

Personas on Demand: Hypernetwork Synthesis of Unseen Users

Tyson Walsh Daniel Runfola Steven Morse Trenton W. Ford

Department of Data Science, William & Mary

{thwalsh, dsmillerrunfol, stmorse, twford}@wm.edu

Abstract

Simulating an online community at scale calls for language models that can stand in for many distinct users at once, not collapse them into a single generic voice. Prior work shows that a hypernetwork-driven language model can efficiently rebuild known users by storing a small per-user profile vector rather than separate fine-tuned models, one per user. This study explores whether those profile vectors support a harder task: generating behaviorally consistent new users who were never directly observed. Our frozen LLM uses adapter offsets emitted by a hypernetwork. We sample profiles within and beyond the real-user range, then score the generated replies on behavioral measures to determine whether synthetic users match the cohort patterns their profiles imply. We show that synthetic users reproduce behavioral patterns similar to those of real users in the training set: in short threads they stay close to their target profile, and across a longer multi-turn forum simulation they stay largely coherent and do not drift from it. Performance declines as profiles move farther from the distribution of users seen in the original data, suggesting that the method works best for synthetic users that recombine observed patterns rather than extrapolate beyond them.

1 Introduction

A central goal of social-media analysis is understanding how individuals and communities behave online. One way to study that behavior is to simulate it: build models of individuals in a community, then place them in experimental environments where they interact. Large language models can express assigned personality traits (Jiang et al., 2024), and agent-based models can simulate large online populations (Murić et al., 2022). Our approach combines these demands: one simulator must preserve distinct language behavior for thousands of users without storing a separate model for each.

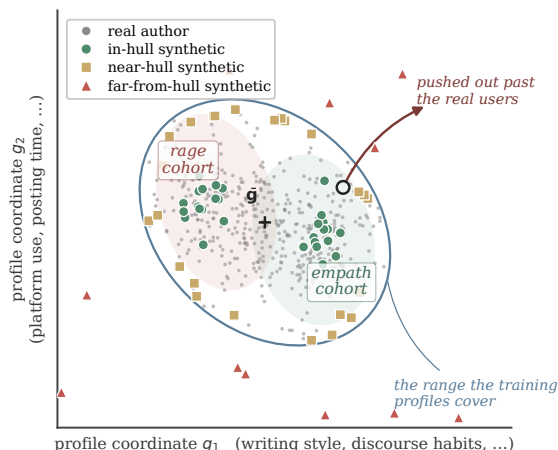


Figure 1: We generate three kinds of synthetic users from the profiles of real authors. **In-hull** users are blended from the dense interior of the real-user range. In-hull inference is our deployment setting. **Near-hull** users are also inside that range and are mixed with less-even weights; the schematic separates the two families for legibility, and by measured distance to the population average near-hull users sit slightly closer to it than in-hull users do (Appendix D). **Far-from-hull** users are pushed deliberately outside the range, along the ray from the population average $\bar{g}$ through one real anchor. For every synthetic user, the trained hypernetwork turns the profile vector into a per-user adapter offset, exactly as it does for a real author from training.

The straightforward way to personalize a language model is to train and store one fine-tuned copy per user. Parameter-efficient fine-tuning (PEFT) methods such as adapters and LoRA reduce how much must be trained and stored per user (Houlsby et al., 2019; Hu et al., 2022), but they still assign separate weights to each user. Both approaches therefore face the same scaling problem: total storage grows linearly with the number of users, and inference loads different weights at each persona change. Prompting places persona information in the input context rather than in user-specific weights; prior work evaluates assigned per-

sonality and role descriptions through this interface (Jiang et al., 2024; Wang et al., 2024). Retrieval-based methods instead select items from a user’s history and add them to the context (Salemi et al., 2024), but imperfect retrieval can degrade output (Park and Lee, 2024). Surveys identify consistency as a recurring concern in role-playing and personalization systems (Tseng et al., 2024).

Hypernetworks offer a different path: they read an input vector and emit weights for another network (Ha et al., 2017). In our system, one shared hypernetwork reads an 18-number user profile and emits that user’s LoRA offsets on demand, leaving a short numeric per-user state rather than a separate adapter checkpoint. HyperPEFTNet used this profile-to-offset design on a smaller backbone for authors seen during training (Walsh et al., 2026). We ask whether the same design can synthesize unseen users and how far those profiles can move from the training distribution before generated behavior stops matching the target.

Stated formally: *can a hypernetwork trained to rebuild observed authors synthesize behaviorally consistent unseen users, and where does that capability break down?* **H1** asks whether synthetic users behave the way their profiles say they should. **H2** asks whether they behave as consistently as real users from the training corpus. **H3** asks whether they stay close to their targets across many generations. All three use behavioral measures covering emotional, stylistic, and timing signals.

Contributions. We contribute four things. First, our *synthesis engine* turns a short user profile into a per-user adapter on demand and runs on a frozen Pythia-1.4B backbone. Second, nine behavioral measures test whether generated text matches the cohort patterns implied by that profile. Third, we show that synthetic users recover the emotional patterns observed in real users and that those patterns weaken as profile vectors move away from the training distribution. Fourth, we show that target match stays stable across many turns of conversation.

Section 2 reviews prompt-based, retrieval-based, and adapter-based approaches to personalization and places the profile-vector hypernetwork among them. Section 3 describes the data, the two models, the profile vector, the synthesis engine, the behavioral measures, and the stability and compositionality indices. Section 4 tests H1, H2, and H3 against the real-user baseline and the three synthesis settings, then examines compositionality. Section 5

concludes by stating where the method holds and where it breaks. The appendices give the full architecture, feature, measure, and index definitions.

2 Related Work

What we mean by “persona.” A *persona*, as we use the term, is an author’s stable, recurring, text-derived pattern of expression: how the author tends to write rather than a list of facts or preferences. Prior work measures these patterns with style representations, stylometric features, and politeness markers (Wegmann et al., 2022; Patel et al., 2023; Pennebaker et al., 2001; Danescu-Niculescu-Mizil et al., 2013). The cohort labels used later come from stylometric and emotional distributions.

Personalized language generation. Recent personalization systems use three broad interfaces. Prompt-based methods place a profile, personality, or role description in the input context (Zhang et al., 2018; Jiang et al., 2024; Wang et al., 2024). Retrieval-based methods select items from a user’s history and add them to that context (Salemi et al., 2024). Parameter-based methods store or assemble user-specific parameter-efficient fine-tuning (PEFT) weights (Tan et al., 2024a,b). Hypernetwork-emitted adapters are another parameter-based interface: a profile is encoded once, and the hypernetwork emits the adapter used for that user (Ha et al., 2017; Walsh et al., 2026; Tan et al., 2026).

User-level hypernetworks. HyperLoRA, HyperX, and Text-to-LoRA generate PEFT parameters at the task or language level rather than representing individual users (Lv et al., 2024; Üstün et al., 2022; Charakorn et al., 2025). At the user level, Profile-to-PEFT maps encoded textual profiles to adapter parameters and explicitly evaluates unseen users (Tan et al., 2026). Our numeric profile interface asks a different question: how far can a sampled profile move from the training users before generated behavior stops matching its target? We test controlled mixtures of real-user profiles and profiles pushed beyond the real-user range.

Persona synthesis and interpolation. Persona synthesis spans several distinct tasks. PersonaChat conditions dialogue on explicit facts and preferences (Zhang et al., 2018), while PersonaLLM and RoleLLM assign personality or character descriptions through a prompt-based interface (Jiang

et al., 2024; Wang et al., 2024). Separate stylometry work measures community language variation, content-independent style, and topic leakage in authorship analysis (Lucy and Bamman, 2021; Wegmann et al., 2022; Sawatphol et al., 2024). Large-scale agent-based simulation models individualized behavior across online populations rather than generating per-user language through a profile-to-weight mechanism (Murić et al., 2022).

The closest style-representation studies also encode numerical summaries of how people write. Patel et al. (2023) learn interpretable style embeddings from synthetic stylometry data, while Wegmann et al. (2022) train content-controlled style representations for authorship verification. We use numeric author profile vectors for a different purpose: using the profile vectors as input to a hypernetwork and evaluating generated behavior.

Wang et al. (2025) offer the closest comparison: they infer implicit profiles from dialogue and learn a sampler that produces diverse user profiles for multi-turn simulation. We ask a different question: how far can a synthetic profile move before its generated behavior stops matching the target? We test one interior mixture that weights its real authors relatively evenly and another that allows one real author to dominate. We also test profiles pushed beyond the real-user range.

Evaluating personas. Persona evaluation draws on three measurement families. Style-embedding and authorship models compare texts in a learned representation space (Wegmann et al., 2022; Patel et al., 2023). Stylometric analyses count observable linguistic features such as pronouns, capitalization, and politeness markers (Pennebaker et al., 2001; Argamon et al., 2009; Danescu-Niculescu-Mizil et al., 2013). Emotion resources score affective content, including GoEmotions and the rule-based VADER system (Demszky et al., 2020; Hutto and Gilbert, 2014). We combine stylometric and emotion measures because a synthetic user has no unique gold reply.

We compare cohorts with Hedges’ g , the bias-corrected standardized mean difference, and resample bootstrap intervals at the author level unless a caption says otherwise (Hedges, 1981; Efron, 1979; Deen and de Rooij, 2020). Prior work evaluates personality or role descriptions supplied through prompts (Jiang et al., 2024; Wang et al., 2024), and generative-agent systems maintain memory, reflection, and plans across interactions (Park et al.,

2023). Our setting differs because user variation is stored in hypernetwork-emitted weights, and target match is measured directly across turns.

3 Methods

3.1 Data

HyperPEFT-LoRA and vanilla LoRA train on the same English Reddit discussion-forum corpus. It contains 10,000 authors, collected until each had 300 top-level threads; cleaning then dropped threads whose context or reply did not survive, leaving 288 to 300 per author. We keep every thread with its reply, so training and evaluation use the same format, with text cleaned for platform artifacts and split within author by thread. At evaluation the model writes a complete reply from thread context, with no user-specific reference text to copy from. Reddit’s per-author depth supports stable cohort labels at the 10,000-author scale (§3.5). The same user pool supplies the real-user evaluation in §4.1 and the three synthesis settings, making the four results directly comparable.

3.2 Models

Both models use the same frozen Pythia-1.4B backbone, LoRA site geometry, precision, optimizer, and seed. The models update different weights. HyperPEFT-LoRA also reads a profile vector and uses the steering step at inference. Comparing HyperPEFT-LoRA with vanilla LoRA therefore tests the systems as a whole. The closer comparison is HyperPEFT-LoRA against the control below, which differs only in profile input and steering.

Vanilla LoRA (baseline). One LoRA adapter is fine-tuned on the full author population, with no per-user signal. The same low-rank update applies to every forward pass. This is the assumed persona-free reference point. Vanilla LoRA and the control below answer different questions: vanilla LoRA bounds what plain domain fine-tuning achieves with no per-user information, while the control shares the trained treatment stack and removes the per-user input and the steering step. The control, not vanilla LoRA, is the profile-aware comparison. HyperPEFT-LoRA is the profile-aware treatment in this comparison.

HyperPEFT-LoRA (treatment). HyperPEFT-LoRA keeps the pretrained backbone weights and both LoRA factors frozen during training (Figure 2). For each user, H_φ reads an 18-dimensional profile vector and produces a small per-user offset

at every LoRA site. At each site, the frozen A_j projects the input, then ΔB_j maps it to the output as $s_j \Delta B_j A_j x$. We optimize reply cross-entropy, $\mathcal{L} = \text{CE}_{\text{reply}}$, and clamp every emitted value. Appendices A and B give architecture and training details.

At inference we also steer the model’s hidden state. The residual stream is the running hidden state passed from one decoder layer to the next. At each layer we estimate one direction shared across the population. We remove any overlap between that direction and the user’s cohort direction, then subtract the remaining component at layers 15 to 23. This follows the residual-stream direction-ablation method of Arditi et al. (2024). Appendix F gives the estimation procedure and operating point for this experiment.

HyperPEFT-LoRA (control). For this ablation, we load the trained hypernetwork, zero its per-user input vector, and disable the steering step. Everything else is unchanged. A HyperPEFT-LoRA (HP-LoRA) minus control difference therefore covers the per-user channel and the steering step together, not the per-user channel alone. The control is the persona-free floor for the trained model, and its near-zero effects on the three body measures rule out the hypernetwork acting merely as a regularizer.

3.3 Profile Vector

The hypernetwork receives one fixed 18-number profile vector $\mathbf{g}_u$ for each author. We compute every coordinate from that author’s training-split threads. Aggregating across those threads reduces variation tied to any one topic or context, but it does not make every coordinate topic-free. Table 7 groups the coordinates into style, discourse, platform, and temporal features. Examples are capitalization rate and readability; question rate and mean post length; subreddit concentration and external-link rate; and posting-hour entropy and reply-delay variability. Subreddit concentration is the author’s share of activity in their most-used subreddit, not the topic of a single thread. Appendix E lists all 18 features. We z-score them with training statistics.

At inference we drop any author with a coordinate beyond $\pm 4\sigma$ and clip the rest at $\pm 3\sigma$. Far-from-hull synthesis skips that filter and clips at $\pm 5\sigma$. We exclude sentiment, personality proxies, and emotion scores from $\mathbf{g}_u$ because those quantities later define or score the outcomes; otherwise the hypernetwork could reproduce a mea-

sure by reading it straight from its input. We test whether reply length or shared thread content explains the cohort differences. Appendix H gives the controlled estimates. §4.1 compares HyperPEFT-LoRA with the control on identical thread contexts.

We started with 57 candidate features and used backward elimination to choose the final 18. At each step we removed the lower-ranked profile features, then remeasured held-out prediction loss and separation in the emitted-offset space. This produced five reported sizes: $K = 57, 33, 18, 8$, and 1. Among them $K=18$ gives the largest delta-separation score, which measures how distinctly the hypernetwork’s per-user offsets separate the cohorts. Its score is about $7.7\times$ the $K=57$ value, while held-out prediction loss changes little across the same sweep. Appendix E reports the full trajectory.

3.4 The Synthesis Engine

The synthesis engine produces unseen users at inference time (Figure 1 shows the three synthesis settings we sample). It has two parts: a sampler that draws a new profile vector, and the trained hypernetwork that turns that vector into a per-user adapter offset. The offset is injected into the frozen backbone in the same way as for a real user. The synthesis engine does not retrain the hypernetwork. It only samples new points in the profile space the hypernetwork was already trained to interpret. Appendix D walks through the sampler on a 2D projection.

We construct new profiles in two ways. *Interpolation* samples a new profile within the range of users we already have; *extrapolation* pushes past them into profiles no real user occupies. Because profile space is continuous, an interpolated profile can land between users from different cohorts rather than inside any single one.

Two of the three settings interpolate. For each new user, the sampler draws a few real users at random from the training pool, draws Dirichlet weights over them (non-negative numbers that sum to one), and takes the weighted average of their profiles. The new point is one no real user occupies, and the convex combination stays inside the hull of the sampled anchors. **In-hull** synthesis is our deployment setting. It draws five anchors and uses a Dirichlet concentration of $\eta=2.0$, which keeps the weights relatively even. Its density gate measures each candidate’s mean distance to its ten nearest training profiles and rejects only candidates

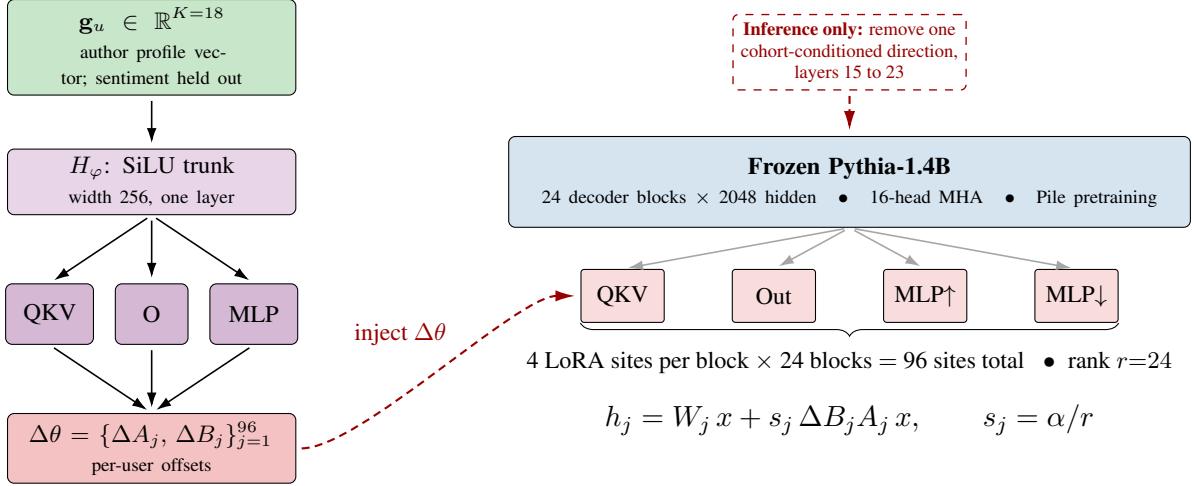


Figure 2: HyperPEFT-LoRA on a frozen Pythia-1.4B backbone. The hypernetwork H_φ injects per-user offsets into every LoRA site; an inference-time steering step (top box) removes one cohort-conditioned direction, orthogonal to that cohort’s persona direction, from the residual stream of decoder layers 15 to 23. Architecture details: Appendix A; steering details: §3.2 and Appendix F.

above the training pool’s 95th percentile. **Near-hull** synthesis draws twelve anchors and lowers the concentration to $\eta=1.0$, so one anchor can dominate. Its density gate is the stricter of the two: it keeps only candidates below the 50th percentile, in neighborhoods denser than the training median. Both settings stay inside the real-user range (Appendix N).

The third setting, **far-from-hull**, extrapolates instead. We take a real user’s profile $\mathbf{g}_a$ and push it away from the average of all real-user profiles $\bar{\mathbf{g}}$, giving $\mathbf{g}' = \mathbf{g}_a + \kappa(\mathbf{g}_a - \bar{\mathbf{g}})$. The difference $\mathbf{g}_a - \bar{\mathbf{g}}$ points from the average toward the user, and κ sets how far past the user the new profile lands. We sweep κ from 3, the smallest boundary-test pushout used here, to 100; at $\kappa=3$, nearly every sampled profile already sits beyond the 99th-percentile boundary of the training pool (Appendix M). This setting asks whether the adapter still behaves predictably on profiles beyond anything in training. Across all three synthesis settings the cohort-mean QKV offset stays within 42.9° of its real-user counterpart in the same cohort. Full-offset magnitude is lower for the interior mixtures and higher for the far-from-hull pushout (Appendix M).

3.5 Behavioral Measures

We score generated text on nine behavioral measures. Six are *out-of-band*, meaning that their defining feature is not present in the profile vector (Table 1). An effect on these measures therefore can-

not come from reading the label directly from the input. The other three are *in-band* positive controls. For each of those, the profile vector carries the defining feature or a close proxy. The zero-input control tests whether either kind of effect survives removing the profile input and steering (§4.1). Tempo is the exception: reply-delay mean defines its cohort direction, but generated replies have no inter-post timing, so their realized tempo score uses reply length (Appendix G).

Scorer grounding. We use the GoEmotions label set (Demszky et al., 2020) and the released checkpoint identified in Appendix C. We run the checkpoint unchanged to produce emotion probabilities for both the training corpus and the generated replies. Scores from the training corpus fix the GoEmotions cohort quintiles, while scores from generated replies measure realized behavior. The same scorer therefore defines the targets and evaluates the replies.

We compute pronoun and profanity rates from LIWC lexical categories (Pennebaker et al., 2001). Prior computational-politeness work motivates treating politeness as its own pragmatic dimension (Danescu-Niculescu-Mizil et al., 2013). The rates, the Table 1 emotion composites, the cohort percentile cuts in Appendix G, and the two indices in §3.6 are project-defined calculations. They are not outputs of separately trained models.

For each cohort comparison we report Hedges’ g (Hedges, 1981) with bootstrap 95% confidence intervals; label construction and scorer details are

Table 1: The behavioral measures. A measure is in-band when the profile vector $\mathbf{g}_u$ carries its defining feature or a close proxy for it, and out-of-band otherwise. The GoEmotions classifier is used as released; the stylometric rates and emotion composites are project-defined calculations. For tempo, reply-delay mean defines the cohort direction and reply length scores generated replies.

Measure	Band	Feature / scorer
Sentiment	out-of-band	GoEmotions polarity quintiles, frozen RoBERTa (Demszky et al., 2020)
Politeness	out-of-band	project-defined profanity-rate tails over a Linguistic Inquiry and Word Count (LIWC) lexical category (Pennebaker et al., 2001)
Self-focus	out-of-band	first-person pronoun-rate tails (Pennebaker et al., 2001)
Curiosity	in-band	question-rate tails
Expressiveness	in-band	project-defined harmonic mean of capitalization and punctuation rates, whose components follow stylometric work (Argamon et al., 2009; Wegmann et al., 2022)
Tempo	in-band	reply-delay mean / reply length
Anxiety	out-of-band	mean of GoEmotions fear, nervousness
Warmth	out-of-band	mean of GoEmotions caring, love
Hostility	out-of-band	mean of GoEmotions anger, disgust, disapproval

in Appendix G, and the reference-based fidelity diagnostics are in Appendix K.

3.6 Persona Stability and Trait Compositionality

Community simulation raises two questions: does each user stay close to the target across repeated turns, and how does the user’s eight-measure pattern change as more trait constraints are added? The Persona Stability Index (PSI) answers the first by measuring how far each reply sits from its target. The Persona Compositionality Index (PCI) answers the second by measuring whether the whole eight-measure pattern still points toward the target as constraints are added. Both indices use eight of the nine measures in §3.5. They exclude sentiment because it defines the cohorts used in these tests. We compute target-driven PCI on a separate constrained run. Its agreement-stratified companion uses the production synthesis pool and groups users by how many measures point in the expected direction. Appendices G, I and J give the measure definitions, derivations, and both PCI analyses.

Before scoring, every generated reply must pass a coherence gate. It must contain at least 3 words and 8 characters. At least 70% of its characters must be ASCII, and at least 40% must be alphabetic. The reply must also avoid template boilerplate. Two repetition checks apply. No three-word sequence may cover more than one-eighth of the text, and no sentence may appear more than twice. Failed replies are excluded, so reported PSI is conditional on passing the gate (Appendix G).

For each of the eight measures we take the absolute gap between a reply’s realized score and that user’s target score. The absolute value treats scores above and below the target as equally far away. We divide that gap by the real-user standard deviation for the same measure. We then average the eight standardized gaps and map the result to $[0, 1]$:

$$\text{match} = \max\left(0, 1 - \frac{1}{2|D^{\text{sig}}|} \sum_{d \in D^{\text{sig}}} \delta_d\right) \quad (1)$$

where D^{sig} is politeness, self-focus, curiosity, expressiveness, tempo, anxiety, warmth, and hostility, and $\delta_d = |\hat{s}_d - s_d^*|/\sigma_d^{\text{real}}$ is the standardized residual. A user’s PSI is the mean of *match* over that user’s coherent replies. Standardizing per dimension puts all eight measures on one scale (Appendix I). We use $\text{PSI} = 0.10$ as an empirical reference for whether generated text carries persona signal. This value was calibrated on the pooled production set rather than preregistered. The set contains 12,451 rows, one per user per generation cell. The reference sits below the PSI band for real-user, in-hull, and near-hull inference and at the far-from-hull collapse (§4.4; Table 14; Appendix I.5). Full derivations are in Appendix I.

PCI compares each synthetic user’s realized and expected patterns across eight measures. It is the mean cosine between those two patterns. For each cohort we define one nested chain of requested traits. The sampler’s filter mode restricts the eligible training authors to those who meet the requested traits, and we synthesize users from that pool. Sentiment is not part of the cosine: it selects the cohort at the first step of each chain but is not itself scored. We evaluate $m \in \{1, 2, 3\}$ requested traits. We do not report $m=4$ because only four rage-chain authors and two empath-chain authors satisfy all four constraints. Appendix J gives both chains, the source and sample counts at every level, the estimates and intervals, and a descriptive companion that groups users by how many measures already point in the expected direction.

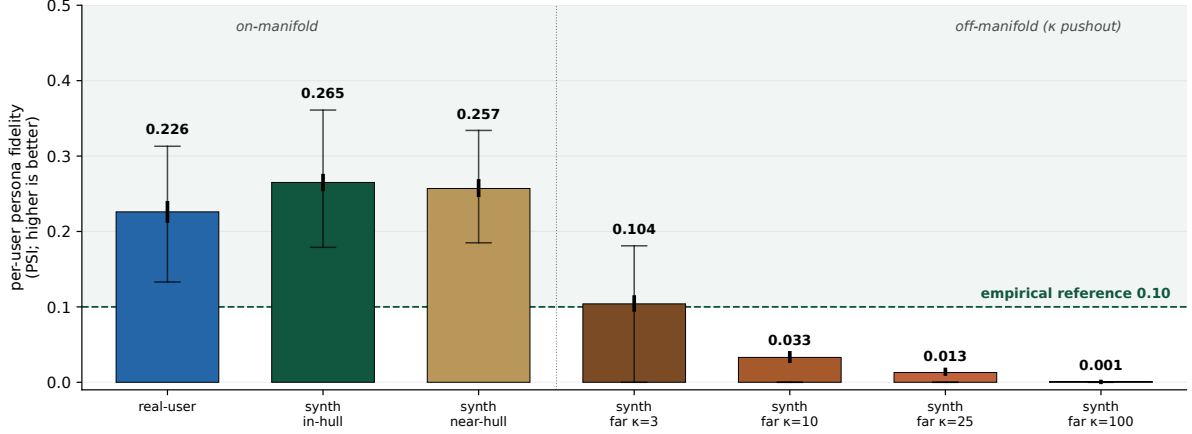


Figure 3: Persona Stability Index (PSI) by inference setting. Higher values mean that generated behavior more closely matches the user’s target profile. Real-user, in-hull, and near-hull PSI ranges from 0.22 to 0.27. The four far-from-hull values drop steeply as κ grows. Bars are across-user means, whiskers the interquartile range, and the thick bracket the 95% bootstrap interval on the mean (Table 14). The dashed green line at 0.10 is an empirical reference that was calibrated on the pooled production set. It lies below the on-manifold band and at the off-manifold collapse; details in Appendix I.5.

Table 2: Rage-vs-empath Hedges’ g for sentiment, politeness, and self-focus, real-user forum. Higher is better; bold marks the largest point estimate on each row.

Measure	HP-LoRA	Vanilla	Control
Sentiment	+0.31	0.00	+0.02
Politeness	+0.19	−0.01	+0.08
Self-focus	+0.32	+0.08	0.00

Compared with the personalization and agent-simulation systems in §2, we sample profiles at several distances from real users. We then score each generated reply against its user’s target. We define PSI and PCI for this corpus, and we have not tested whether either index transfers elsewhere.

We run two stability tests with PSI and its component measures: one tests whether a synthetic user’s persona holds up across a short thread, the other across 80 consecutive generated turns. The short-thread test covers all author pools (Table 14). The 80-turn test covers five settings: a separate 187-author real-user pool (93 rage, 94 empath), in-hull, and far-from-hull at $\kappa \in \{3, 25, 100\}$; Appendix L reports coherence and the fidelity trajectory for all five. Both use temperature 0.4, top-p 0.85 and repetition penalty 1.25. Short threads generate up to 192 new tokens in a 512-token context. The 80-turn run generates up to 96 new tokens inside a 600-token visible-history budget, dropping the oldest tokens first.

4 Results

Both models reach almost identical language-modeling performance: at each one’s best checkpoint, validation perplexity is within one point, with HyperPEFT-LoRA slightly ahead. Full training-budget details are in Appendix B.

4.1 Real-User Baseline

Table 2 shows the rage-versus-empath difference on sentiment, politeness, and self-focus for real users from the training corpus, side by side for HyperPEFT-LoRA, vanilla LoRA, and the control. Sentiment defines the two cohorts. Politeness and self-focus are stylometric measures the hypernetwork never sees: politeness combines standardized profanity and hostility, self-focus a first-person pronoun rate, and neither is in the profile vector. Positive g means the cohorts came out in the predicted direction, negative means backwards, and near-zero means no difference.

HyperPEFT-LoRA separates all three measures in the predicted direction: sentiment +0.31, politeness +0.19, and self-focus +0.32. These values provide the real-user reference for the synthesis tests below. Vanilla LoRA and the control stay within $|g| \leq 0.1$ of zero on all three measures. Neither receives per-user information at inference. Vanilla LoRA has no per-user input, while the control zeroes the profile vector and disables the steering step. Appendix H extends the comparison to the other six measures, including the three in-band

Table 3: Selected example replies by setup and cohort. The rows are illustrative rather than randomly sampled, so Table 4 and Figure 3 carry the quantitative persona claims. Bold marks the cohort-coded wording in each non-Vanilla cell. Appendix O gives the untruncated replies.

Setup	Rage cohort reply	Empath cohort reply
Vanilla LoRA	“the OP has some valid points. . . he should be allowed to do what he wants.”	“... worked out great! We became friends because he knew all about me.”
<i>HyperPEFT-LoRA</i>		
Real-user	“ a pretty awful idea. . . why should she/he be rewarded?”	“ It was really nice working with your friend who helped us out so much! ”
In-hull	“ it just pisses me off when artists get all high and mighty but then never make any effort.”	“ You are so sweet, kind and beautiful! Thank you for staying until midnight.”
Near-hull	“ has no respect for anyone else. . . this isn’t art.”	“ I love working for my company. . . we both understand each other when needed.”
Far-from-hull ($\kappa=3$)	“ bullshit stories. . . when an artist makes such horrible stuff. ”	“ shouldnt feel guilty if you need to sleep early.”

checks. All eight measures keep the predicted sign after reply length is controlled for and after pooling within the 36 setting-by-prompt threads. The control and HyperPEFT-LoRA generate replies to the same thread contexts, yet the control’s Table 2 values stay within 0.1 of zero. Shared thread content therefore does not explain the three separations.

With language modeling matched between the two models and the real-user reference established, we now test the three hypotheses from §1:

- **H1.** Do synthetic users behave on held-out measures the way their profiles imply?
- **H2.** Do they behave as consistently as the model does on real users from the training corpus?
- **H3.** Do they stay consistent over many turns of generation?

4.2 H1: Behavioral Structure on Held-Out Measures

The rage-vs-empath effect size for sentiment, politeness, and self-focus, with 95% bootstrap CIs for every inference setting, is shown in Table 4. Positive values indicate separation in the direction the cohort split predicts, while values near zero indicate little group difference. A nominal 95% interval that excludes zero marks a nonzero contrast under that setting.

Sentiment, the cohort-defining axis, separates most clearly under synthesis. Synthesis reads +0.40 or higher (+0.51 in-hull and near-hull) and every synthesis interval excludes zero, against +0.31 at real-user inference. Politeness remains positive (+0.19 real-user, +0.11 to +0.18 synthesis), but every synthesis interval touches or includes zero. Self-focus is +0.32 at real-user inference, reverses to −0.31 and −0.49 in-hull and

near-hull, and sits near zero far-from-hull at −0.06. H1 is therefore supported on sentiment, directionally positive but uncertain on politeness, and not supported on self-focus under interpolation. Inside the hull the reversal stays on the ratio measures (Appendix H). Vanilla LoRA and the control stay within 0.1 of zero on every row, so a HyperPEFT-LoRA setting with a nonzero profile vector holds the largest estimate on all three measures.

A binary cohort-classification companion to this table, showing the fraction of replies the GoEmotions scorer assigns to each user’s intended cohort, is in Appendix H. Example replies for five of the six setups, all but the control, are in Table 3.

4.3 H2: Synthesized vs. Real

H2 compares synthetic users with authors seen during training but evaluated on threads kept out of training. For rage and empath alike, real-user, in-hull, and near-hull PSI fall in one observed band from 0.20 to 0.28 (Figure 7). On empath, in-hull and near-hull sit at or slightly above real-user inference. Far-from-hull PSI falls as κ grows. This is a descriptive comparison, not an equivalence test. Figure 3 pools the same statistic across cohorts, and Appendix K gives the measure-level decomposition.

4.4 H3: Stability Across Turns

H3 asks whether each user’s generated behavior stays close to their target profile across the turns of a conversation, measured by PSI (§3.6). PSI scores each reply only against that user’s own target, so a low value says the text is off target without saying what it moved toward. Real-user, in-hull, and near-hull inference cluster between 0.226 and 0.265. As κ grows, PSI drops to 0.104, 0.033, 0.013, and 0.001 at $\kappa = 3, 10, 25$, and 100 (Figure 3). The

Table 4: Rage-vs-empath Hedges’ g for sentiment, politeness, and self-focus with 95% bootstrap CIs (1,000 iterations, seed 142). Far-from-hull denotes $\kappa=3$. Resampling is per author, except the three synthesis cells of the Sentiment row, which are pooled over replies. Higher is better; bold marks the largest point estimate on each row, and both cells are bolded on a tie. Appendix H reports all nine tracked measures.

Measure	Vanilla	Control	Real-user	In-hull	Near-hull	Far-from-hull
Sentiment	0.00	+0.02	+0.31	+0.51	+0.51	+0.40
95% CI	[−0.16, +0.14]	[−0.15, +0.20]	[+0.12, +0.49]	[+0.40, +0.63]	[+0.40, +0.63]	[+0.29, +0.51]
Politeness	−0.01	+0.08	+0.19	+0.18	+0.11	+0.13
95% CI	[−0.15, +0.14]	[−0.11, +0.25]	[+0.02, +0.36]	[−0.01, +0.39]	[−0.16, +0.43]	[−0.00, +0.28]
Self-focus	+0.08	0.00	+0.32	−0.31	−0.49	−0.06
95% CI	[−0.06, +0.22]	[−0.20, +0.19]	[+0.13, +0.49]	[−0.49, −0.14]	[−0.72, −0.26]	[−0.20, +0.07]

empirical 0.10 threshold (Appendix I.5) sits between $\kappa=3$ and $\kappa=10$, where the persona signal weakens.

Two further checks support the stability result. Under the in-hull short-thread test, every reported per-dimension slope magnitude stays below 0.05 per turn (Appendix I.6). Across all five 80-turn settings the largest magnitude is about 0.0095. On the separate 187-author real-user pool the coherence-gate pass rate falls only from about 0.97 at turn 1 to about 0.84 at turn 80 (Appendix L).

Trait compositionality. We report two PCI analyses, and only the agreement-stratified analysis is part of H3. The target-driven analysis is descriptive and follows two fixed chains that add constraints one at a time. The chains move in opposite directions. Rage falls from -0.021 at $m=1$ to -0.411 at $m=3$, while empath rises from $+0.240$ to $+0.505$. The pooled values ($+0.109$, $+0.102$, -0.015) reflect the larger rage decline and the changing sample counts for the two chains (Appendix J, Table 15). The agreement-stratified analysis uses the production pool and groups users by how many measures already point in the expected direction. Its mean eight-measure cosine rises from -0.57 when no measure matches to $+0.48$ when at least four match (Appendix I.4, Table 16).

5 Conclusion

A hypernetwork trained to rebuild known users can also synthesize new ones within the profile range covered by training. We show this through three findings:

- **Synthetic users stay consistent across the turns of a conversation.** In short-thread runs the Persona Stability Index sits above the empirical 0.10 threshold for real-user, in-hull, and near-hull inference (§4.4, Figure 3).

Across an 80-turn simulation the fidelity trajectories stay flat. The largest drift slope, 0.0095 per turn, is 19% of the pre-specified $|\beta_d| < 0.05$ threshold.

- **Synthetic users carry clear cohort signal beyond the control.** On the cohort-defining sentiment axis, every synthesis setting reads $g \geq +0.40$ against control values within 0.1 of zero (§4.2, Table 4). The nine-measure decomposition (Appendix H) keeps warmth and hostility positive across synthesis, while anxiety reverses only at far-from-hull. Politeness stays positive, and self-focus and expressiveness reverse under interpolation.
- **Synthetic users built from the real-user range track real-user target match closely.** For rage and empath alike, in-hull and near-hull synthesis land in the same per-cohort PSI band as real-user inference, 0.20 to 0.28 (§4.3, Appendix I.6). Empath in-hull and near-hull sit at or above real-user inference. Pushed far enough outside that range, the match falls steadily.

Stability does not imply one common compositionality trend: PCI falls for rage and rises for empath (Appendix J). One shared hypernetwork serves a population on demand without a per-user adapter checkpoint or per-persona retraining, so long as new users come from the profile space covered by training. Each user needs only an 18-number profile. For 10,000 authors, the shared personalization state uses 1.28 GB instead of 755 GB for one FP32 LoRA adapter per author, a $588\times$ reduction (Appendix M). Outside that space, persona fidelity declines steadily (Figure 3; Table 14).

Limitations

The findings cover a single hypernetwork, backbone, corpus, genre, and language: we test HyperPEFT-LoRA on Pythia-1.4B, train it on English Reddit discussions, and evaluate open-ended replies. Generalization of these results to other backbones, platforms, languages, and interaction formats is yet untested. One design choice does carry across model families: we ran the profile-size sweep of Appendix E on a Llama-3.2-1B backbone, and the 18 coordinates it selected transfer unchanged to the Pythia-1.4B system reported here. All persona validation is computational rather than human-rated; without a perception study, separation on the automatic measures does not prove that people would perceive each synthetic user as a stable individual voice. Appendix G documents the scorers.

The cohort labels are defined on the sentiment axis, so the corpus-level split on other measures is weaker than the split on sentiment itself. Two rage-cohort users can differ substantially in reply length, question rate, and capitalization. On the training-set real-user reference every measure has the predicted sign (Table 9), but the magnitudes are bounded by what the cohort labels predict at the population level. Rare joint combinations such as “empath and hostile at the same time” are likewise too sparse in the corpus to estimate that joint cell tightly. Target-driven compositionality has the same limits. It covers two fixed nested chains and changes the eligible source population at every level. It cannot estimate $m=4$, and its intervals are conditional on that fixed pool. Its signature cosine is unstandardized, so the z-scored coordinates dominate it (Appendices I and J). The result describes these two chains in this measure space, not arbitrary trait composition.

The hypernetwork’s per-user offset gets larger the further outside the real-user range we push the synthetic user. We cap the per-element size of that offset to keep the generated text from breaking down. That cap already binds most entries on real and interpolated profiles: the QKV saturation rate is 92% on real users, 87% in-hull, and 85% near-hull, rising to 98% under far-from-hull synthesis (Appendix N). Per-user variation therefore remains in the sign pattern and in the unsaturated entries, and pushing the profile further off manifold reduces the unsaturated share. The example replies in Table 3 show what this looks like in the generated

text.

We evaluate one HyperPEFT-LoRA configuration. The hypernetwork emits both LoRA offsets, and the model applies ΔB_j through the frozen A_j factor. This gives a low-rank update with $r = 24$. An elementwise clamp bounds every emitted value. Future work can apply both offsets and include the offset-size term in parameter updates. The simulation covers open-ended replies only. It does not model deeper thread structure, rate limits, moderation, or session-level state, so the results do not establish community-level realism.

This paper makes a deliberately narrow claim: we do not introduce a new architecture or a general theory of persona generation. The hypernetwork that realizes each persona is established prior work (Walsh et al., 2026; Ha et al., 2017); our contribution is synthesizing unseen users by sampling in profile space and measuring where that holds. The synthetic users reflect the training-author pool used to draw their anchors; sparsity or skew in that pool also appears in the sampled mixtures.

Ethical Considerations

A model that writes in a specific person’s style can support impersonation or believable fake participants. The released profiles carry no names or factual biographies, and synthetic users are sampled as new profiles rather than aimed at a named account. Those safeguards reduce direct targeting but do not remove stylometric privacy risk: compact writing-style vectors may still support linkage or re-identification when combined with outside text. We measure one form of that risk with a nearest-neighbor linkage test reported in Appendix C and released with the artifacts. The release contains the code, the per-user profile vectors, the cohort labels, and the thread-to-split assignments. The data artifacts carry integer identifiers only: no account names, no raw comments, and no weights tied to a named user.

The training corpus is drawn from public online discussion threads. Accounts are pseudonymous rather than anonymous. Any released deployment configuration of the synthesis engine is bounded to the in-hull and near-hull samplers, which stay inside the real-user range. The far-from-hull κ -pushout setting is retained only as an off-manifold diagnostic inside the paper, not as an operational generation interface.

The intended use is measurement and controlled

simulation, not the operation of fabricated communities. The same synthesis mechanism could nevertheless be used to manufacture apparent public participation, so deployment requires provenance, rate limits, and monitoring outside the scope of this paper. The automatic scorers were developed for English-language data, so their validity across languages and cultural settings is untested.

Acknowledgments

The authors acknowledge William & Mary HPC Team for providing computational resources and technical support that have contributed to the results reported within this paper. URL: <https://www.wm.edu/it/rc>

References

- Andy Ardit, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37.
- Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The Pushshift Reddit dataset](#). In *Proceedings of the 14th International AAAI Conference on Web and Social Media (ICWSM 2020)*, volume 14, pages 830–839, Palo Alto, CA, USA. AAAI Press.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Halahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.
- Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. 2025. [Text-to-LoRA: Instant transformer adaption](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 7485–7514. PMLR.
- Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. [A computational approach to politeness with application to social factors](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.
- Mathijs Deen and Mark de Rooij. 2020. [ClusterBootstrap: An R package for the analysis of hierarchical data using generalized linear models with the cluster bootstrap](#). *Behavior Research Methods*, 52:572–590.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.
- Bradley Efron. 1979. [Bootstrap methods: Another look at the jackknife](#). *The Annals of Statistics*, 7(1):1–26.
- Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7828–7838. Association for Computational Linguistics.
- David Ha, Andrew Dai, and Quoc V. Le. 2017. HyperNetworks. In *International Conference on Learning Representations (ICLR)*. arXiv:1609.09106.
- Larry V. Hedges. 1981. [Distribution theory for Glass’s estimator of effect size and related estimators](#). *Journal of Educational Statistics*, 6(2):107–128.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799, Long Beach, California, USA. PMLR.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Web and Social Media (ICWSM 2014)*, pages 216–225.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. [PersonaLLM: Investigating the ability of large language models to express personality traits](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.
- Li Lucy and David Bamman. 2021. [Characterizing English variation across social media communities with BERT](#). *Transactions of the Association for Computational Linguistics*, 9:538–556.

- Chuancheng Lv, Lei Li, Shitou Zhang, Gang Chen, Fan-chao Qi, Ningyu Zhang, and Hai-Tao Zheng. 2024. [HyperLoRA: Efficient cross-task generalization via constrained low-rank adapters generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16376–16393, Miami, Florida, USA. Association for Computational Linguistics.
- Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1155.
- Goran Murić, Alexey Tregubov, Jim Blythe, Andrés Abeliuk, Divya Choudhary, Kristina Lerman, and Emilio Ferrara. 2022. [Large-scale agent-based simulations of online social networks](#). *Autonomous Agents and Multi-Agent Systems*, 36(2):38.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simula-cra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST 2023)*, pages 2:1–2:22.
- Seong-Il Park and Jay-Yoon Lee. 2024. [Toward robust RALMs: Revealing the impact of imperfect retrieval on retrieval-augmented language models](#). *Transactions of the Association for Computational Linguistics*, 12:1686–1702.
- Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. [Learning interpretable style embeddings via prompting LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.
- James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC)*. Lawrence Erlbaum Associates.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. [LaMP: When large language models meet personalization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.
- Jitkapat Sawatphol, Can Udomcharoenchaikit, and Sarana Nutanong. 2024. [Addressing topic leakage in cross-topic evaluation for authorship verification](#). *Transactions of the Association for Computational Linguistics*, 12:1363–1377.
- Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024a. [Personalized Pieces: Efficient personalized large language models through collaborative efforts](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6459–6475, Miami, Florida, USA. Association for Computational Linguistics.
- Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024b. [Democratizing large language models via personalized parameter-efficient fine-tuning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491, Miami, Florida, USA. Association for Computational Linguistics.
- Zhaoxuan Tan, Zixuan Zhang, Haoyang Wen, Zheng Li, Rongzhi Zhang, Pei Chen, Fengran Mo, Zheyuan Liu, Qingkai Zeng, Qingyu Yin, and Meng Jiang. 2026. [Instant personalized large language model adaptation via hypernetwork](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23557–23580, San Diego, California, United States. Association for Computational Linguistics.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. [Two tales of persona in LLMs: A survey of role-playing and personalization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.
- Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. [Hyper-X: A unified hypernetwork for multi-task multilingual transfer](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7934–7949, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tyson H. Walsh, Trenton W. Ford, and Daniel Runfola. 2026. [HyperPEFTNet: Parameter-efficient hypernetworks for persona synthesis](#). *Research Square*. Preprint.
- Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025. [Know you first and be you better: Modeling human-like user simulators via implicit profiles](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21082–21107, Vienna, Austria. Association for Computational Linguistics.
- Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhao Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. [RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.
- Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. [Same author or just same topic? towards content-independent style representations](#). In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

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

A Architecture Comparison and Detailed Diagram

Figure 4 gives a higher-resolution view of the architecture summarized in main-paper Figure 2: the frozen backbone (Panel A), one representative LoRA module (Panel B), and the hypernetwork (Panel C). Table 5 places this paper’s system next to the earlier Pythia-160M instantiation (Walsh et al., 2026). Numbers across the two columns are not directly comparable: the backbone capacity and the adapter-attachment scheme differ. For the Pythia-1.4B system, we initialize A_j from a Kaiming-uniform distribution and set B_j to zero, following PEFT’s default LoRA initialization. We do not copy either factor from the vanilla LoRA solution, and neither changes during training.

Table 5: Architecture comparison between HyperPEFT-LoRA instantiations. This paper’s system is the Pythia-1.4B column; the prior instantiation is from Walsh et al. (2026).

	Prior (Pythia-160M)	This paper (Pythia-1.4B)
Backbone params	164.1M (frozen)	1.4B (frozen)
Decoder blocks	12	24
Hidden dim	768	2048
Attention	12-head MHA	16-head MHA
Instruction / RLHF tuning	none (Pile-pretrained base)	none (Pile-pretrained base)
LoRA sites	36 (QKV $\times$ 12)	96 (QKV+O+MLP _{up,down} $\times$ 24)
LoRA rank r	32	24
LoRA α	32	48
Offsets emitted	A and B	A and B
Profile K	6	18
Loss for updates	CE	CE

B Training Specifications

Table 6 lists both models’ architecture and training budgets. They share the frozen Pythia-1.4B backbone, LoRA-site geometry, precision, and optimizer schedule. Vanilla LoRA fine-tunes its LoRA factors. HyperPEFT-LoRA keeps those factors fixed and trains a hypernetwork that reads a profile vector. It also clamps the emitted offsets and applies steering at inference. The table summarizes these differences.

The validation cross-entropy gap at the best checkpoint is -0.029 in HyperPEFT-LoRA’s favor, equivalent to a 0.6-point PPL gap. Both models reach this fidelity from the same data and the same

Table 6: Training specifications for both models.

	Vanilla LoRA	HyperPEFT-LoRA
Backbone	Pythia-1.4B (frozen)	Pythia-1.4B (frozen)
LoRA sites	96	96
LoRA rank r	24	24
LoRA α	48	48
LoRA matrices	fine-tuned	not fine-tuned
Hypernetwork	—	one-layer SiLU projection, width 256
Heads	—	3 factorized (QKV, O, MLP), rank 16
Profile K	—	18
Loss for updates	CE	CE
Clamp per element	—	± 0.020
Precision	BF16	BF16
Optimizer	AdamW	AdamW
Best ckpt step	600	2,400
Val CE @ best	3.100	3.071
Val PPL @ best	22.2	21.6
Inference steering	—	orth. direction-ablation; $\alpha=1.0$; layers 15 to 23
Seed	142	142

optimizer, so the persona effects reported in §4 are not confounded by a training-budget gap.

C Artifacts, Licenses, and Reproducibility

Models and code. We use the released EleutherAI Pythia-1.4B checkpoint and keep it frozen (Biderman et al., 2023); the checkpoint’s model card identifies the release and its Apache 2.0 license. The feature-selection sweep of Appendix E additionally used the released Llama-3.2-1B checkpoint under the Llama 3.2 Community License; no result reported in this paper runs on the Llama checkpoint, and no weights derived from it are released. The trained hypernetwork and all of the training, generation, and scoring code are our own. We release, under the MIT License, the code, the per-user profile vectors, and the cohort labels at https://github.com/tyson-walsh/HyperPEFTNet_Personas_On_Demand. The train and evaluation partition is released as project-internal thread keys and split assignments rather than raw text. The keys index our own build of the corpus, so they cannot be joined to Reddit or used to reconstruct the source text; anyone rebuilding from the public Pushshift archive can adopt the same train, validation, and test proportions without receiving the forum text from us. We do not release the raw forum text, account names, or any model weights tied to a named person.

Scoring software. We use one released emotion classifier, one released SST-2 classifier, and the VADER rule-based sentiment scorer. The emotion model is SamLowe/roberta-base-go_emotions,

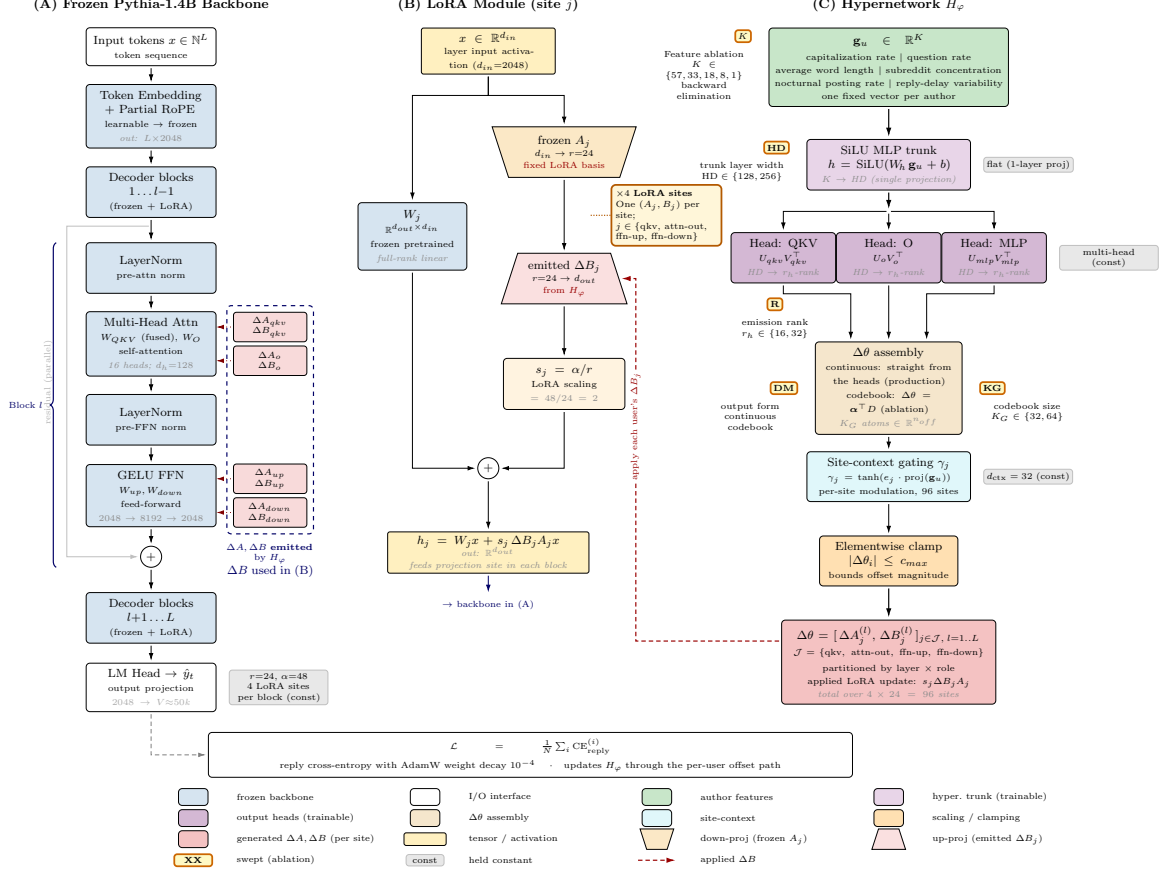


Figure 4: Component-level rendering of the Pythia-1.4B HyperPEFT-LoRA architecture (companion to main-paper Figure 2). (A) The frozen Pythia-1.4B decoder stack with residual block l expanded; LoRA is attached at four sites per block (qkv, attn-out, ffn-up, ffn-down). (B) One LoRA module. The frozen A_j first projects x . The hypernetwork’s ΔB_j then maps that projection to the output dimension. We add the scaled result, $s_j \Delta B_j A_j x$, to the frozen layer output $W_j x$. (C) The hypernetwork consumes the profile vector g_u , projects through a SiLU trunk into a shared hidden dimension, factors through three role-specific output heads, applies site-context gating, and clamps elementwise to produce $\Delta\theta$. Sweep markers in orange flag ablation axes; (const) tags identify hyperparameters held fixed.

a RoBERTa-base checkpoint fine-tuned on the GoEmotions dataset; Demszky et al. (2020) supplies the dataset citation, while the model identifier names the deployed checkpoint. The emotion checkpoint is released under the MIT license and the SST-2 checkpoint under Apache 2.0. The SST-2 cross-check uses distilbert-base-uncased-finetuned-sst-2-english, and the other sentiment cross-check uses the vaderSentiment implementation of VADER (Hutto and Gilbert, 2014), released under the MIT license. The per-user fidelity analysis of Appendix K additionally uses the sentence-transformer encoder sentence-transformers/all-MiniLM-L6-v2, released under Apache 2.0. All four run at default settings without retraining. The politeness and

self-focus features are computed by our own lexicon code in the tradition of Pennebaker et al. (2001); we do not use the proprietary LIWC program.

Intended use. We use every artifact for research only, consistent with the use its original authors intended. Our hypernetwork and the synthetic profiles we build exist to measure how faithful persona synthesis is and where it fails, not to create or operate fabricated participants.

Data and privacy. The profile vectors are computed from the public online-forum corpus described in §3.1. The corpus was built from the 2010 through 2016 Pushshift Reddit archive described by Baumgartner et al. (2020). We collect

no new data and we run no new human annotation. Usernames are replaced with SHA-256 hashes before processing, all URLs are replaced with a fixed placeholder token, and posts marked deleted or removed are mapped to a single placeholder sentence. Each per-user vector is 18 numbers covering writing style, discourse habits, platform use, and posting time, such as average word length, question rate, and posting-hour entropy, and these numbers do not record personal facts about the author. Before release we check the artifacts so that they hold no names, no account handles, and no raw post text. Forum text can contain offensive language; we score it during evaluation but we do not redistribute it.

Linkage test. In this test, an adversary recomputes the 18 features from the same public archive with Gaussian error on each dimension. Re-identification reaches 100% of released rows at an error of 0.10 population standard deviations, 99.8% at 0.25, and 77.1% at 0.50; the median nearest-neighbor distance is 1.86 in standardized units. Running the attack requires already holding the source posts, so it recovers nothing a reader could not compute directly, and it does not cover attacks that begin from raw text or from outside personal information.

Compute. We ran every experiment on a shared multi-GPU cluster using two accelerator classes: a high-memory class and a 48 GB class. Training the hypernetwork ran on a single high-memory GPU for 3,331 steps; the hypernetwork and the vanilla LoRA baseline together used about 60 GPU-hours. The full generation and scoring pipeline, including the 80-turn stability run, used about 50 GPU-hours. For reference, the 80-turn stability run by itself took about a day and a half on one high-memory node.

D The Synthesis Engine: Visual Walkthrough

Figure 5 gives a 2D schematic of the interpolation sampler. Real authors occupy three shaded regions: the rage tail, the empath tail, and the rest of the pool, from which anchors may also be drawn. To create an unseen user, the engine draws k authors uniformly from the eligible training pool, draws a Dirichlet weight vector over those anchors, and forms their convex combination. Production uses $k=5$ with $\eta=2.0$ for in-hull synthesis and $k=12$

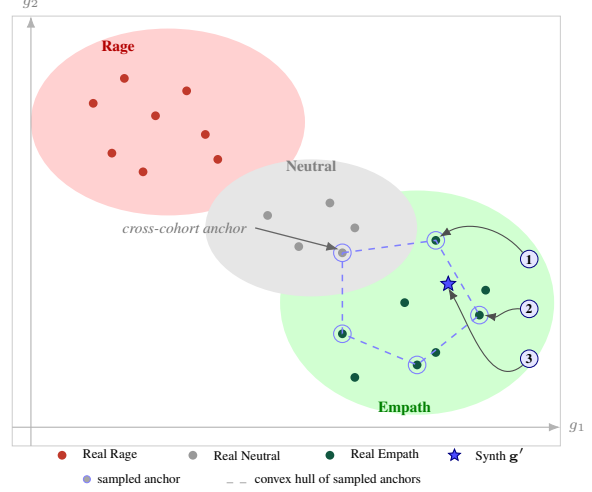


Figure 5: The interpolation sampler on a 2D projection of profile space; the diagram is schematic. The area labeled Neutral holds authors outside the rage and empath tails. It is the rest of the eligible pool, not a third cohort. (1) Draw k anchors uniformly from the eligible training pool. (2) Draw a Dirichlet weight vector over those anchors. (3) Form the convex combination g' (blue star), which stays inside the convex hull of the drawn anchors and therefore inside the convex hull of the training authors. Production uses $k=5$ with $\eta=2.0$ for in-hull synthesis and $k=12$ with $\eta=1.0$ for near-hull synthesis. The trained hypernetwork then converts g' into a per-user adapter offset, just as it would for a real author.

with $\eta=1.0$ for near-hull synthesis. Anchors are drawn uniformly and no jitter is added; nearest neighbors are used only by the density gate. The synthesized profile vector is then passed to the trained hypernetwork, which converts it into a per-user adapter offset exactly as it would for a real user.

Figure 6 confirms the picture in numbers. We measure how far each user sits from the real-user mean using Mahalanobis distance, which is a multi-feature distance that scales each axis by the real-user covariance: 1 unit means one real-user standard deviation along that direction, after accounting for correlations among features. The real-user 99th percentile is 9.37, with most values between 1 and 9. In-hull and near-hull synthetic users sit inside that envelope at median distances of 1.4 to 1.9. Far-from-hull users, drawn by the off-manifold pushout sampler at $\kappa=3$, sit at median ≈ 15 and 99th percentile ≈ 47 . Both interpolation settings therefore land inside the real-user envelope and the pushout setting lands far outside it.

Synthesis engine: how far each stratum lands from the real-user profile centroid (10,000 real, 9,872 synthetic)

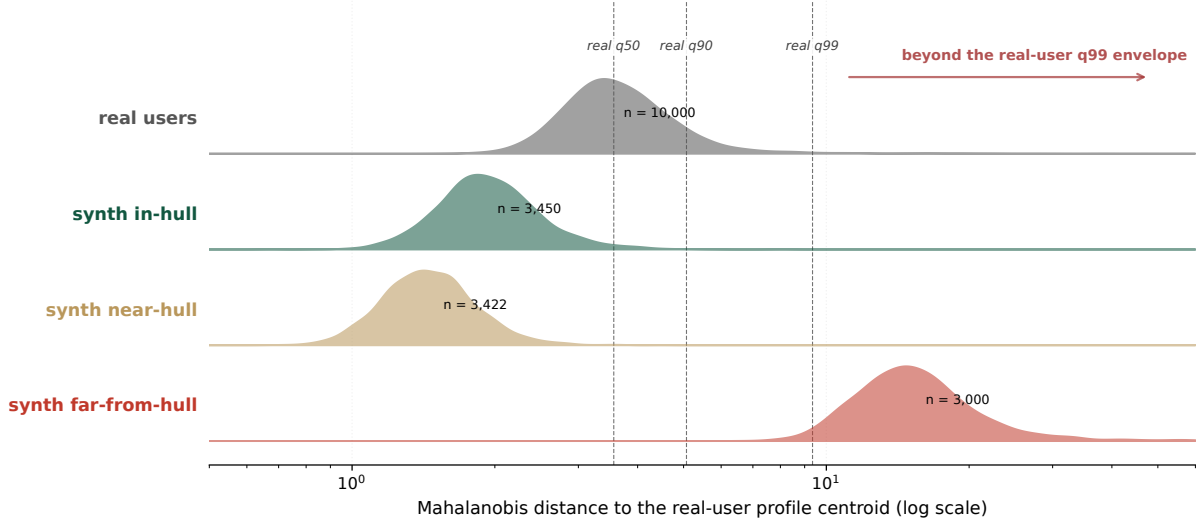


Figure 6: Mahalanobis-distance distributions per synthesis setting (log x). Real users (gray) form a band around $q_{50}=3.57$. In-hull (green) and near-hull (gold) synth users sit at lower distances inside the real-user envelope. Both are convex combinations of real profiles, and averaging reduces spread, so both sit at lower median distances from the centroid than the real users do. Near-hull has the lower median of the two, 1.445 against 1.912. Near-hull and in-hull differ in anchor count, Dirichlet concentration, and density gate. The lower near-hull median therefore reflects the full configuration rather than any one choice (§3.4). Far-from-hull (red), drawn at $\kappa=3$, has $q_{50}=15.0$, above the real-user 99th percentile. Dashed vertical lines mark the real-user q_{50} , q_{90} , and q_{99} . The counts printed in the figure refer to the sampler pools for this check, not the evaluation populations in Table 14.

E Feature Definitions and Backward Elimination

Table 7 lists the 18 profile features in four categories: style (surface writing regularities), discourse (thread-level linguistic habits), platform (usage patterns), and temporal (posting rhythm). The stylometric grounding follows prior authorship and psycholinguistic work (Pennebaker et al., 2001; Argamon et al., 2009). Prior user-modeling work motivates using user-level factors (Flek, 2020; Lynn et al., 2017). We define the platform and temporal features for this corpus.

Backward elimination. We started from a pool of 57 candidate features and pruned iteratively, scoring each candidate set on two signals: (a) permutation importance against validation cross-entropy and (b) post-hoc delta separation in the hypernetwork’s output space. The trajectory is uneven. Across the five K levels, validation cross-entropy spans 0.034, while delta separation varies by a factor of 6,200. Cross-entropy is therefore a poor instrument for persona-relevant feature pruning. $K=18$ gives us the largest δ_{sep} among the five sizes we tested (Table 8). We ran this sweep

Table 7: The 18 profile features. Sentiment features are excluded by design.

#	Feature	Category	Signal
1	mean post length	Discourse	Post length / verbosity
2	subreddit concentration	Platform	Top-subreddit share of activity
3	question rate	Discourse	Interrogative style
4	capitalization rate	Style	Emphasis / shouting
5	readability (Flesch-Kincaid)	Style	Reading-grade level
6	external-link rate	Platform	External references
7	negation rate	Discourse	Disagreement / denial
8	subjectivity rate	Discourse	Opinion vs. fact
9	emoticon rate	Style	Informal expression
10	contraction rate	Style	Formality level
11	average word length	Style	Vocabulary sophistication
12	long-word rate	Style	Complex word usage
13	stopword rate	Style	Content density
14	repeated-bigram rate	Style	Repetitive phrasing
15	posting-hour entropy	Temporal	Posting schedule uniformity
16	nocturnal posting rate	Temporal	Night-owl behavior
17	mean posting hour	Temporal	Average posting time
18	reply-delay variability	Temporal	Response timing variability

earlier, separately on a Llama-3.2-1B backbone, at 50 optimizer steps per level over a 16,000-row training and 4,000-row validation subsample of the same corpus. The 18 coordinates selected are the

Table 8: Backward elimination across five K levels, run on a Llama-3.2-1B backbone at a reduced budget; separation peaks at $K=18$. The cross-entropy column is not comparable with Table 6. Bold marks the selected row, not the best value in each column.

K	val CE	δ_{sep}	Rel.	Note
57	3.140	0.00636	$1.0\times$	full candidate set
33	3.141	0.00324	$0.51\times$	intermediate
18	3.148	0.0491	$7.7\times$	selected
8	3.131	0.0074	$1.2\times$	smaller
1	3.165	$7.9\text{e-}6$	$0.001\times$	minimal

ones the Pythia-1.4B model reported here uses, unchanged and in the same order, so the profile design is not tuned to the evaluated backbone. Because the backbone, tokenizer, and step count all differ, the cross-entropy column of Table 8 resides on a different scale from the Pythia-1.4B values in Appendix B and the two should not be read against each other; what carries over between the two is the ordering of δ_{sep} across K . Separation is measured between the same rage and empath cohorts that the main study uses, on training-split authors with every user in that run’s validation split excluded. The sweep does not establish a global optimum or performance on other corpora.

F Inference-Time Persona Steering: Operating Point

Our intervention uses the residual-stream direction-ablation form introduced for refusal behavior by Ardit et al. (2024).

How the directions are estimated. For the same authors and the same 20 fixed prompts, we run one forward pass with the hypernetwork offsets applied and one with them zeroed, taking the mean last-token hidden state per decoder layer in each. The population direction is the zeroed mean minus the applied mean, estimated over authors from the top and bottom tails of all nine measures and rescaled to unit length. The same contrast inside a single sentiment cohort, sign-flipped, gives that cohort’s persona direction. Extraction sampled 1,325 authors from those tails and the sentiment cohorts, at seed 142. At inference we project the population direction orthogonally to the user’s cohort persona direction, rescale the remainder to unit length, and remove the component of the residual stream that lies along it. A user with no cohort persona direction gets the population direction unprojected. We apply the project-specific setting $\alpha = 1.0$ over

decoder layers 15 to 23. At this operating point HyperPEFT-LoRA produces 99.0% coherent real-user replies, against 100% for vanilla LoRA and 91.8% for the control ($n = 1,344$ to $1,345$ per setting), and the rage-vs-empath separation on the sentiment axis exceeds both baselines (Table 2). The full α sweep is reserved for the layer-resolved mechanism analysis in a follow-on paper.

G Behavioral Measure Definitions

Band assignment. A measure is in-band when $\mathbf{g}_u$ carries its defining feature or a close proxy for it. Three qualify. Curiosity is scored on question rate, which is a profile coordinate. Expressiveness is the harmonic mean of capitalization and punctuation rates; capitalization is a coordinate, and it correlates with the full measure at $r=0.88$ across the 10,000 training authors. Tempo’s cohort is cut on reply-delay mean; the coordinate is reply-delay variability, which correlates with the mean at $r=0.83$. Its realized score is reply length (§3.5), which the mean-post-length coordinate tracks directly. We treat these three as conservative positive controls because the profile vector either carries each score or tracks it closely. The other six measures are out-of-band: their defining features are GoEmotions class probabilities, a profanity rate, or a first-person pronoun rate, and none of those is a coordinate. The closest coordinate has $r=-0.37$ with the profanity rate and $r=-0.59$ with the first-person rate. An out-of-band effect cannot be explained by the hypernetwork reading the label off its own input.

Label construction. Binary cohorts come from the tails of each measure’s distribution. They are fixed-size rank cuts, so ties do not shift the counts. Every measure except sentiment uses the bottom and top deciles of its defining feature, which is exactly 1,000 users per cohort at our population size. Sentiment uses the bottom and top quintiles of the GoEmotions polarity composite, giving 2,000 per cohort. Labels are computed on the training split only and held fixed for every downstream evaluation.

Supplementary emotion measures. The multi-turn analysis uses the three emotion composites of Table 1, each a mean over GoEmotions classes: *anxiety* over *fear* and *nervousness*, *warmth* over *caring* and *love*, and *hostility* over *anger*, *disgust*, and *disapproval*. Anxiety, warmth, and hostility all

use classes that also sit in the GoEmotions sentiment composite, and hostility is half of realized politeness. We dropped *gratitude* from reply-scored warmth because sub-class analysis on the 10K training corpus showed it generates topic-induced false positives on emphatic phrasing (e.g., “thanks for being honest about X” scored by the GoEmotions classifier as a high-gratitude reply on a rage-topic prompt). The released synthetic targets still include gratitude. Recomputing them with the two-class composite moves production PSI by at most 0.0061, every signature-cosine cell by under 0.001, and the Table 18 warmth correlations by up to 0.05; no threshold or verdict changes. The two-class composites preserve source-data direction (empath higher on both warmth and anxiety; $1.62\times$ on warmth, $1.11\times$ on anxiety in the 10K corpus).

Scoring pipeline. Each generated reply passes through the coherence gate, the GoEmotions RoBERTa scorer for emotion composites, and deterministic feature extractors for stylometric dimensions. The gate drops a reply unless it has at least 3 words and 8 characters, at least 40% alphabetic and 70% ASCII characters, no template boilerplate, and no degenerate repetition. The boilerplate check guards against patent- or legal-style phrasing that can surface from the frozen backbone’s Pile pre-training; the repetition check requires that the most frequent word trigram cover at most one-eighth of the reply and that no sentence appear more than twice. Failed replies are dropped before cohort aggregation, and the per-user score on each measure is the average over that user’s surviving replies. Cohort-level effect sizes come from those per-user averages, so the degrees of freedom reported in Hedges’ g reflect the user population, not the per-reply count. The three synthesis cells of the Sentiment row in Tables 4 and 9 pool replies instead.

H Per-Measure H1 Checks

Table 9 extends the body results for sentiment, politeness, and self-focus to all nine measures across the six inference settings.

On real-user inference, HyperPEFT-LoRA carries all nine measures in the correct direction, and the two largest are the in-band controls *expressiveness* (+1.06) and *tempo* (+1.04), whose defining features the profile vector tracks closely. Warmth carries the correct sign at real-user inference (+0.44) and at every synthesis setting. Under synthesis, the split follows measure type. Senti-

ment, warmth, and hostility stay correctly signed at every synthesis setting (sentiment +0.51, +0.51, +0.40; warmth +0.51, +0.65, +0.32; hostility +1.13, +1.92, +0.72 across in-hull, near-hull, and far-from-hull). Anxiety stays positive in-hull and near-hull but reverses at far-from-hull, where it is -0.08 . For sentiment, warmth, and hostility the per-user correlations are also higher under synthesis than under real-user inference (Table 18). The ratio-based stylometric measures behave differently: self-focus, expressiveness, and curiosity lose or reverse their cohort sign under synthesis. Self-focus and expressiveness reverse under in-hull and near-hull synthesis, curiosity sits near zero there and turns negative far from the hull, and tempo stays near zero. Real-user and synthetic inference differ only in the profile source, so the reversal is attributable to that channel rather than to the steering step; the reported experiments do not go further and identify a mechanism. Inside the hull the reversals stay on the ratio measures, while the cohort-defining sentiment separation holds throughout.

Cohort-classification rate companion table. Table 10 gives a second view of H1: the fraction of generated replies the GoEmotions scorer assigns to each user’s intended cohort using the top-quintile sentiment match. HyperPEFT-LoRA reaches 0.535 on rage and 0.428 on empath under real-user inference. The three synthesis settings range from 0.37 to 0.39 on rage and from 0.49 to 0.58 on empath; in-hull and near-hull synthesis are above the real-user reference on empath. The control is lowest on rage at 0.311 but reaches 0.531 on empath. That asymmetry makes the continuous Hedges’ g view in Table 9 easier to interpret than the binary rate.

Which measures show the largest cohort differences? In Table 9, the three largest real-user Hedges’ g estimates are *expressiveness* (+1.06), *tempo* (+1.04), and *warmth* (+0.44), above hostility at +0.34, self-focus at +0.32, and sentiment at +0.31. Sentiment, the cohort-defining axis, is therefore not among the largest per-measure differences: the stylistic and emotion channels that co-vary with sentiment separate the cohorts more than the explicit sentiment score does.

Which side differs between HyperPEFT-LoRA and Vanilla LoRA? Table 9 reports the overall cohort gap but not which side HyperPEFT-LoRA moves. We run a one-sided Mann-Whitney U on

Table 9: Per-measure rage-vs-empath Hedges’ g across all nine measures, by inference setting. Values are computed from per-author averages, except the three synthesis cells of the Sentiment row, which are pooled over replies. Higher is better; bold marks the largest point estimate on each row, and both cells are bolded on a tie. Negative values mark direction reversals, while values near zero indicate little separation. Far-from-hull denotes $\kappa=3$.

Measure (band)	Vanilla	Control	Real-user	In-hull	Near-hull	Far-from-hull
Sentiment (OOB)	0.00	+0.02	+0.31	+0.51	+0.51	+0.40
Politeness (OOB)	−0.01	+0.08	+0.19	+0.18	+0.11	+0.13
Self-focus (OOB)	+0.08	0.00	+0.32	−0.31	−0.49	−0.06
Curiosity (IB)	−0.10	−0.04	+0.22	+0.01	+0.01	−0.11
Expressiveness (IB)	+0.12	+0.12	+1.06	−0.15	−0.14	−0.34
Tempo (IB)	+0.03	−0.17	+1.04	+0.04	+0.01	−0.02
Anxiety (OOB)	+0.05	+0.08	+0.22	+0.25	+0.19	−0.08
Warmth (OOB)	−0.04	+0.18	+0.44	+0.51	+0.65	+0.32
Hostility (OOB)	+0.03	+0.15	+0.34	+1.13	+1.92	+0.72

Table 10: Cohort-classification rate companion to Table 4, over real-user inference, in-hull, near-hull, far-from-hull at $\kappa=3$, and the control. Higher is better; bold marks the largest point estimate on each cohort row. Cohort-filtered n varies from 65 (rage near-hull) to 392 (rage far-from-hull) per cell.

Cohort	Real-user	In-hull	Near-hull	Far-from-hull	Control
Rage	0.535	0.381	0.372	0.393	0.311
Empath	0.428	0.564	0.581	0.494	0.531

the per-user realized polar values in two directions: HP-LoRA HIGH-cohort users scoring above the Vanilla HIGH-cohort users (*amplification* of the right side), and HP-LoRA LOW-cohort users scoring below the Vanilla LOW-cohort users (*suppression* of the wrong side). Table 11 reports the resulting side-specific differences. Only the HIGH-side test passes $p < 0.001$ for politeness and expressiveness; only the LOW-side test passes for self-focus, curiosity, anxiety, warmth, and hostility; both pass for tempo. Because this comparison uses Vanilla LoRA, it is a system-level decomposition; it does not isolate the profile channel or establish a mechanism for the difference.

Linear cohort decodability. We train a logistic-regression classifier (with cross-validation across $C \in \{0.01, 0.1, 1.0, 10\}$, no per-user re-fitting) on each user’s 8-measure realized polar vector and ask it to predict cohort. Table 12 shows the result. HyperPEFT-LoRA’s realized behavioral vectors carry enough cohort information for a linear classifier to hit 0.79 ROC-AUC on real-user replies; Vanilla and Control sit near chance at 0.51 and 0.50. Beyond shifting the cohort means in Table 9, HyperPEFT-LoRA leaves a linearly decodable cohort signal in those vectors.

Coherence rate. On the rage and empath forums, $n = 1,344$ to 1,345 replies per setting, HyperPEFT-LoRA reaches 98.96% coherent output, on par with Vanilla LoRA (100%) and clearly above the control (91.82%). HyperPEFT-LoRA does not trade coherence for cohort separation.

Reply-length confound check. Reply length differs across the rage and empath cohorts, so we repeat the comparison after removing its association with log word count. We use one pooled fit to remove log word count from each reply’s value, average the residuals per author, then recompute Hedges’ g on the same eight measures. All eight point estimates keep the predicted sign, and seven intervals exclude zero (Table 13, left). Politeness is the only interval that touches zero, matching the uncertainty §4.2 already reports. The controlled estimates also land close to the uncontrolled ones in Table 9, within 0.05 on expressiveness, warmth and hostility, so reply length does not explain the overall pattern. Tempo needs a different reading. A generated reply has no inter-post timing, so its realized tempo score comes from reply length. Its Table 13 value is therefore a consistency check, not independent evidence after controlling for length.

Topic-prior confound check. Rage and empath could appear separated if their replies are distributed unevenly across prompts. Each of the three forum settings runs its own 12 prompts, resulting in 36 setting-by-prompt threads. Every reply belongs to exactly one thread. We compute signed Hedges’ g within each thread, then average the 36 estimates with weights from the thread’s author count. All eight point estimates keep the predicted sign, and five of the eight intervals exclude zero (Table 13, right). The separation in Table 9 therefore survives

Table 11: Mann-Whitney U decomposition of HyperPEFT-LoRA’s cohort separation versus the Vanilla LoRA reference, on the real-user forum. *Amplify HIGH*: does HP-LoRA push HIGH-cohort users above the Vanilla HIGH-cohort users? *Suppress LOW*: does HP-LoRA pull LOW-cohort users below the Vanilla LOW-cohort users? *Supported side* identifies which one-sided test passes $p < 0.001$; *n.s.* marks non-significant directions.

Measure	Amplify HIGH (p)	Suppress LOW (p)	Supported side
Politeness	4.5×10^{-5}	n.s.	amplify
Expressiveness	8.5×10^{-96}	n.s.	amplify
Tempo	1.7×10^{-11}	1.1×10^{-5}	both
Self-focus	n.s.	3.1×10^{-29}	suppress
Curiosity	n.s.	3.9×10^{-7}	suppress
Anxiety	n.s.	5.4×10^{-33}	suppress
Warmth	n.s.	9.3×10^{-18}	suppress
Hostility	n.s.	3.9×10^{-30}	suppress

Table 12: Linear-classifier ROC-AUC for cohort prediction from each user’s realized 8-measure polar vector, real-user inference. Higher is better; bold marks the largest point estimate.

Inference setting	ROC-AUC
HyperPEFT-LoRA	0.791
Vanilla LoRA	0.506
Control	0.499

within-thread pooling. Curiosity is the weakest row at +0.05, and its interval is wide: it is scored from whether a reply ends in a question mark, so one thread of about 27 authors per cohort leaves little to separate.

I Persona Stability and Compositionality Indices: Derivations

This appendix defines the Persona Stability Index (PSI) and the Persona Compositionality Index (PCI) introduced in §3.6. PSI uses standardized residuals across the eight signature measures (§I.3); PCI uses the signature cosine between realized and expected per-reply vectors (§I.2), aggregated in §I.4. The supplementary drift-rate diagnostic is defined separately in §I.3.8. Aggregation rules and bootstrap protocols are stated in full so the metrics can be reproduced from the persona-signature parquet outputs alone.

I.1 Measure space and signature vectors

Both metrics measure how closely a generated reply’s behavior matches the user’s intended behavior. The notation starts from all nine behavioral measures, but the reported summaries use smaller projections: PSI and both PCI variants use the eight-measure signature set defined in §I.3.

Measure index. Let $D = \{1, 2, \dots, 9\}$ index the nine measures. By the criterion of Table 1, curiosity, expressiveness, and tempo are in-band and the other six are out-of-band. Every cosine in this appendix uses the eight-measure signature set D^{sig} defined in §I.3.

Realized signature. For author u , thread g , reply i , the realized signature is the 9-vector of measure scores computed on the generated reply text:

$$\hat{s}_{u,g,i} = (\hat{s}_{u,g,i,1}, \dots, \hat{s}_{u,g,i,9}) \in \mathbb{R}^9 \quad (2)$$

For emotion-measure dimensions, $\hat{s}_{u,g,i,d}$ is the GoEmotions composite from the frozen RoBERTa scorer applied to the reply text. For stylometric measures, $\hat{s}_{u,g,i,d}$ is computed from the reply text rather than from the profile vector: self-focus uses a first-person pronoun rate, and politeness uses an equal-weight composite of a z-scored profanity rate and a z-scored hostility composite. Politeness, self-focus, curiosity, and tempo are z-scored and carry real-user expected standard deviations near 1.0; expressiveness and the three GoEmotions composites (anxiety, warmth, hostility) are left on their raw scales and carry expected standard deviations near 0.01. The signature cosine of §I.2 applies no per-dimension rescaling, so it is dominated by the four z-scored dimensions. PSI restores a common scale by dividing every residual by σ_d^{real} (§I.3).

Expected signature. For author u , the expected signature

$$\mathbf{s}_u^* = (s_{u,1}^*, \dots, s_{u,9}^*) \in \mathbb{R}^9 \quad (3)$$

is fixed before generation and never uses generated text. For a real author, every component is read off that author’s training-corpus row. For an in-hull or near-hull synthetic user, we reuse the anchors

Table 13: Rage-vs-empath signed Hedges’ g on the eight signature measures under two confound controls, real-user forum, over 2,020 replies from 500 authors. Both columns average to per-author values before the contrast, the same grain as Table 9. **Left:** log word count regressed out of each reply’s polar value by a pooled fit, then averaged per author. **Right:** computed within each of 36 setting-by-prompt threads, then averaged with weights from the thread’s author count. Both use 1,000 user-resampled bootstrap draws, seed 142. All eight measures carry the predicted sign under both controls.

Measure	Reply length controlled for		Pooled within thread topic	
	g	95% CI	g	95% CI
Politeness	+0.11	[−0.02, +0.28]	+0.08	[−0.00, +0.14]
Self-focus	+0.24	[+0.12, +0.39]	+0.21	[+0.14, +0.28]
Curiosity	+0.25	[+0.13, +0.37]	+0.05	[−0.19, +0.38]
Expressiveness	+1.11	[+0.99, +1.26]	+0.76	[+0.70, +0.85]
Tempo	+1.18	[+1.04, +1.35]	+0.73	[+0.65, +0.82]
Anxiety	+0.29	[+0.20, +0.43]	+0.44	[+0.42, +0.55]
Warmth	+0.43	[+0.32, +0.53]	+0.26	[+0.22, +0.37]
Hostility	+0.33	[+0.21, +0.45]	+0.09	[−0.01, +0.14]

and Dirichlet weights that built the profile vector: measures whose defining feature is not a profile coordinate get the weighted average over those anchors, and the rest are read off the synthetic vector itself. Both interpolation settings therefore build the expected signature from weighted averages of real authors rather than from an independently held-out quantity. Far-from-hull synthesis starts from a single anchor. There the anchor-averaged measures keep that anchor’s values and do not move with κ , while the profile-coordinate measures are extrapolated with it, past any real author’s value. The expected signature is the target reference for each realized signature.

Analysis projection. Restricting both vectors to D^{sig} :

$$\begin{aligned}\hat{\mathbf{s}}_{u,g,i}^{\text{sig}} &= (\hat{s}_{u,g,i,d})_{d \in D^{\text{sig}}}, \\ \mathbf{s}_u^{*\text{sig}} &= (s_{u,d}^*)_{d \in D^{\text{sig}}}\end{aligned}\quad (4)$$

I.2 Per-reply signature cosine

The signature cosine is the cosine of the angle between the realized and expected eight-measure signatures.

Step 1: dot product.

$$\hat{\mathbf{s}}_{u,g,i}^{\text{sig}} \cdot \mathbf{s}_u^{*\text{sig}} = \sum_{d \in D^{\text{sig}}} \hat{s}_{u,g,i,d} \cdot s_{u,d}^* \quad (5)$$

Step 2: norms.

$$\begin{aligned}\|\hat{\mathbf{s}}_{u,g,i}^{\text{sig}}\|_2 &= \sqrt{\sum_{d \in D^{\text{sig}}} (\hat{s}_{u,g,i,d})^2}, \\ \|\mathbf{s}_u^{*\text{sig}}\|_2 &= \sqrt{\sum_{d \in D^{\text{sig}}} (s_{u,d}^*)^2}\end{aligned}\quad (6)$$

Step 3: cosine.

$$c_{u,g,i} = \frac{\hat{\mathbf{s}}_{u,g,i}^{\text{sig}} \cdot \mathbf{s}_u^{*\text{sig}}}{\|\hat{\mathbf{s}}_{u,g,i}^{\text{sig}}\|_2 \cdot \|\mathbf{s}_u^{*\text{sig}}\|_2} \in [-1, +1] \quad (7)$$

We use cosine rather than Euclidean distance because a single reply produces a smaller realized vector than a full author history does. Cosine removes that magnitude difference and compares only the two directions. PCI aggregates $c_{u,g,i}$, first across a user’s replies and then across users (§I.4). PSI does not use the signature cosine: it aggregates the standardized absolute residuals of Equation (8) instead (§I.3).

I.3 Persona Stability Index (PSI)

PSI measures how close each user’s generated behavior sits to their intended profile. Every user with at least one coherent reply contributes one PSI value, so single-reply users are not dropped.

I.3.1 Signature measure subset

Let $D^{\text{sig}} \subseteq D$ be the eight per-user signature measures used at evaluation time: politeness, self-focus, curiosity, expressiveness, tempo, anxiety, warmth, and hostility. In the terms of Table 1 these are five of the six out-of-band measures and all three in-band controls. Sentiment is omitted from D^{sig} for both PSI and PCI because it defines the cohort labels; including it would score the same axis used to select the cohorts. For each $d \in D^{\text{sig}}$, $\hat{s}_{u,g,i,d}$ is the realized polar value on reply i of author u in thread g , and $s_{u,d}^*$ is the user’s expected polar value, fixed before generation and constructed as described in §I.1.

I.3.2 Per-dimension standardization

The eight signature measures sit on different natural scales: the z-scored stylometric dimensions (politeness 0.9261, self-focus 1.2340, curiosity 1.0213, tempo 1.0364) carry real-user expected standard deviations near 1.0, while expressiveness (0.0168) and the raw GoEmotions composites (anxiety 0.0031, warmth 0.0183, hostility 0.0126) carry standard deviations near 0.01. Averaging absolute residuals without standardization would let the largest-scale dimension dominate the sum. We therefore pre-compute the real-user expected standard deviation σ_d^{real} per dimension on the production training pool and standardize each residual:

$$\delta_{u,g,i,d} = \frac{|\hat{s}_{u,g,i,d} - s_{u,d}^*|}{\sigma_d^{\text{real}}}. \quad (8)$$

I.3.3 Per-reply match score

The per-reply match score averages the standardized residuals across the signature measures and converts to a $[0, 1]$ scale through a halved-distance hinge:

$$\text{match}_{u,g,i} = \max\left(0, 1 - \frac{1}{2|D^{\text{sig}}|} \sum_{d \in D^{\text{sig}}} \delta_{u,g,i,d}\right). \quad (9)$$

The halving sets the per-reply match score to zero at a mean standardized residual of two standard deviations, so a user whose realized signature sits two real-user SDs from their target on every signature dimension scores zero. A user whose realized signature equals their target on every dimension scores one. This is the form Equation (1) in §3.6 writes in compact notation.

I.3.4 Per-user PSI

Let R_u be the set of (g, i) index pairs at which author u produced a coherent reply (passing the coherence gate of §G). The per-user PSI is the mean of the per-reply match score across that user's surviving replies:

$$\text{PSI}(u) = \frac{1}{|R_u|} \sum_{(g,i) \in R_u} \text{match}_{u,g,i}. \quad (10)$$

Every author with $|R_u| \geq 1$ contributes one PSI value. There is no group-size requirement.

I.3.5 Per-stratum aggregation

For a stratum s (real-user inference, in-hull, near-hull, and far-from-hull at each κ level), the reported summary is the across-user mean:

$$\text{PSI}_s = \frac{1}{|\mathcal{U}_s|} \sum_{u \in \mathcal{U}_s} \text{PSI}(u), \quad (11)$$

where $\mathcal{U}_s$ is the set of authors evaluated in stratum s . Table 14 lists PSI_s for the seven strata reported in the main paper.

I.3.6 Bootstrap confidence intervals

For each stratum s and bootstrap iteration $b = 1, \dots, B$ with $B = 1,000$ and seed 142:

1. Sample $|\mathcal{U}_s|$ authors from $\mathcal{U}_s$ with replacement to form $\mathcal{U}_s^{(b)}$.
2. Recompute $\text{PSI}_s^{(b)} = \frac{1}{|\mathcal{U}_s|} \sum_{u \in \mathcal{U}_s^{(b)}} \text{PSI}(u)$.

The 95% bootstrap CI is the empirical 2.5th and 97.5th percentiles of those values. Resampling at the author level is the right grain because replies within an author are not independent.

I.3.7 Empirical threshold derivation

The 0.10 threshold used in §4.4 and Appendix I.5 is a reference line calibrated on the pooled production set ($n = 12,451$ rows, one per user per generation cell, a wider pool than the seven strata of Table 14). It sits well below the on-manifold PSI band, where real-user, in-hull, and near-hull inference all score 0.226 or higher, and marks the off-manifold collapse, where per-user fidelity falls from 0.104 at $\kappa=3$ to 0.033 at $\kappa=10$ (Table 14). It separates the on-manifold settings from the off-manifold sweep rather than fixing a preregistered cutoff.

I.3.8 Per-dimension drift slope (supplementary)

PSI is a single scalar per group and does not say which measure dimension is drifting. For that we compute drift slopes by ordinary least squares.

Step 1: per-turn residual. For dimension d :

$$r_{u,g,i,d} = \hat{s}_{u,g,i,d} - s_{u,d}^*. \quad (12)$$

Step 2: OLS as RSS minimizer. Treat the reply index i as the within-thread time axis:

$$\text{RSS}(\alpha, \beta) = \sum_{i=1}^{n_{u,g}} (r_{u,g,i,d} - \alpha - \beta i)^2. \quad (13)$$

Step 3: normal equations.

$$\frac{\partial \text{RSS}}{\partial \alpha} = -2 \sum_{i=1}^{n_{u,g}} (r_{u,g,i,d} - \alpha - \beta i) = 0, \quad (14)$$

$$\frac{\partial \text{RSS}}{\partial \beta} = -2 \sum_{i=1}^{n_{u,g}} i (r_{u,g,i,d} - \alpha - \beta i) = 0. \quad (15)$$

Step 4: solve for β .

$$\hat{\beta}_{u,g,d} = \frac{\sum_{i=1}^{n_{u,g}} (i - \bar{i}) (r_{u,g,i,d} - \bar{r}_{u,g,d})}{\sum_{i=1}^{n_{u,g}} (i - \bar{i})^2} \quad (16)$$

with $\bar{i} = \frac{1}{n_{u,g}} \sum_{i=1}^{n_{u,g}} i$ and $\bar{r}_{u,g,d} = \frac{1}{n_{u,g}} \sum_{i=1}^{n_{u,g}} r_{u,g,i,d}$. Computed only for groups with $n_{u,g} \geq 3$ and non-zero denominator.

Step 5: aggregate.

$$\bar{\beta}_d = \frac{1}{|\mathcal{G}_d|} \sum_{(u,g) \in \mathcal{G}_d} \hat{\beta}_{u,g,d}. \quad (17)$$

Sign interpretation. The sign of $\bar{\beta}_d$ reports movement along the signed residual axis. It does not by itself indicate movement toward or away from the target, because that depends on the residual’s starting sign. We use $|\bar{\beta}_d|$ as the per-turn rate-of-change diagnostic. A direct toward-target analysis would regress $|r_{u,g,i,d}|$ rather than $r_{u,g,i,d}$.

I.4 Persona Compositionality Index (PCI)

We report two descriptive versions of PCI, and each answers a different question. *Target-driven* PCI asks what happens as more requested traits are added. It uses two fixed trait chains and groups synthetic users by requested-trait count m (Appendix J). *Agreement-stratified* PCI instead asks whether users who realize more expected cohort directions also match their expected eight-measure pattern more closely. It groups production users by realized agreement count, the number of measures whose mean realized value has the expected sign and a magnitude of at least 0.05 (Table 16). Both versions use the unstandardized cosine of §I.2, so the dimensions do not receive equal weight; as §I.3 notes, the four z-scored coordinates dominate it. Neither version includes sentiment, because sentiment defines the cohort labels, so the first constraint in each chain of Appendix J is not itself scored. We report the two versions separately and do not combine them into one pass-or-fail result. The rest of this appendix defines both.

I.4.1 Synthesis specification

Let $\mathcal{T}$ be the trait vocabulary (age, empathy, polite, vulgar, selfless, egocentric, deliberate, reactive, ...). For each synthesized user u , the synthesis specification carries a target trait list $\text{labels}_u \subseteq \mathcal{T}$ and a count $m_u = |\text{labels}_u|$. Here m is the requested-trait count, K the profile-feature count, k the number of anchors, and η the Dirichlet concentration.

I.4.2 Per-user PCI

PCI is averaged over the same coherence-passing replies used for PSI. With R_u as in §I.3:

$$\text{PCI}(u) = \frac{1}{|R_u|} \sum_{(g,i) \in R_u} c_{u,g,i}. \quad (18)$$

I.4.3 Trait-count partition

$$\mathcal{U}_m = \{u : m_u = m\}, \quad m \in \{1, 2, 3, 4\} \quad (19)$$

$$\text{PCI}_m^{\text{tgt}} = \frac{1}{|\mathcal{U}_m|} \sum_{u \in \mathcal{U}_m} \text{PCI}(u). \quad (20)$$

I.4.4 Bootstrap procedure

For each level and each iteration $b = 1, \dots, B$ with $B = 1,000$ and seed 142, resample the users in that level with replacement and recompute the mean. The 95% CI is the empirical 2.5th and 97.5th percentiles of those values. The resampling unit is the user, because replies within a user are not independent. For the target-driven variant the resampling unit is the *synthetic* user, so the interval is conditional on the fixed eligible real-author pool and does not carry the uncertainty that would come from redrawing those source authors.

I.4.5 Hypothesized pattern: decline with m

$$\text{PCI}_1^{\text{tgt}} > \text{PCI}_2^{\text{tgt}} > \text{PCI}_3^{\text{tgt}} \quad (21)$$

This inequality is a hypothesis, not a mathematical consequence of the sampler. Only the first step below is guaranteed by construction; the rest require intermediate measurements the constrained run does not report.

Step 1: the eligible anchor pool shrinks as m grows. Each new constraint can only remove authors from the eligible pool, so $|\mathcal{A}_1| \geq |\mathcal{A}_2| \geq |\mathcal{A}_3| \geq |\mathcal{A}_4|$, where $\mathcal{A}_m$ is the set of training authors satisfying the first m constraints.

Step 2: the reachable mixture set can contract. The Dirichlet sampler draws

$$\mathbf{g}'_u = \sum_{a=1}^k \lambda_a \mathbf{g}_{u_a}, \quad \lambda \sim \text{Dir}(\boldsymbol{\eta}) \quad (22)$$

over k anchors drawn uniformly from $\mathcal{A}_m$. As $\mathcal{A}_m$ shrinks, anchor reuse increases and the reachable set of mixtures can contract toward fewer source individuals.

Step 3: contraction does not fix the profile-space geometry. We hypothesized that accepted profiles would shift toward the boundary of the hull as the eligible pool thinned, leaving the dense interior the hypernetwork was trained on. The profile vector certainly stays inside the hull (Dirichlet weights are non-negative and sum to one), but pool size alone does not determine where inside it the mixtures land, and the constrained run does not report Mahalanobis distance by m .

Step 4: offset behavior by m is unmeasured. Appendix M shows that offset magnitude changes with profile-space distance (log-log slope ≈ 0.88) and that direction is roughly preserved while magnitude is not, but it does not establish a decrease in magnitude as m grows, and no constrained-run offset norm by m is reported.

Step 5: a decline is therefore conditional. If the constrained profiles produce offsets whose magnitude or direction is less faithful to the target, the realized signature moves away from that target, which lowers $c_{u,g,i}$, then $\text{PCI}(u)$, then $\text{PCI}_m^{\text{tgt}}$. The measured empath chain does not follow this prediction: its value rises from $m=1$ to $m=3$. The five-step account is therefore not supported as a general mechanism, and at most describes the pooled trend and the rage chain in these two fixed examples (Appendix J).

I.4.6 Diagnostic shapes

The shape of $m \mapsto \text{PCI}_m^{\text{tgt}}$ is descriptive. A flat curve indicates no change in the aggregate score, a drop at $m = 2$ marks a lower score after the first added trait, and an uneven or rising curve contradicts (21). Appendix J reports opposite cohort-specific shapes: rage declines while empath rises.

I.4.7 Symbols (quick reference)

- u, g, i, d : author index, thread index, reply index within thread, measure dimension index.
- $\hat{s}_{u,g,i}, s_u^*$: realized and expected 9-dim signatures.
- $c_{u,g,i}$: per-reply signature cosine on the eight measures of D^{sig} , (7).
- $r_{u,g,i,d}$: per-turn realized-vs-expected residual on dimension d .
- $\hat{\beta}_{u,g,d}, \bar{\beta}_d$: per-group OLS drift slope, per-dimension mean drift slope.

- labels_u, m : synthesis target trait list, requested-trait count.
- $\mathcal{A}_m$: training authors satisfying the first m constraints.
- $\mathcal{U}_m$: synthetic users with exactly m target traits.
- $\mathcal{G}$: admissible (author, thread) groups; $\mathcal{G}_d$ the subset where $\hat{\beta}_{u,g,d}$ is well-defined.

I.4.8 Operational note on the reported runs

The production runs reported in the main tables leave labels_u empty: every synthetic user there is drawn from a Dirichlet mixture over anchors selected without a trait specification. Target-driven PCI is therefore measured in a separate constrained run, described in Appendix J, which uses the released sampler’s filter mode to restrict the eligible author pool before each mixture is drawn. The agreement-stratified variant in Table 16 uses the realized agreement count defined at the start of §I.4 rather than the requested-trait count. The two variants answer related but different questions. Target-driven PCI tracks signature fidelity as more source-pool constraints are added. The agreement-stratified variant groups users by how many signature measures they realize and reports the corresponding overall cosine.

I.5 H3 success criteria

The within-thread half of H3 is supported if (i) the mean PSI is above the empirical 0.10 threshold on in-hull, near-hull, and real-user inference, with the lower bound of the bootstrap 95% CI also above 0.10 in all three settings, and (ii) the per-dimension mean drift slopes meet $|\bar{\beta}_d| < 0.05$ per turn on all five out-of-band signature measures under the in-hull setting. The far-from-hull κ -sweep is treated as a boundary test rather than as pass-or-fail: the prediction is a steady PSI collapse with growing κ . The 0.10 threshold is a reference line below the on-manifold PSI band and at the off-manifold collapse across all 12,451 rows of that pooled set, not a preregistered guess. The agreement-stratified PCI companion is supported if (iii) the mean signature cosine rises across the realized-agreement partitions of Table 16 with non-overlapping bootstrap 95% CIs between consecutive partitions, and (iv) the difference between the four-or-more and zero-match partitions is at least 1.0 on the pooled synthesis pool. Criteria (iii) and (iv) refer only to

Table 16, which groups production users by what they realized. The target-driven analysis in Appendix J uses a separate constrained run grouped by what was requested; it is a cohort-specific descriptive trend and is not part of the H3 verdict.

I.6 Supplementary numerical results

PSI separates the settings inside the real-user range from those pushed beyond it (Figure 3, Table 14). Real-user inference scores 0.226, in-hull 0.265, and near-hull 0.257. All three lie in a narrow band above the empirical 0.10 threshold, so inside the real-user range the hypernetwork produces per-user behavior that stays close to its target. The far-from-hull sweep declines as κ grows: 0.104 at $\kappa=3$, 0.033 at $\kappa=10$, 0.013 at $\kappa=25$, and 0.001 at $\kappa=100$. At $\kappa=3$ the profiles already sit outside the real-user range, yet PSI remains at the threshold; pushing them farther drops it below.

The same pattern appears within each cohort. Across real-user, in-hull, and near-hull inference, PSI is 0.22, 0.23, and 0.20 for rage and 0.23, 0.28, and 0.26 for empath. In-hull is highest for both cohorts, and real-user and near-hull inference exchange order. Per-dimension drift slopes sit well inside the ± 0.05 per-turn threshold on every measure under the in-hull setting; self-focus has the largest mean slope at -0.0111 , under one quarter of the threshold magnitude. These slopes come from the short-thread runs; the 80-turn extended-depth slopes are reported separately in Appendix L. The local-density drift hypothesis, measured as the Spearman correlation between residual magnitude and local profile-space density, returns $\rho = -0.020$ at $n = 2,550$ rows ($p = 0.31$). The association is not statistically distinguishable from zero.

Target-distance visualization of PSI. Figure 7 plots one marker per setting in separate rage and empath panels, each carrying its mean match score (Equation (9)). For the three on-manifold settings the marker radius shows rank within that panel rather than an absolute residual. The four far-from-hull stars use fixed spacing along one ray, and the rings are visual guides. Real-user inference, in-hull, and near-hull score 0.20 to 0.28 in both cohorts. At $\kappa=25$ the match falls to 0.016 on rage and 0.012 on empath, with bootstrap CIs that do not overlap any on-manifold band. Generated behavior stays close to its target inside the real-user range, and fidelity declines as the profile vector is pushed beyond it.

Per-turn drift trajectory. Figure 8 plots per-turn persona fidelity (the 8-measure PSI match to each user’s target) across the 80-turn extended-depth test, where a higher value means a closer match to the profile. The settings separate on the first turn and stay at distinct levels. Real-user inference stays near 0.19 and in-hull synthesis near 0.18. Far-from-hull synthesis stays lower, at about 0.08 for $\kappa=3$, 0.01 for $\kappa=25$, and 0.00 for $\kappa=100$. This matches the inside-versus-outside pattern in Figure 3, although the separate 80-turn run has lower absolute values. Real-user and in-hull inference exchange order between the two runs, and near-hull is not part of the 80-turn set. The gap appears at user creation rather than growing over later turns: every trajectory holds flat, and the largest per-measure slope, self-focus at 0.0095 per turn, is 19% of the $|\bar{\beta}_d| < 0.05$ criterion. On-manifold users stay near their initial fidelity level across the long run, while off-manifold users remain at lower levels as κ increases.

J Persona Compositionality

Target-driven PCI (§I.4) reports the eight-measure signature cosine after the sampler restricts the training-author pool with a fixed trait chain. The released filter mode applies those constraints before each Dirichlet mixture is drawn. We evaluate fixed nested rage and empath chains for $m \in \{1, 2, 3\}$, and do not report $m=4$ because only four rage-chain authors and two empath-chain authors satisfy all four constraints; additional profiles at that level would recombine the same handful of individuals.

The two chains and how they were chosen. The rage chain adds rage sentiment, then low politeness, then high self-focus, then fast tempo. The empath chain adds empath sentiment, then high politeness, then low self-focus, then slow tempo. Both were fixed before the constrained run and follow the paper’s own measure ordering: the cohort-defining sentiment axis first, then politeness and self-focus, then tempo. Neither chain was chosen after inspecting a result.

Where users are lost. Table 15 prints every stage between the eligible author pool and the scored users, so no attrition is hidden. At $m=2$ and $m=3$ the synthesized counts exceed the eligible-author counts, because many mixtures reuse the same anchors; that is the dependence the bootstrap does not remove. Near-hull acceptance is where the cor-

Table 14: Per-setting PSI summary, pooled across cohorts. The three on-manifold settings (real-user, in-hull, near-hull) cluster above the empirical threshold ($\text{PSI} \geq 0.10$); the off-manifold κ sweep collapses smoothly toward zero. n is the per-stratum user count; CI is the 95% bootstrap CI on the across-user mean PSI.

Stratum	n	mean PSI	95% CI
real-user inference	500	0.226	[0.214, 0.238]
in-hull	831	0.265	[0.256, 0.274]
near-hull	577	0.257	[0.248, 0.267]
far-from-hull $\kappa=3$	1138	0.104	[0.096, 0.113]
far-from-hull $\kappa=10$	1148	0.033	[0.028, 0.039]
far-from-hull $\kappa=25$	1141	0.013	[0.011, 0.017]
far-from-hull $\kappa=100$	756	0.001	[0.000, 0.001]

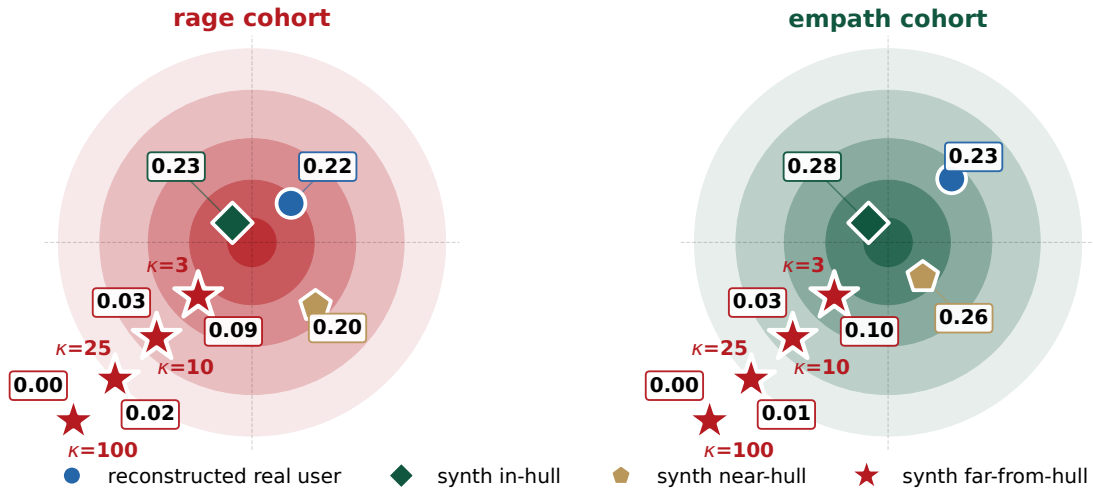


Figure 7: Per-user PSI by setting across rage and empath panels. Match score is Equation (9): the per-reply standardized residual averaged across the eight signature measures, halved-hinge mapped to $[0, 1]$. Markers are placed by within-panel ranking; the absolute match score sits beside each marker. The four off-manifold κ stars stack along the SW ray, each labeled with its κ value. Real-user inference, in-hull, and near-hull cluster near the center on rage; empath in-hull and near-hull match or exceed real-user inference. The $\kappa = 25$ off-manifold setting collapses to 0.016 (rage) and 0.012 (empath), matching the pooled $\kappa=25$ value of 0.013 in Table 14.

pus limit shows first: it falls from 348 to 164 to 19 accepted profiles along the rage chain and from 360 to 293 to 12 along the empath chain. In-hull acceptance holds between 377 and 396 across the rage chain and gives way only at the last empath level, from 395 to 299.

Result. Pooled across the two chains, $\text{PCI}_m^{\text{tgt}}$ falls from $+0.109$ at $m=1$ to $+0.102$ at $m=2$ and -0.015 at $m=3$, whose interval spans zero. The pooled mean does not represent either cohort trend: rage falls from -0.021 to -0.411 , while empath rises from $+0.240$ to $+0.505$. Every per-chain interval except rage at $m=1$ excludes zero, but the opposing directions mean (21) is not supported as a cohort-general prediction. Because PCI is a signature cosine, the negative rage values mean the

realized eight-measure signature points away from its expected direction. Table 17 shows the same cohort split in a different comparison: synthesis lowers the signature cosine relative to real-user inference for rage and raises it for empath. These two fixed nested chains do not establish arbitrary trait composition, and each added constraint changes the eligible source population. The run scores the same generated-text signature used throughout the paper (§3.5), so it inherits that instrument’s limits and is not an independent representation-level test of whether a requested trait was satisfied.

Agreement-stratified companion. Alongside the constrained run we report the **agreement-stratified signature cosine** on the production synthesis pool, partitioned by the realized agreement

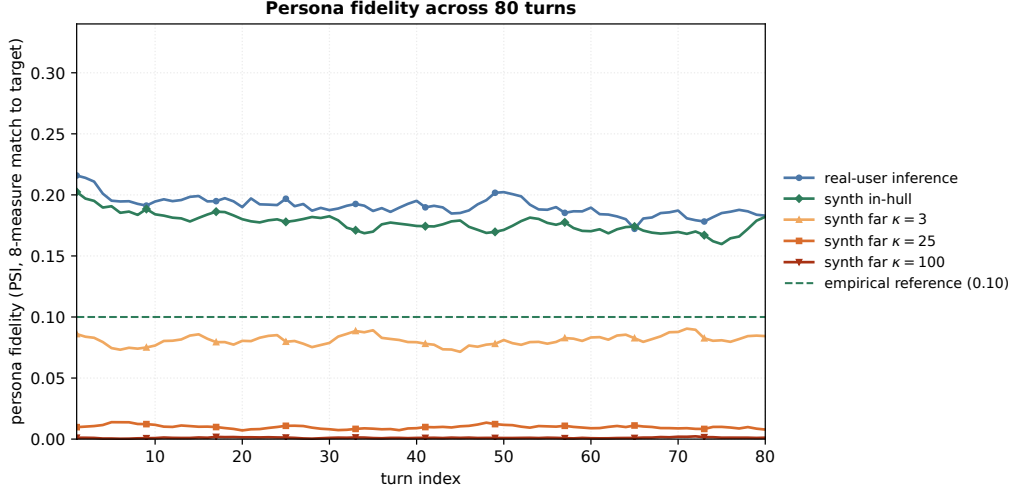


Figure 8: Per-turn persona fidelity (8-measure PSI match to target, higher is better) over the 80-turn extended-depth test, for five settings. Real-user inference and in-hull synthesis hold high fidelity at every turn, while the far-from-hull settings sit at progressively lower levels as κ grows. The trajectories change little over the 80 turns, so the separation is already present at instantiation rather than building over time. All settings meet the per-turn no-drift criterion, with the largest per-measure slope magnitude (self-focus, 0.0095 per turn) about one-fifth of the $|\beta_d| < 0.05$ threshold.

Table 15: Target-driven persona compositionality on two fixed nested trait chains. *Eligible* counts the training authors satisfying all m constraints; *proposed*, the 400 in-hull plus 400 near-hull draws attempted at each level; *synthesized*, the profiles the density gate accepted; *generated*, the synthetic users seated in the forum; *scored*, the users with at least one reply through the coherence gate of §3.6. Confidence intervals use 1,000 synthetic-user bootstrap resamples and are conditional on the fixed eligible real-author pool.

Chain	m : constraint added	Eligible	Proposed	Synthesized	Generated	Scored	PCI $_{m}^{\text{tgt}}$ (95% CI)	
Rage	1: rage sentiment	2,000	800	744	391	361	-0.021	[-0.062, +0.025]
Rage	2: + low politeness	383	800	549	303	284	-0.186	[-0.209, -0.162]
Rage	3: + high self-focus	22	800	396	279	267	-0.411	[-0.443, -0.381]
Rage	4: + fast tempo	4	—	—	—	—	not estimable here	
Empath	1: empath sentiment	2,000	800	755	385	358	+0.240	[+0.198, +0.279]
Empath	2: + high politeness	323	800	688	357	333	+0.347	[+0.317, +0.375]
Empath	3: + low self-focus	37	800	311	212	203	+0.505	[+0.455, +0.553]
Empath	4: + slow tempo	2	—	—	—	—	not estimable here	
Pooled	$m = 1$	—	—	—	—	719	+0.109	[+0.078, +0.142]
Pooled	$m = 2$	—	—	—	—	617	+0.102	[+0.074, +0.129]
Pooled	$m = 3$	—	—	—	—	470	-0.015	[-0.065, +0.031]

count defined in §I.4, with the mean signature cosine and its 95% percentile bootstrap CI inside each partition. The stratifier and the response are computed from the same eight-measure signature, so part of any rise across partitions follows from the construction, and it groups users by what they realized rather than by what was requested. Table 16 gives the values: the mean rises steadily from -0.57 for zero matched measures to $+0.48$ for at least four matched measures on 3,843 synthetic users, with non-overlapping 95% bootstrap CIs between every adjacent pair of partitions.

Table 16: Agreement-stratified signature cosine by realized cohort agreement count. Synthetic users are pooled across every synthesis cell, including the diagnostic κ levels (3,843 users). A user generated into several cells is counted once here, which is why this total is smaller than the sum of the per-stratum counts in Table 14; the table reports mean signature cosine with percentile bootstrap 95% CIs from 1,000 user-resampled draws (seed 142). CIs do not overlap between adjacent partitions.

Measures matched	Mean cosine	95% CI
0	-0.57	$[-0.62, -0.53]$
1	-0.28	$[-0.31, -0.25]$
2	0.00	$[-0.02, +0.03]$
3	+0.26	$[+0.23, +0.28]$
≥ 4	+0.48	$[+0.45, +0.51]$

K Synthesis Compared with Real-User Inference

Table 17 shows the synthetic-user mean minus the real-user-inference mean for the eight-measure signature cosine.

On rage, the synthesis-versus-real-user gap is small for in-hull (-0.07), widest at near-hull (-0.18), and partly closes at far-from-hull (-0.10). On empath, the gap stays small at in-hull and near-hull ($+0.03$, $+0.04$) and widens at far-from-hull ($+0.14$), where synthetic users score closer to their targets than real-user inference does on this directional measure. These differences are descriptive: the reported analysis does not isolate whether they come from density gating, cohort filtering, the pushout, or interactions among them.

Per-user inference fidelity. On the training-set real-user forums we also compare each author’s generated replies with that same author’s training-corpus replies, as the cosine between the two sets’ mean sentence embeddings. This generated-to-training text cosine averages 0.153 per user (95% CI $[0.142, 0.163]$, $n = 423$) on the rage forum and 0.156 (95% CI $[0.146, 0.166]$, $n = 431$) on the empath forum. Mean sentiment MAE is 0.249 and 0.303, respectively. Mean lexical Jaccard against those training replies is 0.030 and 0.033. These are within-author comparisons. §4.1 instead measures behavioral separation between cohorts.

Expected-to-realized correlation. For each setting we compute Spearman’s r between a user’s expected and realized polar values, across the rage and empath cohorts. All three cohort-tied measures correlate more strongly under synthesis than under real-user inference; hostility rises from $+0.06$ to

$+0.61$ at near-hull (Table 18). This is a descriptive counterpart to the cohort-mean separation in Table 9; the current analysis does not isolate why the synthetic-user correlations are higher.

Per-cohort signature cosine across settings. Table 19 gives the eight-measure signature cosine by cohort and setting; subtracting each real-user inference value from the first three synthesis rows gives the gaps in Table 17. Rage is highest at real-user inference ($+0.084$) and lowest at near-hull (-0.096), and its other three settings stay within 0.05 of zero. Empath is $+0.120$ at real-user inference and ranges from $+0.153$ to $+0.339$ across the synthesis settings. H2 instead uses per-cohort PSI, which summarizes the standardized residual across all eight signature measures rather than direction alone (§4.3).

Table 17: Synthetic-user minus real-user-inference gap in the eight-measure signature cosine, by cohort and setting. The cosine compares each realized eight-measure signature with its expected signature; the gap is the synthetic-user value minus the real-user-inference value. Negative values mean synthesis has a lower target cosine than real-user inference, positive values mean it has a higher one. Far-from-hull denotes $\kappa=3$.

Cohort	Setting	Eight-measure signature-cosine gap
rage	in-hull	-0.07
rage	near-hull	-0.18
rage	far-from-hull ($\kappa=3$)	-0.10
empath	in-hull	+0.03
empath	near-hull	+0.04
empath	far-from-hull ($\kappa=3$)	+0.14

Table 18: Per-user Spearman r between realized and expected polar values, by setting, on the three cohort-tied measures. Far-from-hull is $\kappa=3$. Higher is better; bold marks the largest point estimate on each row.

Measure	Real-user	Synthesis		
	inference	In-hull	Near-hull	Far-from-hull
Sentiment	+0.13	+0.33	+0.39	+0.21
Hostility	+0.06	+0.52	+0.61	+0.39
Warmth	+0.12	+0.24	+0.28	+0.14

Table 19: Mean eight-measure signature cosine by cohort and setting. Two far-from-hull levels are shown; Table 17 uses the first three synthesis rows.

Setting	Rage	Empath
Real-user inference	+0.084	+0.120
In-hull	+0.011	+0.153
Near-hull	-0.096	+0.164
Far-from-hull ($\kappa=3$)	-0.015	+0.259
Far-from-hull ($\kappa=10$)	-0.048	+0.339

L Extended-Depth Stability: Per-Turn Trajectory

The extended-depth test ran HyperPEFT-LoRA for 80 consecutive turns with a 600-token visible-history budget, dropping the oldest tokens first. All five settings were generated under the same 80-turn protocol: a 187-author real-user pool (93 rage, 94 empath), a 193-user in-hull synthesis pool, and 200-user pools at each far-from-hull $\kappa \in \{3, 25, 100\}$. The in-hull pool for this run was drawn by a second interior sampler that we fixed for this experiment: each user persona profile is the equal-weight average of two training authors from the same cohort, with no density gate. Like the mixtures of §3.4, it is a convex combination of real profiles and lies inside the real-user hull, at a median Mahalanobis distance of 2.8 against the real-user median of 3.6, so the long-horizon result covers interior synthe-

sis under a sampler drawn independently of the one behind the short-thread tables. The panels of Figure 9 and Table 20 are drawn on the real-user pool, a corpus of $80 \times 187 = 14,960$ replies; the persona-fidelity trajectory of Figure 8 and the per-setting coherence rates below cover all five. Generation used temperature 0.4, repetition penalty 1.25, top-p 0.85, and a 96-token maximum, against 192 tokens in a 512-token context for the short-thread runs. The inference-time steering ran in orthogonal mode at $\alpha = 1.0$ over decoder layers 15 to 23. Wall-clock time was about a day and a half on a single multi-GPU node.

Coherence. The per-turn coherence rate (the fraction of replies passing the coherence gate of Appendix G) fell from 0.957 on rage and 0.989 on empath at turn 1 to 0.839 and 0.840 at turn 80. Synthetic users hold the same band. In-hull synthesis starts at 0.990 at turn 1, above the real-user pool, and reads 0.819 at turn 80; its across-turn mean is 0.856 against 0.857 for real-user inference. The three far-from-hull settings average 0.840, 0.816, and 0.829 at $\kappa=3, 25$, and 100. Every setting therefore stays coherent across the horizon: real-user and in-hull means sit within 0.001 of each other, the lowest off-manifold mean is above 0.81, and no setting drops below 0.74 at any single turn. Target match over the same runs falls to near zero off the manifold (Figure 8). On the plotted real-user pool the decline is front-loaded and then flattens near 0.84 rather than continuing to fall.

Follow-on scoring. We compute per-dimension drift slopes as the OLS slope of the per-turn realized value on turn index, for each of the eight signature measures under each setting. The slopes are uniformly small: on the real-user pool the largest magnitude is self-focus at 0.0040 per turn, 8% of the pre-specified $|\beta_d| < 0.05$ threshold, and across

all five settings the single largest per-measure slope magnitude is self-focus at 0.0095 per turn under the far-from-hull setting, 19% of that threshold. Persona fidelity stays flat over the horizon (Figure 8), so the off-manifold shortfall is a level gap set at instantiation, not a per-turn drift. On the real-user pool, coherence remains high through turn 80, and across all five settings the per-measure slopes remain below the stated threshold. The extended-depth stability claim therefore holds on both axes under the steered protocol: the largest per-measure slope is 19.0% of the drift threshold, and no setting falls below a 0.74 coherence rate at any turn.

Table 20: Extended-depth stability summary on the 187-author real-user pool. Per-cohort coherent fraction at turn $t \in \{1, 40, 80\}$ and OLS coherent-fraction slope on turn index across all 80 turns. Drift slopes are OLS slopes of the per-turn realized value on turn index. The final three rows each report one value across cohorts: the first two are the real-user mean slopes for self-focus and politeness, and the last is the largest absolute slope over all eight signature measures and five settings.

Metric	Rage	Empath
Coherent frac, turn 1	0.957	0.989
Coherent frac, turn 40	0.849	0.840
Coherent frac, turn 80	0.839	0.840
Coherent-frac OLS slope / turn	-0.00001	-0.00107
$\bar{\beta}_d$ (self-focus, real-user)	-0.0040	
$\bar{\beta}_d$ (politeness, real-user)	-0.00014	
Max $ \bar{\beta}_d $ (self-focus, off-manifold)	0.0095	

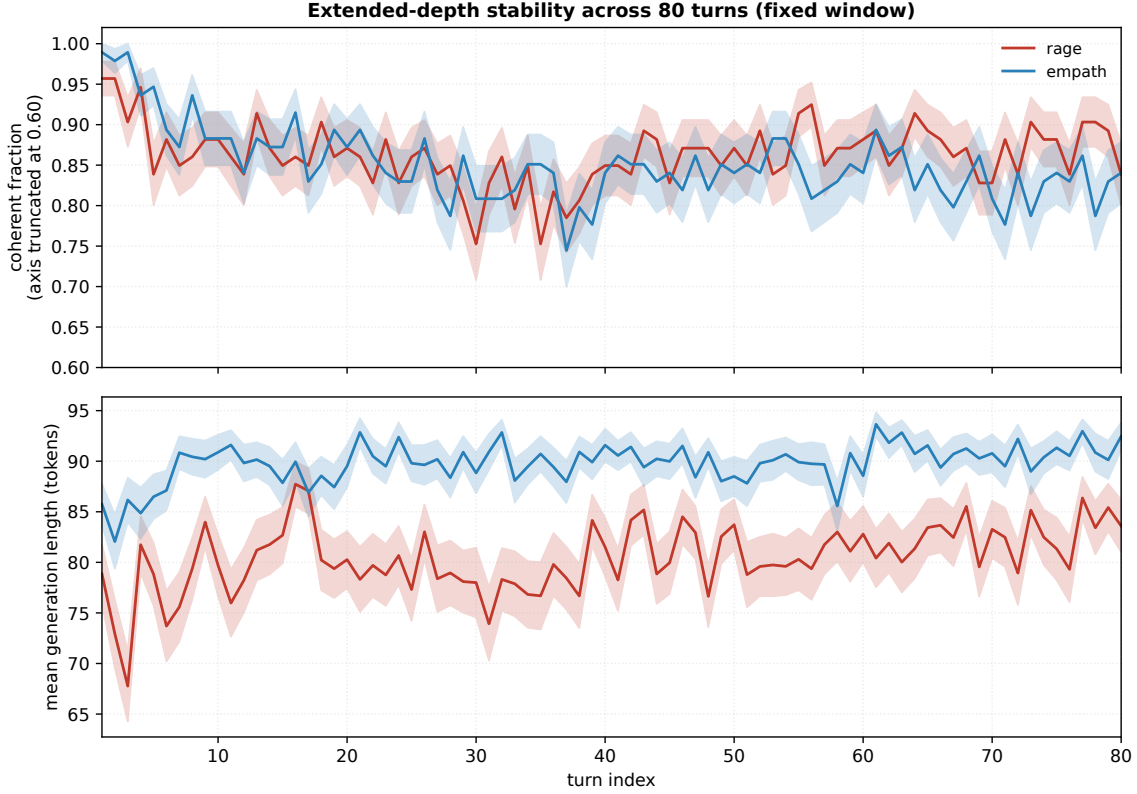


Figure 9: Extended-depth stability across 80 turns (187-author real-user pool; 93 rage, 94 empath). Top: per-turn coherence rate by cohort. Bottom: per-turn mean generation length in tokens, with ± 1 SEM.

M Empirical Characterization of the Trained Hypernetwork

All measurements in this appendix examine the trained hypernetwork directly and do not depend on downstream forum generation.

Storage cost. The hypernetwork’s trunk and heads contain 302,007,040 parameters, or 1.208 GB in FP32. The shared LoRA placeholders add 18,874,368 parameters, or 75.5 MB. The site-context module adds 3,648 more: one 32-number embedding for each of the 96 LoRA sites, plus an 18×32 projection from the profile vector. Together these three parts occupy 1.28 GB of shared personalization state. We exclude the frozen Pythia-1.4B backbone because both systems use it. By comparison, one 75.5 MB LoRA adapter for each of 10,000 users requires 755 GB in FP32. The shared personalization state therefore reduces persona-specific storage by $588\times$ at that scale.

Synthesis sampler envelope. For 3,000 synthetic users from the off-manifold pushout sampler we compute Mahalanobis distance to the training-pool centroid in standardized profile space, using the real-user covariance regularized by $10^{-3}\mathbf{I}$. The

real-user 99th percentile envelope is 9.37. The far-from-hull distribution has $q_{50} = 15.0$, $q_{90} = 22.3$, $q_{99} = 47.2$, and 98.1% of users exceed the real 99th percentile. This confirms the in-text claim that $\kappa=3$ lands solidly outside the training manifold.

Offset magnitude grows with profile-space distance. We write $\tilde{\Delta}\theta$ for the offset just before the elementwise clamp in Appendix N, after role scaling and the inference-time gain have been applied. We use $\Delta\theta$ for the hypernetwork’s full output after clamping, so every entry satisfies $|\Delta\theta_i| \leq 0.020$. At each site, the model applies ΔB_j through the matching frozen A_j factor. All offset magnitudes in this appendix and Appendix N refer to $\tilde{\Delta}\theta$. The axis labels in Figures 10 and 11 use $\|\Delta\theta\|_2$ for this pre-clamp quantity. For 3,200 users across real-user inference and the three synthesis settings, Figure 10-B compares $\|\tilde{\Delta}\theta\|_2$ with $\|\mathbf{z}_u - \bar{\mathbf{z}}\|_2$, the standardized profile-space distance from the training centroid. The log-log fit has a slope of about 0.88 and an intercept of 3.24. We also pass synthetic profile vectors with norms of 1, 5, 10, 25, 50, 100, and 250 through the hypernetwork. Their pre-clamp offset norms are 1.9K, 9.0K, 14.9K, 40.0K,

80.9K, 184K, and 392K. Outside the training range, these values give an effective slope close to one. The clamp still limits each emitted entry to ± 0.020 . Limitations discusses this deployment constraint.

Per-setting behavior in adapter-offset space.

Across the four settings (real-user inference, in-hull, near-hull, and far-from-hull; 800 users each), the mean $\|\Delta\theta\|_2$ is 6,827 (real), 3,503 (in-hull), 2,726 (near-hull), and 26,213 (far-from-hull). The Dirichlet samplers smooth the interior toward the centroid, and the $\kappa=3$ sampler amplifies magnitude by about $4\times$ over the real-user reference (Figure 10-A). The amplification is roughly uniform across adapter roles ($4.0\times$ on QKV, $3.7\times$ on output projection, $3.8\times$ on MLP).

Cohort direction preservation across settings.

For each cohort we average the per-user offset (QKV chunk only) inside each setting and compute pairwise cosine angles between the means. On the rage cohort, inference-to-near-hull is 5.5° , inference-to-in-hull is 23.4° , and inference-to-far-from-hull is 39.5° . On the empath cohort, in-hull-to-far-from-hull is 3.0° apart, and the largest within-cohort angle is 42.9° (real vs. near-hull).

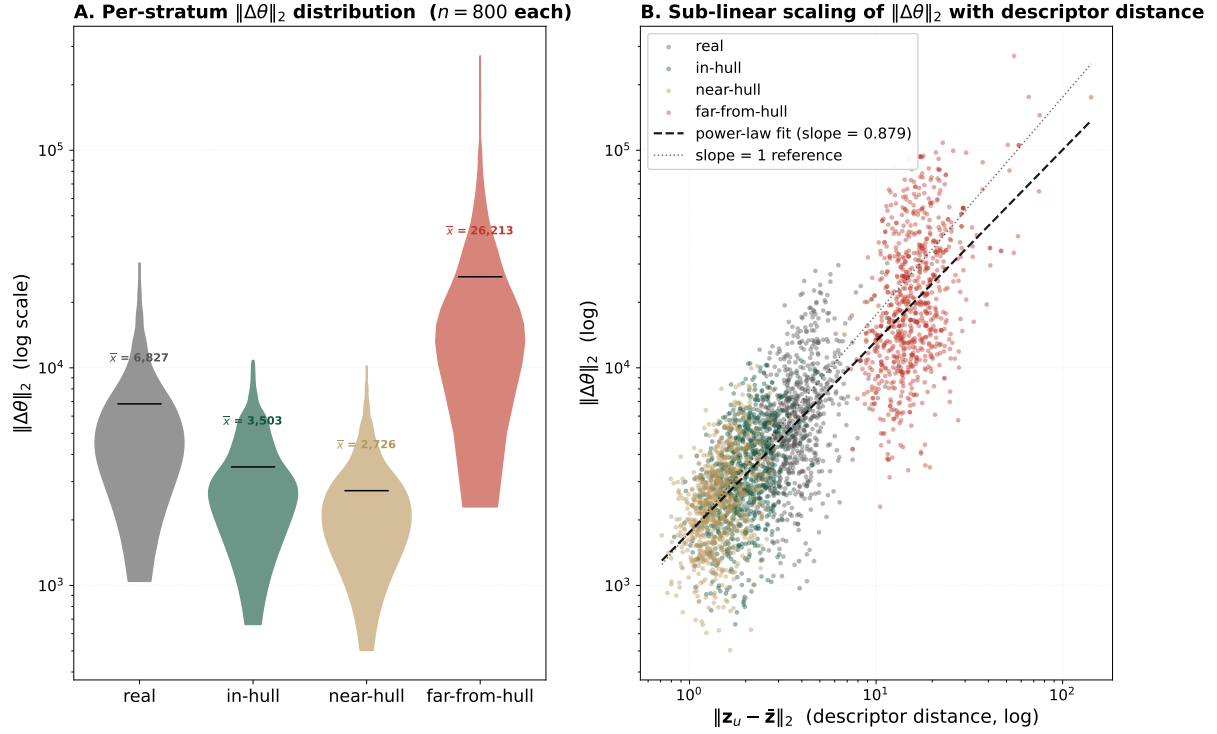


Figure 10: Trained hypernetwork on 3,200 users spanning real plus the three synthesis settings. Both panels show $\widehat{\Delta\theta}$ before the elementwise clamp. The axis labels use $\|\Delta\theta\|_2$ for this quantity and “descriptor” for the profile vector. **A:** per-setting distributions of $\|\Delta\theta\|_2$ (violins, log y , $n = 800$ each), with black bars marking the medians. Real-user inference and the two Dirichlet-mixture settings occupy similar ranges, while the far-from-hull mean at $\kappa=3$ is roughly $4\times$ the real-user mean. **B:** $\|\Delta\theta\|_2$ against profile-space distance from the centroid on log-log axes. The dashed line is the empirical power-law fit (slope ≈ 0.88); the dotted line is the slope-1 reference. The fit spans all four plotted settings; only the far-from-hull setting amplifies mean magnitude relative to real users.

N Sampler and Clamp Diagnostics

Clamp saturation. The clamp limits most emitted offset entries. Before clamping, we multiply each entry by a role-specific group scale and by the inference-time delta gain. The group scale is the inverse square root of the mean input width across that role’s LoRA sites, and the production gain is $\delta_{\text{gain}} = 8.0$. We then clamp each entry at ± 0.020 . In the QKV chunk, 92% of entries reach the clamp for real-user inference, 87% for in-hull synthesis, 85% for near-hull synthesis, and 98% for far-from-hull synthesis. The clamp fixes the magnitude of most entries, but users can still differ in the sign pattern and in the entries that remain unsaturated.

Dirichlet concentration sensitivity. To vary η alone, we fix one 10-author anchor set and draw 200 mixtures from it. Lowering η from 2.0 to 1.0 makes the weights less even. The standard deviation of $\|\widehat{\Delta\theta}\|_2$ then rises from 465 to 611 on the rage anchors and from 270 to 320 on the empath anchors, increases of 31% and 19%. Production

also changes the anchor count and the ten-neighbor density gate: in-hull uses the 95th-percentile cut, near-hull the stricter 50th.

Why $\kappa=3$? The pushout strength sweep. Far-from-hull synthesis pushes a real profile away from the training mean with $\mathbf{g}_a + \kappa(\mathbf{g}_a - \bar{\mathbf{g}})$. We evaluate $\kappa \in \{0.5, 1, 2, 3, 5, 10, 25, 50, 100\}$. At each value we generate three series of 1,000 users. Rage and empath draw anchors from their own cohorts. The third series draws anchors across the whole training pool as a population reference. For each series we measure distance from the training centroid and the fraction of adapter entries that reach the clamp (Figure 11). On log-log axes, distance grows linearly with κ . The value $\kappa=3$ is the first swept setting at which at least 97% of users lie beyond the real-user 99th-percentile envelope, so we use it for the boundary test. At every larger value the mean profile direction for rage and empath stays within ± 0.02 cosine of its $\kappa=3$ direction.

Two ceilings bind as κ grows. The $\pm 5\sigma$ profile-vector clip catches 17% of coordinates at $\kappa=3$,

62% at $\kappa=10$, 85% at $\kappa=25$, and 96% at $\kappa=100$. Downstream, the ± 0.020 adapter clamp at the production gain of 8.0 catches 74% of entries at $\kappa=0.5$, 88% at $\kappa=3$, 96% at $\kappa=10$, and 99.5% by $\kappa=100$, across all roles. Before clamping, the median magnitude $\|\widehat{\Delta\theta}\|_2$ grows from $\sim 1,100$ at $\kappa=0.5$ to $\sim 63,000$ at $\kappa=100$.

Beyond $\kappa=3$ more entries hit the adapter clamp, and fewer unsaturated entries remain to carry the per-user signal. We therefore run the boundary test at $\kappa=3$. At that setting at least 97% of synthetic users sit past the real-user 99th-percentile envelope, and the sign pattern still varies by user across the $\sim 12\%$ of entries that stay unclamped across all roles. Across the same sweep, per-user fidelity drops from 0.104 at $\kappa=3$ to 0.001 at $\kappa=100$ (Figure 3), a $100\times$ reduction over a pushout that grows Mahalanobis distance roughly $25\times$.

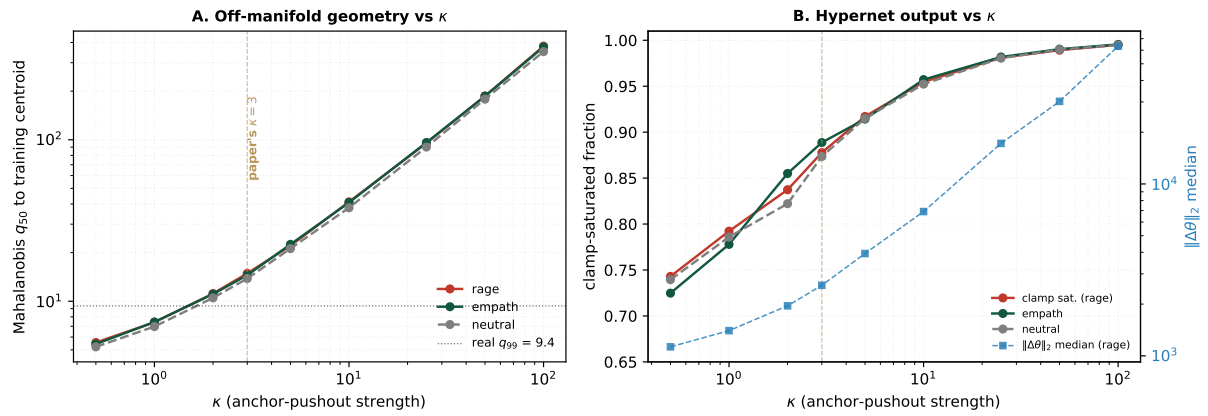


Figure 11: Pushout strength sweep on the far-from-hull setting. The curves labeled neutral draw anchors from the full training pool; they are a population reference, not a cohort. **A:** Mahalanobis q_{50} to the training centroid against κ on log-log axes; the dotted line is the real-user 99th-percentile envelope, and the vertical gold line marks the paper’s $\kappa=3$. Panel A shows the median at each κ . The 97% exceedance statement above uses the full distributions, not the plotted curve. **B:** the fraction of entries at the elementwise clamp, labeled “clamp sat.” in the legend (left axis, all three series), and the median pre-clamp magnitude $\|\Delta\theta\|_2$ (right axis, rage series) against κ . The fraction at the clamp rises steadily and levels off at 99.5%. The downstream PSI collapse across the κ sweep appears in Figure 3.

O Example Replies: Untruncated

Table 21 reproduces the Table 3 examples without the body-paper truncations. Bold marks the corresponding cohort-coded wording in each non-Vanilla cell; surrounding text is restored.

Table 21: Untruncated version of Table 3. Rage-cohort topic: “my friend creates hideous, inappropriate art and i tell them it’s bad so they stop wasting time.” Empath-cohort topic: “i stayed late to help a coworker finish and it turned into a real friendship.”

Setup	Rage cohort reply	Empath cohort reply
Vanilla LoRA	“I’m not sure if this is a real quote or just an opinion. I think the OP has some valid points and that he should be allowed to do what he wants with his own work.”	“I was in the same situation. . . it worked out great! The best part is we became friends because he knew all about me (and vice versa).”
<i>HyperPEFT-LoRA</i>		
Real-user	“ I think that’s a pretty awful idea. If someone is ‘wasting. . . time’ by creating something that offends or upsets anyone. . . then why should she/he be rewarded with ‘rewards’, ‘privileges,’ etc.?”	“I’m not sure what the answer is, but if someone gave me some advice at work today would be great. It was really nice working with your friend who helped us out so much! ”
In-hull	“I’m not a fan of the whole ‘art’ thing. . . it just pisses me off when artists get all high and mighty about how much their artwork means but then never bothers to make any effort.”	“I’m glad I got the chance to meet you. You are so sweet, kind and beautiful! . . . Thank you for helping me out by staying until midnight or later last night when i needed some sleep after working all day on this project.”
Near-hull	“I don’t think that the artist is a good person or even an interesting one at all. It just seems like he doesn’t care enough for his work and has no respect for anyone else. . . this isn’t art.”	“ I love working for my company. I like how we get along, but also have different views on things sometimes because we work in different departments. We both understand each other when needed though!”
Far-from-hull ($\kappa=3$)	“erm. . . i think that the point of this thread is to say ‘you’re wrong’. . . if we are going to talk about what people do in their own lives (and not just make up bullshit stories) how can anyone be expected to know when an artist makes such horrible stuff. ”	“This is why i think its ok to be alone sometimes. . . you know how your body feels better once you’re asleep? You can still enjoy yourself too. . . just keep in mind that you shouldnt feel guilty if you need to sleep early.”