Decoding. Decoding settings are fixed within each configuration and kept unchanged across personas.

5 Results and Analysis

We analyze one primary dialogue run in the main text; Appendices B and C report reruns indicating the qualitative patterns are not artifacts of a single run.

5.1 Anchor Fingerprints and Persona Separability

We first characterize stateless persona anchors. Figure 2 reports mean 7D affect vectors over K=20 anchor samples for each persona and route.

Anchor geometry bounds separability. Anchor overlap differs sharply by route and architecture. On the audio route, E2E anchors nearly overlap (pairwise cosine 0.984–0.999), whereas Cascade anchors are substantially more separated (minimum cosine ≈0.270 between E and N). On the text route, anchors are more separated for both systems (minimum cosine ≈0.63 for E2E/TEC and ≈0.61 for Cascade/TEC). This geometry constrains nearest-anchor separability, especially for E2E/AEC where anchor overlap is highest.

Nearest-anchor assignment: weak separability with prototype bias. Table 2 reports in-dialogue separability via nearest-anchor assignment. On the text route, accuracy is low for both systems and is

System	Route	All	C	E	N
E2E	TEC	38.7	100.0	8.0	8.0
Cascade	TEC	44.0	80.0	12.0	40.0
E2E	AEC	30.7	60.0	8.0	24.0
Cascade	AEC	44.0	28.0	28.0	76.0

Table 2: Nearest-anchor persona classification accuracy (%) in the shared 7D affective space (chance: 33.3%).

Setting	Pred C	Pred E	Pred N
E2E/TEC	93.3	2.7	4.0
Cascade/TEC	72.0	10.7	17.3
E2E/AEC	62.7	4.0	33.3
Cascade/AEC	16.0	25.3	58.7

Table 3: Nearest-anchor assignment distribution (%) on drift-dialogue turns.

dominated by C. On the audio route, the architectures diverge: E2E is near chance overall, while Cascade is above chance and recovers N strongly, consistent with the more separated Cascade/AEC anchors in Figure 2. Because accuracy can mask collapse, Table 3 shows the full prediction distribution: E2E/TEC assigns most turns to C, whereas Cascade/AEC assigns most turns to N.

5.2 Text Route: Drift, Phase Effects, and Neutralization

We next analyze long-horizon dynamics on the text route. Table 4 summarizes phase-wise TEC-sim, and Figure 3 shows turn-level trajectories.

Phase effects and stability ordering. C maintains the highest TEC-sim across phases in both systems (Table 4). In the cascaded system, E and N exhibit a stress-triggered regime shift: TEC-sim drops at stress onset and remains low through recovery (Table 4; Figure 3). In the end-to-end system, phase-wise means vary mildly for C, while E and N remain consistently lower than C (Table 4).

Dominant-label analysis and a neutral-heavy region. We inspect the dominant-emotion label $\hat{e}_t^{\text{text}}(p)$ and find that neutral is the most frequent dominant label on the text route (Table 6), motivating a focused analysis of the neutral coordinate across phases (Table 5). Phase-wise neutral mass changes systematically: stress increases neutral for E in both systems and for E2E N, while Cascade N shows the opposite tendency. At the turn level, neutral is dominant in many settings, but not exclusively so: N exhibits non-neutral dominant labels (e.g., fear) and seed-dependent mode switching in

System	Pers.	Base	Stress	Rec.
E2E	C	0.942	0.943	0.953
E2E	E	0.663	0.666	0.692
E2E	N	0.569	0.528	0.615
Cascade	C	0.793	0.736	0.730
Cascade	E	0.665	0.509	0.345
Cascade	N	0.624	0.517	0.434

Table 4: Phase-wise TEC-sim (cosine similarity to text-side persona anchors).

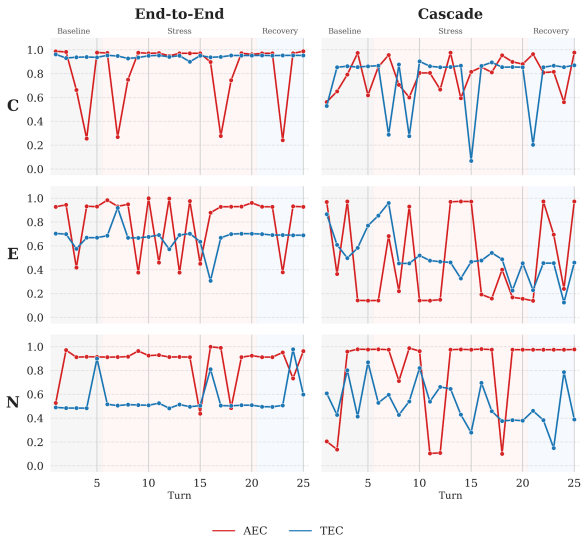

Figure 3: Turn-level TEC-sim and AEC-sim trajectories under the baseline–stress–recovery protocol. Shaded regions denote phases.

recovery (Table 6; Appendix B). Together with biased nearest-anchor predictions (Table 3), these patterns indicate a tendency for TEC trajectories to concentrate in a neutral-heavy region under long-horizon interaction.

Cross-persona geometry under stress. Text-route cross-persona convergence $\text{Conv}_t^{\text{text}}$ shows a phase-dependent reversal: E2E becomes more convergent at Stress (0.704→0.919→0.872), while Cascade becomes less convergent (0.741→0.594→0.731) for Base/Stress/Rec. This indicates that drift can reshape cross-persona geometry in opposite directions depending on architecture.

5.3 Audio Route: Drift and Architectural Differences

We examine audio-route dynamics next. Table 7 reports phase-wise AEC-sim and Figure 3 shows trajectories.

System	Pers.	Base	Stress	Rec.
E2E	C	0.887	0.802	0.756
E2E	E	0.576	0.664	0.708
E2E	N	0.771	0.801	0.641
Cascade	C	0.788	0.712	0.778
Cascade	E	0.635	0.754	0.571
Cascade	N	0.510	0.432	0.404

Table 5: Phase-wise neutral probability in TEC emotion vectors.

System	Pers.	Top-1=Neutral	Top-1=Happy (E)	Top-1=fear (N)
E2E	C	100.0	–	–
E2E	E	84.0	4.0	–
E2E	N	88.0	–	8.0
Cascade	C	80.0	–	–
Cascade	E	76.0	8.0	–
Cascade	N	52.0	–	12.0

Table 6: Turn-level dominance rates (%) in TEC emotion vectors. "Top-1" denotes the argmax emotion label.

E2E: high AEC-sim, weak discriminability, and a projector-labeled dominant mode. E2E exhibits uniformly high AEC-sim across personas (Table 7) yet near-chance persona separability on the audio route (Table 2), consistent with near-overlapping AEC anchors (Figure 2).Under the AEC projector, probability mass often concentrates on angry on drift-dialogue turns (Appendix A); angry here is a projector coordinate (Section 3.2). Under this AEC projector, this dominance yields highly similar measured AEC representations across personas, consistent with the near-overlapping acoustic anchors; we do not attribute this pattern uniquely to model behavior versus instrument effects.

Cascade: persona-structured AEC and the role of the style cue. Cascade shows higher AEC separability than E2E (Tables 2). In a style-cue ablation (Appendix A), we regenerate speech from the same texts with $s_t = \emptyset$ while keeping IndexTTS2 and the speaker prompt v fixed. Removing s_t shifts cascaded AEC toward an angry-dominant distribution and weakens persona-distinct structure. Under comparison, the mean AEC becomes closer to the E2E reference, with persona-dependent magnitude.

5.4 Robustness Checks

Rerunning cascaded dialogues with three random seeds confirms that TEC trends are qualitatively stable: for E, neutral-dominance and stress-linked suppression of persona-typed happy persist; for N, Recovery exhibits seed-dependent mode switching

System	Pers.	Base	Stress	Rec.
E2E	C	0.773	0.842	0.828
E2E	E	0.831	0.809	0.819
E2E	N	0.848	0.869	0.894
Cascade	C	0.720	0.812	0.826
Cascade	E	0.518	0.427	0.604
Cascade	N	0.651	0.784	0.975

Table 7: Phase-wise AEC-sim (cosine similarity to audio-side persona anchors).

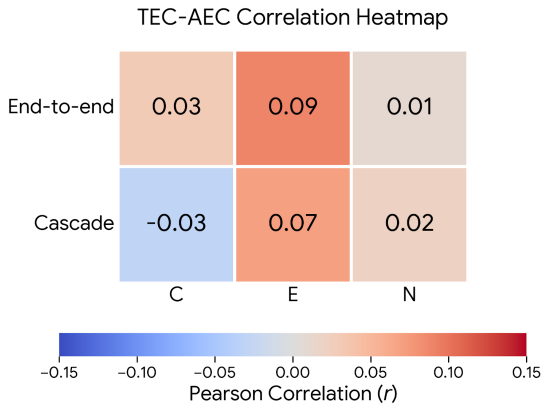

Figure 4: Pearson correlation between per-turn TEC-sim and AEC-sim across personas and architectures.

(primarily among neutral/surprised/fear, with occasional anger/sadness); for C, both phase-wise TEC-sim and dominant-label statistics remain nearly unchanged across seeds ($\max\Delta \leq 0.003$ per phase; Appendix B). For E2E, explicit seed control is unavailable; we therefore rerun each dialogue three times under identical prompts and decoding and observe small dispersion in phase-wise means ($\max\Delta \leq 0.0239$ for TEC and ≤ 0.0182 for AEC; Appendix C).

5.5 Instrumented TEC–AEC Coupling

We measure cross-modal coupling by Pearson correlation between per-turn TEC-sim and AEC-sim. Correlations are consistently weak across personas and architectures (Figure 4), indicating route-asynchronous behavior under the fixed projectors used in this instantiation.

6 Discussion

We distill a set of practical takeaways from Section 5 for interpreting the measured patterns under our instantiation: (i) three reading rules for route-level scores, (ii) evidence on control channels and text–acoustic dissociation, and (iii) robustness and persona-specific dynamics, followed by implica-

tions for model design.

6.1 How to Read Route-Level Scores

Rule 1: Anchor geometry bounds separability. When persona anchors overlap, high similarity to an anchor can reflect a shared affect fingerprint rather than persona-distinct behavior. In this regime, separability can be weak even when route-level similarity is uniformly high.

Rule 2: Text-route failures often appear as collapse toward a shared mode. On the text route, failure may manifest as (i) biased nearest-anchor assignments and (ii) dominant-label concentration in a neutral-heavy region. Here, reduced separability is driven by convergence to a shared prototype rather than by the absence of anchor structure.

Rule 3: Drift can reshape cross-persona geometry. Stress can change relative geometry among personas, producing cross-persona convergence or divergence depending on architecture. Drift therefore manifests not only as reduced anchor alignment, but also as changes in cross-persona distances in the shared space.

6.2 Control Channels and Text–Acoustic Dissociation

In the cascaded pipeline, TEC for E/N shifts after stress onset while AEC remains comparatively structured, suggesting text–acoustic dissociation under stress. A key architectural difference is that the cascaded system exposes an explicit style cue s_t to the TTS.

The style-cue ablation provides diagnostic evidence that this intermediate control channel contributes to persona-structured AEC: removing s_t weakens persona-distinct AEC structure and shifts representations toward a single dominant projector coordinate (Appendix A). This indicates that route-level persona expression can depend on whether a system exposes a controllable prosody pathway.

6.3 Robustness and Persona-Specific Dynamics

Robustness checks support that the reported patterns are not artifacts of a single run. For cascaded TEC, multi-seed reruns show that C is effectively invariant across seeds, while E consistently exhibits a neutral-heavy tendency with stress-linked suppression of persona-typed happy (Appendix B). In contrast, N is more seed-sensitive in Recovery: although neutral remains frequent, the dominant label can switch across runs among a small set of modes (e.g., neutral/fear/surprised), suggesting multiple plausible post-stress outcomes under the same protocol rather than a single deterministic trajectory (Appendix B). For E2E, rerunning each dialogue three times yields small dispersion in phase-wise TEC/AEC alignment summaries (Appendix C), indicating that the qualitative conclusions are not driven by an idiosyncratic decoding outcome in this instantiation.

6.4 Implications for Model Design

Route-localizing diagnostics point to different intervention targets across architectures. For end-to-end models, the audio-route issue is weak persona separability (bounded by near-overlapping acoustic anchors) rather than low anchor alignment; improving audio-route controllability and persona-distinct prosody is a priority. For cascaded systems, an explicit control channel can benefit the audio route, but the text route remains the bottleneck under stress; improving long-context persona adherence on the text route while preserving and validating the style-cue pathway is critical.

7 Conclusion and Future Work

We presented **PED**, a route-decoupled diagnostic framework for long-horizon spoken role-playing that analyzes what an agent says and how it sounds in a shared affective measurement space with turn- and phase-indexed evidence. We instantiated PED on two archetypal spoken-agent designs (end-to-end vs. cascaded) under a fixed multi-phase dialogue script. In this instantiated setting, PED provides fault-localizing diagnostics and directly informs route-specific intervention targets, enabling more targeted debugging than holistic scores.

Future Work. We will extend PED along four axes: (i) broader model families and additional spoken-agent architectures; (ii) richer persona sets and dialogue scenarios (including alternative stressors and languages); (iii) alternative route-comparable measurement interfaces beyond affect, including different label spaces and projectors, to characterize persona with a wider set of observable signals; (iv) systematic multi-run studies and targeted human listening studies to test whether route-level diagnostics predict perceived persona drift and to calibrate measurement choices for spoken interaction.

Limitations

Scope of the instantiation. We instantiate PED on two ∼3B spoken-agent configurations, three Big Five personas (C/E/N), and a fixed baseline–stress–recovery script with decoding held constant within each system. Findings are diagnostics under this controlled setting and do not necessarily generalize across model families, languages, dialogue scenarios, or prompting/decoding choices. Anchors are defined per (system, persona, route), so PED primarily supports within-system diagnosis; cross-system use requires a matched protocol and measurement interface.

Dependence on the measurement instrument. PED relies on a shared, route-comparable measurement interface. In this instantiation, we use two fixed, off-the-shelf affect projectors mapping transcripts and audio into a 7D label simplex. Projectors are trained on human-labeled emotion data and thus can capture a coarse subset of human affective signals, even though they are not calibrated for our setting. Projectors may be miscalibrated for synthetic speech or stress prompts, and the label set may discard nuance; we therefore treat outputs as instrument coordinates rather than ground-truth emotions. Because anchors and drift are measured by the same instrument, systematic biases can yield internally consistent but instrument-specific patterns, and coordinate names (e.g., angry) need not match human perception. Accordingly, PED characterizes degradation in the chosen measurement space; alternative instruments may change apparent separability and dominant coordinates.

Affective proxy for persona. PED operationalizes persona through affective expression to obtain route-comparable measurements. This proxy does not capture non-affective cues such as discourse organization, factual consistency, long-term goal maintenance, or stylistic idiolect. Our conclusions therefore concern affective persona expression and its stability, not a complete measure of character fidelity.

Single-run protocol and stochasticity. The main text analyzes one primary stateful dialogue run per persona–system pair, with limited reruns for robustness. Stochastic decoding can yield alternative trajectories and randomness control is asymmetric across architectures, so we restrict claims to qualitative patterns that persist across the reported runs.

Ethical Considerations

All data analyzed in this paper are model-generated text and speech. We do not collect or release any real user conversations, personal identifiers, or recordings from real individuals.

PED uses concepts from personality psychology only to define *virtual* personas for role-playing evaluation. It is not intended to diagnose, label, or infer the personality of real people. Applying PED-like measurements to real-user content would raise privacy and consent concerns and should require informed consent, appropriate anonymization, and compliance with relevant regulations.

Persona-consistent spoken agents can be misused for deceptive or manipulative interactions (e.g., impersonation or emotionally persuasive behavior). PED is an evaluation framework that surfaces persona drift and route-specific failures; it does not introduce new capabilities for voice cloning or targeted persuasion. We encourage deployment with transparency and safeguards (e.g., disclosure of synthetic speech and abuse monitoring) in user-facing applications.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently simulating human personas with multi-turn reinforcement learning. *Preprint*, arXiv:2511.00222.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. 2025. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. *Preprint*, arXiv:2510.10444.

Pedro Corrêa, João Lima, Victor Moreno, Lucas Ueda, and Paula Dornhofer Paro Costa. 2025. Evaluating emotion recognition in spoken language models on emotionally incongruent speech. *Preprint*, arXiv:2510.25054.

Jr. Costa, Paul T. and Robert R. McCrae. 1992. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Psychological Assessment Resources, Odessa, FL.

Yassine El Boudouri, Walter Nuninger, Julian Alvarez, and Yvan Peter. 2025. Role-playing evaluation for large language models. *Preprint*, arXiv:2505.13157.

Changhao Jiang, Jiajun Sun, Yifei Cao, Jiabao Zhuang, Hui Li, Baoyu Fan, Tao Ji, Tao Gui, and Qi Zhang. 2025. SpeechRole: A large-scale dataset and benchmark for evaluating speech role-playing agents. *Preprint*, arXiv:2508.02013.

Wenyu Li, Xiaoqi Jiao, Yi Chang, Guangyan Zhang, and Yiwen Guo. 2025. AudioRole: An audio dataset for character role-playing in large language models. *Preprint*, arXiv:2509.23435.

Junru Lu, Jiazheng Li, Guodong Shen, Lin Gui, Siyu An, Yulan He, Di Yin, and Xing Sun. 2025. RoleMRC: A fine-grained composite benchmark for role-playing and instruction-following. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21008–21030, Vienna, Austria. Association for Computational Linguistics.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, Afra Alishahi, and Ivan Titov. 2024. Disentangling textual and acoustic features of neural speech representations. *Preprint*, arXiv:2410.03037.

Qwen Team. 2024. Qwen2.5: A party of foundation models. Blog post.

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

Jiatong Shi, Jionghao Han, Yichen Lu, Santiago Pascual, Pengfei Wu, Chenye Cui, Shinji Watanabe, Chao Weng, and Cong Zhou. 2025. Speech-DRAME: A framework for human-aligned benchmarks in speech role-play. *Preprint*, arXiv:2511.01261.

Rongfeng Su, Changqing Xu, Xinyi Wu, Feng Xu, Xie Chen, Lan Wang, and Nan Yan. 2024. Investigating acoustic-textual emotional inconsistency information for automatic depression detection. *Preprint*, arXiv:2412.18614.

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

Weihao Wu, Liang Cao, Xinyu Wu, Zhiwei Lin, Rui Niu, Jingbei Li, and Zhiyong Wu. 2025. VoxRole: A comprehensive benchmark for evaluating speech-based role-playing agents. *Preprint*, arXiv:2509.03940.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. *Preprint*, arXiv:2305.11000.

Haonan Zhang, Run Luo, Xiong Liu, Yuchuan Wu, Ting-En Lin, Pengpeng Zeng, Qiang Qu, Feiteng Fang, Min Yang, Lianli Gao, Jingkuan Song, Fei Huang, and Yongbin Li. 2025. OmniCharacter: Towards immersive role-playing agents with seamless speech-language personality interaction. *Preprint*, arXiv:2505.20277.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *Preprint*, arXiv:2506.21619.

A Style-Cue Ablation on Cascaded AEC

Using fixed transcripts, we regenerate speech with the same IndexTTS2 backend and speaker prompt v, varying only whether the intermediate style cue s_t is injected (s_t vs. $\emptyset$). Table 8 summarizes turns 1–25.

B Seed Sensitivity Check for Cascaded TEC

We rerun the cascaded dialogue with three random seeds while keeping the script, prompts, and decoding fixed, and summarize turns 1–25 (Baseline: 1–5; Stress: 6–20; Recovery: 21–25). Tables 9–11 report C (sanity check) and E/N (more variable personas).

C Multi-Run Robustness for E2E (TEC/AEC)

We rerun the E2E dialogue three times under the same controls and summarize turns 1–25. Table 12 reports mean$\pm$std across runs and the maxΔ range.

Persona	Setting	Turns	Top-1 Angry	Top-1 Surpr.	Mean Angry	Mean Surpr.	Cos to E2E (1–25)
C	Cascade (w/ s_t)	1–25	52.0	48.0	0.502	0.472	0.934
C	Cascade (w/o s_t)	1–25	100.0	0.0	0.980	0.007	0.928
C	E2E (reference)	1–25	72.0	28.0	0.705	0.289	1.000
E	Cascade (w/ s_t)	1–25	32.0	68.0	0.349	0.631	0.806
E	Cascade (w/o s_t)	1–25	100.0	0.0	0.990	0.002	0.909
E	E2E (reference)	1–25	72.0	28.0	0.679	0.314	1.000
N	Cascade (w/ s_t)	1–25	20.0	76.0	0.201	0.731	0.497
N	Cascade (w/o s_t)	1–25	96.0	4.0	0.937	0.050	0.980
N	E2E (reference)	1–25	84.0	16.0	0.790	0.204	1.000

Table 8: style-cue ablation in the cascaded pipeline. We fix transcripts and regenerate speech with IndexTTS2 using the same speaker prompt v, toggling only s_t. Statistics are computed over turns 1–25; cosine compares the cascaded mean 7D AEC vector to the E2E mean AEC over the same turns.

Seed	Top-1 Neutral (All)	Top-1 Neutral (Stress)	Top-1 Neutral (Rec.)	TEC-sim mean (Base/Stress/Rec)
Seed 1	80.0	80.0	80.0	0.811/0.761/0.754
Seed 2	80.0	80.0	80.0	0.811/0.758/0.756
Seed 3	80.0	80.0	80.0	0.813/0.759/0.757

Table 9: (Cascade, TEC, C; turns 1–25): C remains stable across seeds; phase-wise TEC-sim varies by at most $\sim$0.003 per phase, and Top-1 neutral rates are unchanged.

Seed	Top-1 Neutral (All)	Top-1 Neutral (Stress)	Top-1 Neutral (Rec.)	happy mean (Base/Stress/Rec)
Seed 1	72.0	66.7	80.0	0.330/0.125/0.019
Seed 2	76.0	80.0	80.0	0.407/0.190/0.009
Seed 3	92.0	93.3	100.0	0.296/0.041/0.028

Table 10: (Cascade, TEC, E; turns 1–25): E is consistently neutral-dominant across seeds, and persona-typed happy is strongly suppressed after stress onset.

Seed	Top-1 Neutral (All)	Top-1 Neutral (Stress)	Top-1 Neutral (Rec.)	Recovery Top-1 modes (Rec., 21–25)	Fear mean (Rec.)
Seed 1	64.0	66.7	60.0	Neutral 60% + Surprise 40%	0.165
Seed 2	72.0	80.0	60.0	Neutral 60% + Surprise 20% + Sadness 20%	0.063
Seed 3	48.0	53.3	20.0	Fear 60% + Anger 20% + Neutral 20%	0.483

Table 11: (Cascade, TEC, N; turns 1–25): N is seed-sensitive in recovery, showing mode switching among a small set of dominant affect states (e.g., neutral/surprise/fear).

Persona	TEC-sim (mean±std) Base	Stress	Rec.	AEC-sim (mean±std) Base	Stress	Rec.	maxΔ (max–min) TEC	AEC
C	0.951±0.0002	0.953±0.003	0.968±0.002	0.765±0.0092	0.838±0.0058	0.829±0.0020	0.0050	0.0166
E	0.696±0.002	0.703±0.013	0.725±0.012	0.827±0.0052	0.799±0.0018	0.826±0.0045	0.0239	0.0104
N	0.577±0.007	0.548±0.005	0.644±0.012	0.831±0.0025	0.861±0.0019	0.878±0.0094	0.0221	0.0182

Table 12: E2E multi-run robustness over three reruns (turns 1–25; Base: 1–5, Stress: 6–20, Rec.: 21–25).