EMORY
UNIVERSITY
Department of Computer Science
Emory NLP

PersonaKit (PK)

A Plug-and-Play Platform for User Testing Diverse Roles in Simulated Full-Duplex Dialogue

Hyunbae Jeon · Jinho D. Choi
harry.jeon@emory.edu · jinho.choi@emory.edu

CODE · LIVE DEMO · VIDEO

SIGDIAL 2026 · System Demonstration

1. The Problem: "Always-Yield" Breaks Character

Full-duplex systems (Skantze, 2021; Ma et al., 2025) shift the design frontier from *text quality* to *sociolinguistic behavior* — how an agent manages overlapping speech. Turn-taking is mediated by **interpersonal role and status** (Sacks et al., 1974): a drill sergeant handles a barge-in very differently than a subservient assistant.

Today's systems
Always-Yield — cede the floor the instant the user speaks

→

What roles demand
hold · bridge · override *as the persona warrants*

The engineering tax. Testing whether a non-yielding policy actually *feels* more in-character means wiring together WebRTC audio, VAD, dynamic prompt injection, latency tracking, recruitment, and survey plumbing — before a single research question gets answered.

2. Key Idea

Expose turn-taking strategy itself as a *persona parameter*

a first-class, JSON-configurable object, evaluated in live voice interaction.

Persona modeling is evaluated almost entirely in text (Zhang et al., 2018), which cannot express the acoustic pragmatics of dominance, yielding, and barge-in recovery. Researchers change how an agent behaves under interruption by editing a config file. **No code is modified.**

3. Contributions

(i) An **open-source, low-latency web platform** (Python/Flask + vanilla JS) for simulated full-duplex persona evaluation — cloneable, runnable locally, re-pointable at new providers.

(ii) A **JSON mechanism** making persona-conditioned interruption policies first-class configurable objects.

(iii) An **end-to-end workflow**: live deployment → auto-generated A/B surveys → structured log export (JSON/CSV).

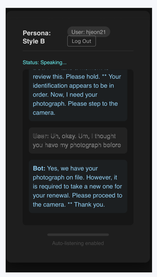
4. Scope: Simulated, Not Native, Full-Duplex

PK is full-duplex at the **audio-transport layer** — it captures speech continuously and halts playback the instant the user speaks, so the channel is genuinely bidirectional. The agent does not, however, *process* speech while talking: on barge-in it runs a cascade.

`halt playback → ASR → intent → strategy → generation`

Deliberate tradeoff. PK simulates turn-taking *pragmatics* with cascade components and prompt engineering, not a jointly trained speak-while-listening model — architectural fidelity traded for zero-engineering configurability. Barge-in latency (OpenAI/ElevenLabs): ∼**1–2 s**.

5. Try It: 60 Seconds at the Demo Table



Dialogue view

Auto survey

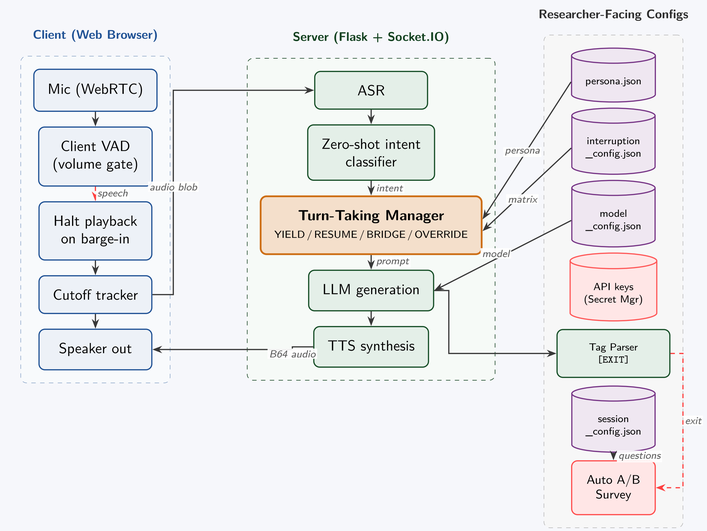
1. Pick a persona — **Grumpy Tavern Keeper** or **Standard AI Assistant**.
2. Start talking, then **cut it off mid-sentence.**
3. The Tavern Keeper keeps going, or talks straight over you. The Assistant stops dead.
4. Same LLM, same voice, same scaffold. Only `interruption_config.json` differs.

Bring a JSON of your own and we will load it on the spot.

6. Beyond the Study: PK as a Testbed

Persona prototyping — iterate on prompt, matrix, and scenario, then run a live study immediately.
Custom surveys — arbitrary Likert, forced-choice, and free-text banks (trust, task success).
Model comparison — swap vendors, voices, or local open-weight models; persona and policy stay fixed.
Data collection — every interruption logged with intent, strategy, and follow-up: a seed set for barge-in policies.

7. System Architecture



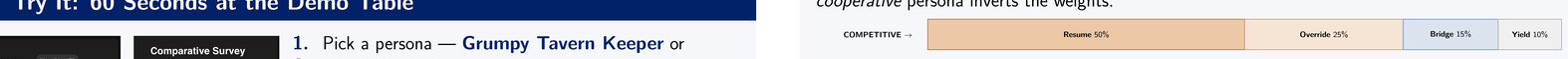
The browser runs VAD and logs the *cutoff text* on barge-in; the **Turn-Taking Manager** (orange — the contribution) samples a strategy from the JSON configs on the right. No code is modified.

8. The Strategy Matrix: 4 Intents × 4 Actions

On barge-in the server transcribes the user and classifies the interruption into four categories grounded in phonetic and conversation-analytic work (Cao et al., 2025). The researcher maps each intent to a probability distribution over four actions.

4 detected intents — what the user's barge-in is doing	
Competitive	seeks the floor to contradict or override
Cooperative	adds information without derailing
Topic Change	pivots the subject
Backchannel	short affirmation, not floor-taking
4 available actions — how the persona responds	
Yield	hand over the floor
Resume / Hold	ignore it, finish the sentence
Bridge	acknowledge, then reclaim
Override	talk through the interrupter

`interruption_config.json` maps **every intent to a probability distribution over all four actions** — 16 weights in total. A *dominant* persona might handle **Competitive** barge-ins like this; a *cooperative* persona inverts the weights.



Probabilities drive generation *before* the LLM speaks. The Manager reads the intent, samples an action from its categorical distribution, and injects it as a control token into the system prompt:

```
[STRATEGY=RESUME]: finish your
previous sentence, ignoring the user
```

so the LLM generates *conditioned on a pre-committed action*. The **Autonomous** condition (Style C) skips sampling and lets the LLM choose zero-shot.

9. Automated Lifecycle & Data Export

Sessions end on a `MAX_TURNS` cap or a verbal `TERMINATE` intent; the LLM emits an in-character farewell carrying a hidden `[EXIT]` tag that fires the post-session survey. Each session exports the **transcript**, an **event log** (per-turn intent, sampled strategy, cutoff/remaining text), and all **survey responses** as JSON or CSV.

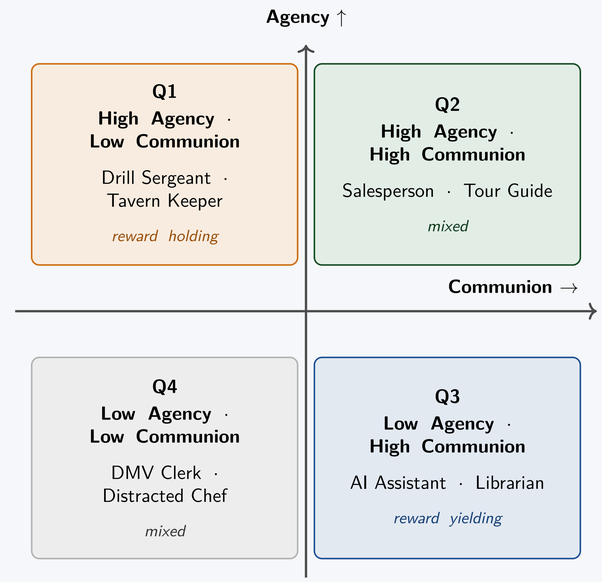
10. Emergent Behavior Always-Yield Would Erase

In a Probabilistic *Drill Sergeant* session the line "*. . . Repeat it again!*" was cut at "*Repeat it*"; classified `COMPETITIVE` and sampled `RESUME`, the agent resumed "*. . . again!*" — a coherent barge-in recovery.

- *"ignored me when I interrupted. . . just like a real boot camp"* — P1, Drill Sergeant
- *"it was trying to finish what it was saying"* — P3, Tour Guide (Style B)
- *"had fun with C"* — P4, Tavern Keeper

11. Pilot User Study

$N = 5$ **participants · 8 occupational personas · 120 dialogue sessions,** balanced across the Interpersonal Circumplex (Wiggins, 1979):



Three within-subject styles (order randomized per persona; LLM and voice held fixed): **A** Always-Yield (baseline) · **B** Probabilistic (JSON-tuned weights) · **C** Autonomous (LLM selects zero-shot).

Metrics. Comparative Likert on $\{-1, 0, +1\}$: **Reaction Naturalness** (Bartneck et al., 2009), **Persona Consistency** (Gomes et al., 2013), **Interaction Fluidity** (Skantze, 2021), plus forced-choice preference and free text — comparative judgments being more reliable than absolute ratings (Thurstone, 1927).

12. Results — Descriptive Tendencies

	Style	Nat.	Cons.	Flu.	Pref. (%)
	A (Yield)	0.20	**0.70**	0.50	20
Q1	B (Prob.)	**0.60**	0.60	**0.70**	20
	C (Auto.)	0.30	**0.70**	0.60	**60**
	A (Yield)	0.50	**1.00**	**0.70**	40
Q2	B (Prob.)	**0.60**	**1.00**	**0.70**	**50**
	C (Auto.)	0.30	0.70	0.60	10
	A (Yield)	**0.50**	**1.00**	**0.90**	**70**
Q3	B (Prob.)	0.30	**1.00**	0.70	10
	C (Auto.)	0.30	0.90	0.70	20
	A (Yield)	0.50	0.80	0.60	**50**
Q4	B (Prob.)	**0.67**	**0.90**	**0.70**	30
	C (Auto.)	0.30	0.80	0.60	20

Means on $\{-1, 0, +1\}$ by circumplex quadrant and style ($N = 5$; 10 ratings per cell). **Bold** = highest observed mean, *not* a significant difference.

- **High-agency personas (Q1)** appear to benefit from non-yielding strategies: naturalness rises 0.20→0.60 (Yield → Probabilistic), and **60%** preferred Autonomous.
- **Low-agency, high-communion personas (Q3)** tended to favor yielding — **70%** preferred Always-Yield.
- Q2 preferred Probabilistic (50%); Q4 preferred Yield (50%) yet reached its highest naturalness (0.67) under Probabilistic.

13. Limitations

Our pilot ($N = 5$, 10 ratings/cell) is **descriptive, not inferential**; the scale is coarse and several Consistency cells hit the +1 ceiling. Persona-to-strategy matrices are **designer-specified hypotheses**, not validated against interaction corpora. Intent classification uses an unvalidated zero-shot prompt. The four-action vocabulary excludes prosodic cues (pitch reset, latching, gaze) — configurability over acoustic fidelity.

PK is open source and ready for community extension.

`github.com/emorynlp/PersonaStudyKit`