Personas on Demand: Hypernetwork Synthesis of Unseen Users

Anonymous ACL submission

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. We test this using a frozen LLM driven by a hypernetwork, generating synthetic users from the real-user manifold and evaluating them on behavioral measures to determine whether synthetic users match the cohort patterns their profiles imply. We show that synthetic users reproduce persona 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 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.

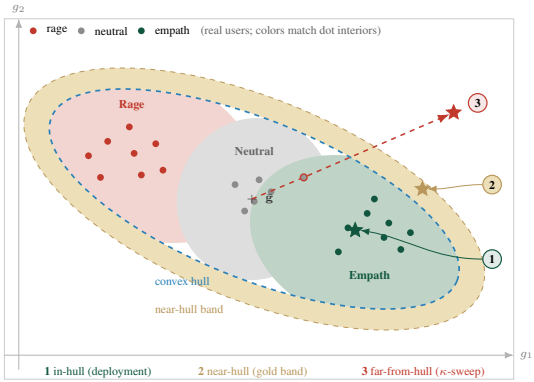

Figure 1: We generate three kinds of synthetic users. **(1)** *In-hull* users are drawn from the dense interior of the real-user manifold. **(2)** *Near-hull* users still sit inside the manifold but are drawn with looser sampling and a looser density gate, so they serve as a boundary-style robustness test rather than a literal hull-edge sample. **(3)** *Far-from-hull* users are pushed deliberately outside the manifold. For each synthetic user, the trained hypernetwork turns the user's profile vector into a per-user adapter offset, exactly as it does for a real user from training.

1 Introduction

A central goal of social-media analysis is to understand how individuals and communities behave online. One way to study that behavior is to simulate it: build models of individuals in a community and create experimental environments where those models can interact to simulate the community. Researchers increasingly turn to large language models for this purpose, since a single model can, in principle, model the language of varied individuals. However, current LLM implementations struggle to do this faithfully at population scale, where a simulator must represent thousands of distinct users without collapsing them into a single generic voice.

The straightforward way to personalize a language model is to train and store one fine-tuned copy per user, but this fails to scale: storage grows linearly with the number of users and, at inference, incurs loading a different fine-tuned model at every persona change (Houlsby et al., 2019; Hu et al., 2022). Prompting one shared model with several personas avoids that cost but does not hold up over a conversation: persona consistency is a known weak spot, and during a conversation, the model drifts back toward a default voice as the exchange continues (Tseng et al., 2024). Prompt-based persona descriptions and retrieval over per-user history reduce some of these problems but introduce others: prompts dilute stylistic signatures (Jiang et al., 2024; Wang et al., 2024), and retrieval pipelines shift the burden to embedding and index infrastructure (Park and Lee, 2024).

Hypernetworks offer a different path. A single shared hypernetwork reads a compact user profile and emits a small adapter on demand. Stated plainly, the persona becomes a short numeric vector rather than a stored model checkpoint (Ha et al., 2017; Walsh et al., 2026). Prior work shows that this kind of hypernetwork can recreate users seen during training. What remains open is the more useful version: can the same mechanism generate behaviorally consistent new users who were never observed? Without that capacity, a population-scale simulator cannot grow beyond its training data.

The research question we seek to answer is, *can a hypernetwork trained to rebuild observed authors synthesize behaviorally consistent unseen users, and where does that capability break down?* Three sub-questions organize the rest of the paper. First, do synthetic users behave the way their profiles say they should? Second, do synthetic users behave as consistently as real users from the training corpus? Third, do synthetic users remain consistent across many generations rather than drifting back toward the base model's default voice? We label these **H1**, **H2**, and **H3** in the rest of the paper. Each is tested on a set of nine behavioral measures that cover emotional, stylistic, and timing signals.

Contributions. In this manuscript, we contribute four things. First, a *synthesis engine* that turns a short user profile into a per-user adapter on demand, running on a frozen Pythia-1.4B backbone. Second, a set of nine behavioral measures to assess whether each new user's generated text matches the cohort patterns implied by their profile. Third, evidence that synthetic users in our deployment setting recover the emotional patterns we see in the training set's real users, and that those patterns weaken as user vectors move away from the observed distribution. Fourth, a stability check that tracks how steady each user's responses remain across many turns of conversation.

2 Related Work

What we mean by "persona." A *persona*, as we use the term, is an author's stable, text-derived way of writing: how they communicate rather than what they believe (Wegmann et al., 2022; Patel et al., 2023; Pennebaker et al., 2001; Danescu-Niculescu-Mizil et al., 2013). To quantify persona, the cohort labels we use later in the paper come from stylometric and emotional distributions.

Personalized language generation. Personalization for generative models is generally done through one of three approaches: prompt-based profiles (Zhang et al., 2018; Li et al., 2016; Jiang et al., 2024), retrieval over user histories (Salemi et al., 2024; Tan et al., 2024b; Park and Lee, 2024), and per-user adapter weights inserted into the backbone (Tan et al., 2024a,b; Tseng et al., 2024). The three approaches differ in where per-user state lives at inference time: re-encoded from text for every reply, looked up in an embedding index, or cached inside the model itself. Per-user adapter weights emitted by a hypernetwork (Ha et al., 2017; Walsh et al., 2026) sit in the third family and avoid the prompt-dilution and retrieval-infrastructure costs of the other approaches.

User-level hypernetworks. A hypernetwork (Ha et al., 2017) reads an input vector and emits the weights for another network. In effect, a hypernetwork influences the behavior of another downstream model. In most prior work, the hypernetwork is auxiliary: another network is trained alongside it and carries the main training load, usually at the task level rather than the user level (Lv et al., 2024; Üstün et al., 2022; Charakorn et al., 2025). Two prior works, instead, run at the user level. Profile-to-PEFT (Tan et al., 2025) feeds long text profiles and a retrieval store through a 4B-parameter sentence encoder, then emits per-user adapter weights from the result; the encoder is the bulk of the trainable footprint. HyperPEFTNet (Walsh et al., 2026) reverses the auxiliary framing: the hypernetwork is the only trainable component, the backbone stays frozen, and the input is a fixed-length vector of stylometric numbers per author on a smaller backbone. Neither prior user-level study has tested whether the same mechanism can extend to unseen users.

Persona synthesis and interpolation. Most persona-aware models reproduce users seen during training rather than generating new ones (Zhang et al., 2018; Lucy and Bamman, 2021; Sawatphol et al., 2024). Some agent-simulation work builds agent populations from sociodemographic templates without checking whether the resulting behavior is faithful (Murić et al., 2022). The closest related work blends numerical summaries of how people write: Patel et al. (2023) learn a representation that supports smooth blending between users while Wegmann et al. (2022) test whether such representations describe people who weren't

in the training data.

Evaluating personas. Three instrument families dominate persona evaluation. The first measures how closely the text generated for an author resembles a real, held-out reply by that same author, typically using a learned similarity model (Wegmann et al., 2022; Patel et al., 2023); this is precise but needs a ground-truth reply to compare against, which is not available when the model is generating freely. The second counts hand-defined surface features in the text, such as how often the author uses profanity, first-person pronouns, or questions (Danescu-Niculescu-Mizil et al., 2013; Pennebaker et al., 2001); these are easy to compute and interpret, but they describe surface form rather than meaning. The third passes the text through a pretrained classifier that predicts an emotional or attitudinal label (Demszky et al., 2020; Hutto and Gilbert, 2014); this method reads emotional tone directly at the cost of inheriting whatever biases live in the classifier's training data.

For comparing two groups of users on these instruments, the field typically uses a standardized difference in means: Cohen's d, or the bias-corrected variant Hedges' g (Cohen, 1988; Hedges, 1981), paired with bootstrap confidence intervals to capture sampling uncertainty (Efron, 1979; Deen and de Rooij, 2020). Drift in a generated persona across many turns has been studied in two settings: personas described to the model through a prompt (Jiang et al., 2024; Wang et al., 2024), and personas tracked as state variables in agent-based social simulations (Park et al., 2023; Zhang et al., 2018). In both cases, the persona is built from natural-language input rather than per-user model weights, so neither characterizes drift directly when the persona lives in trained parameters.

3 Methods

3.1 Data

Both models train on the same English-language discussion-forum corpus drawn from Reddit. The corpus contains 10,000 authors, each with at least 300 top-level threads, and we keep every thread together with its reply, so training and evaluation use the same format. The text is cleaned for platform artifacts and split into each author at the thread level. At evaluation time, the model writes a complete reply from the thread context, with no user-specific reference text to copy from. Reddit's per-author depth lets us build stable cohort labels at the 10,000-author scale (§3.5). Because the same pool of users supplies both the real-user evaluation in §4.1 and the three synthesis settings, those four results are directly comparable.

3.2 Models

Two models trained on the same data are used in every comparison in the paper and differ in one aspect: whether a per-user signal is included in the model. Both share the same initialization, backbone, and training schedule.

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.

HyperPEFT-LoRA (treatment). The LoRA matrices, as well as the entire backbone, stay frozen; the hypernetwork H_φ is the only trainable component. For each user, H_φ reads an 18-dimensional profile vector and produces a small per-user offset for every LoRA site. Training minimizes $\mathcal{L} = \mathrm{CE}_{\text{reply}} + \lambda_\Delta \|\Delta\theta\|_2^2$, cross-entropy on the reply tokens, plus a regularizer that keeps the offsets small. The backbone, LoRA geometry, precision, and optimizer schedule are identical to the vanilla-LoRA baseline, so any difference between the two models reflects the per-user signal. Full architecture and training spec are in Appendix B.

At inference, we remove a learned per-cohort direction from the residual stream of the upper decoder layers. This mitigates training biases that can otherwise prevent the model from generating certain personas, e.g., the rage cohort (Arditi et al., 2024). How we picked the operating point is in Appendix F.

For ablation, we also report the *control*: the trained hypernetwork is loaded but the per-user input vector is zeroed before it reaches the hypernetwork. Everything else about HyperPEFT-LoRA stays the same, *including the identical residual-stream steering step*; only the per-user signal is missing. This isolates the per-user channel from the steering step: differences between HyperPEFT-LoRA (HP-LoRA) and the control in any downstream measure attribute cleanly to the per-user signal, not to the steering intervention. The control is also the persona-free floor for the trained model and rules out the possibility that the hypernetwork is merely acting as a regularizer.

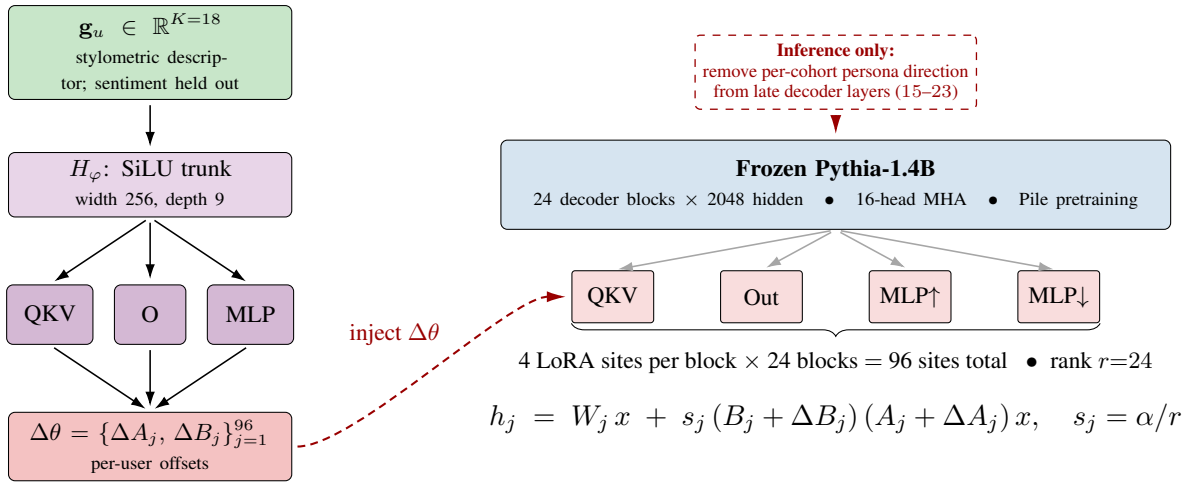

Figure 2: HyperPEFT-LoRA on a frozen Pythia-1.4B backbone. The hypernetwork H_φ injects per-user offsets into every LoRA site; an inference-time steering step (top box) removes per-cohort persona directions from the upper decoder layers. Details: §3.2 and Appendix A.

3.3 Profile Vector

The hypernetwork's only input is a fixed-length profile vector $\mathbf{g}_u \in \mathbb{R}^{18}$ summarizing each author's posting history. We use two families of features. The first is surface writing style: capitalization rate, formality, lexical diversity, readability. The second is platform behavior: posting rhythm, reply-delay variability, subreddit concentration. The full list, with definitions, is provided in Appendix E. Every feature is z-scored against training statistics and clipped at $\pm 5\sigma$.

We deliberately leave sentiment, personality proxies, and emotion scores out of the profile vector. These dimensions are used by our out-of-band measures later in the paper. If we included them in $\mathbf{g}_u$, any measure effect we measured would be circular: the hypernetwork would simply be reading the measure label off its own input.

We started with 57 candidate features and trimmed the profile vector to 18 by repeatedly removing the least-useful one and remeasuring via backwards greedy elimination. The result was not a clean trend: the delta-separation score peaks at 18 features, reaching about 7.7× the delta-separation score at the original 57-feature starting point, while the model's prediction loss on held-out text barely changed at any setting we tried (Appendix E). The 18-feature profile vector therefore gives us the strongest persona signal at no cost to how well the model writes.

3.4 The Synthesis Engine

The synthesis engine produces unseen users at inference time. 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.

Two ways exist to create a new user from a population of real ones: *interpolation*, which samples a new profile within the range of users we already have, and *extrapolation*, which pushes past them into profiles no real user occupies. Interpolation matters here because profile space is continuous: a sampled point can land between users from different cohorts rather than inside any single one, and a chain of such points can wander across cohort regions without leaving the real-user range. Extrapolation leaves it entirely.

Two of the three settings interpolate. For each new user, the sampler picks a few real users with similar profiles, draws Dirichlet weights over them (non-negative numbers that sum to one), and samples a new profile near that weighted center, with a small jitter drawn from the local spread of those neighbors. The new point is one no real user occupies, and the weighted center keeps it within the range of those neighbors. **In-hull** synthesis is our deployment setting: it picks the five most similar

users, weights one heavily, and only keeps samples that fall in the densely-populated middle of the profile space. **Near-hull** synthesis loosens both controls: it picks twelve neighbors, weights them more evenly, and keeps samples from sparser parts of the profile space. Both stay inside the real-user range; "near-hull" refers to the looser sampling, not to where in the profile space the new users land (Appendix N).

The third setting, **far-from-hull**, extrapolates instead. We pick a real user and push their profile further from the population average, walking outward in the same direction the user already sits relative to that average. The result lands outside the real-user range entirely, at a distance set by a strength parameter κ. The user's profile is $\mathbf{g}_a$ and the average of all real-user profiles is $\bar{\mathbf{g}}$; the new profile is $\mathbf{g}' = \mathbf{g}_a + \kappa(\mathbf{g}_a - \bar{\mathbf{g}})$, where the difference $\mathbf{g}_a - \bar{\mathbf{g}}$ points from the average toward the user, and κ scales how far we walk in that direction. We sweep κ from 3 (just outside the range) to 100 (very far outside); at κ=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 when the profile is genuinely beyond anything in training. Across all three settings, the hypernetwork's offset for a new user points in roughly the same direction it would for a real user in the same cohort; what changes is its size, which grows as the profile sits further from the population average.

3.5 Behavioral Measures

We score generated text along six behavioral measures, plus three auxiliary emotion composites for the multi-turn analysis. The six split into two groups based on whether the measure's defining feature is part of the hypernetwork's input vector $\mathbf{g}_u$.

Out-of-band measures are dimensions the profile vector does not encode. Any effect we measure on them cannot be explained by the hypernetwork reading the label off its own input. We use three:

- **Sentiment.** Rage vs. empath cohorts from GoEmotions polarity quintiles, scored by a frozen RoBERTa classifier (Demszky et al., 2020).
- **Politeness.** Tails of each author's profanity rate; a long-established stylometric marker (Danescu-Niculescu-Mizil et al., 2013; Pennebaker et al., 2001; Argamon et al., 2009).
- **Self-focus.** Tails of first-person pronoun rate; the LIWC marker for attentional focus on the self (Pennebaker et al., 2001; Argamon et al., 2009).

In-band measures are positive controls: their defining features are in the profile vector, so a working hypernetwork should produce strong effects on them. We use three:

- **Curiosity.** Question-rate tails.
- **Expressiveness.** Harmonic mean of capitalization and punctuation rates (Argamon et al., 2009; Wegmann et al., 2022).
- **Tempo.** Reply-delay mean.

Scorer grounding. The behavioral measures are anchored in three published, peer-reviewed instruments rather than project-defined ones: GoEmotions (Demszky et al., 2020) for emotion-based axes (sentiment, anxiety, warmth, hostility), the LIWC stylometric tradition (Pennebaker et al., 2001; Argamon et al., 2009) for politeness and self-focus, and Danescu-Niculescu-Mizil et al. (2013) for the politeness-feature tradition. We use these scorers without retraining; the cohort labels themselves are derived from corpus-level GoEmotions quintiles fixed at training time, so the same instruments anchor both the cohort targets and the realized measurement. The Persona Stability Index (PSI) defined in §3.6 is a fixed transform of these scorer outputs, not a new model.

For each cohort comparison we report Hedges' g (Hedges, 1981) with bootstrap 95% confidence intervals; further details comprising how we drew the rage/empath/polite/etc. cohort lines, how we built the user-level labels, and a supporting style-similarity score against held-out reference replies are all in Appendix G.

3.6 Persona Stability Across Turns

A persona that holds for one reply but decays by turn five is useless for community simulation. We test persona durability using the nine behavioral measures from §3.5 (three out-of-band, three in-band controls, three emotion composites: anxiety, warmth, hostility) plus two summary indices that capture persona stability and compositionality. Details for the nine behavioral measures are in Appendix G; PSI and PCI derivations are in Appendix I.

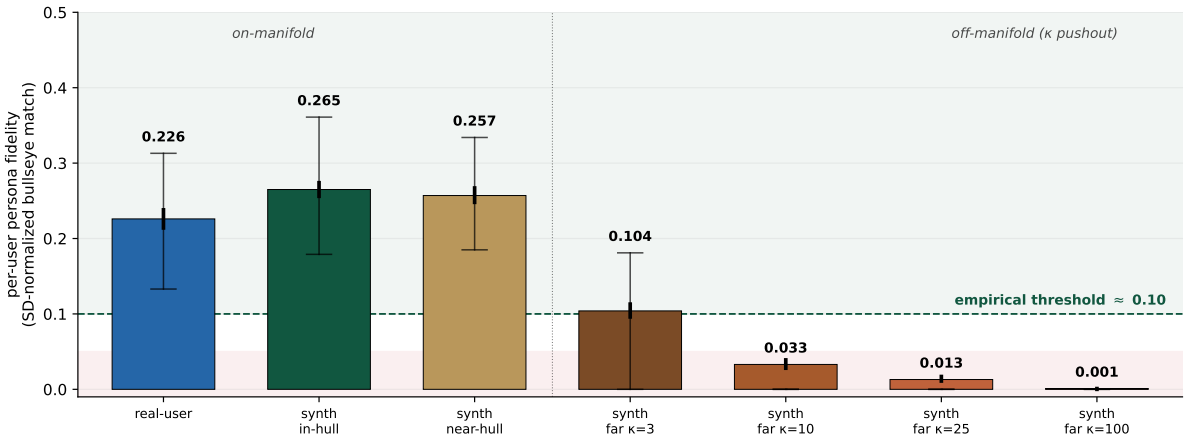

Figure 3: How close each user's generated behavior sits to their target profile, by setting. Higher is better. The three on-manifold settings (real, in-hull, near-hull) cluster around 0.22 to 0.27. The four off-manifold settings drop steeply as κ grows. The dashed green line at 0.10 is the cut where the persona signal noticeably weakens; details in Appendix I.5.

The **Persona Stability Index** (PSI) measures how close each user's generated behavior sits to their intended profile. For every coherent reply we take the absolute residual between realized and intended polar values on each of the 8 signature measures, divide by the real-user expected standard deviation on that measure, average across measures, and convert to a $[0, 1]$ score:

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

where $\delta_d = |\hat{p}_d - p^*_d|/\sigma^{\text{real}}_d$ is the standardized residual on measure d. PSI for a user is the mean of *match* across that user's coherent replies. Per-dim standardization is needed because the 8 signature measures are on different scales (z-scored stylometric ratios versus raw GoEmotions probabilities); without it, one measure (question-rate curiosity) dominates the residual when synthetic users are pushed off-manifold. The threshold 0.10 we adopt for "carries persona signal" is not a preregistered guess: it is the empirical D3-to-D4 inflection of cohort-agreement rate against PSI on the pooled production set ($n = 12{,}451$ user-cell rows; Appendix I.5, file `psi_threshold_derivation.csv`). Full derivations are in Appendix I.

The **Persona Compositionality Index** (PCI) asks whether users who realize more of their intended cohort directions also produce a more faithful overall signature. We partition synthetic users by the number of held-out measures whose realized polar value matches the user's expected sign with magnitude at least 0.05, then report the mean direction match between each user's realized and intended behavioral signatures (defined in Appendix I.2) inside each partition. A flat curve across partitions would mean that stacking more matching traits does not help; a rising curve means realizing more cohort directions is associated with higher overall fidelity. Our current sampler does not accept per-user trait targets, so the target-driven version of PCI is reserved for future work; the observational variant we report here uses realized agreement counts instead of imposed specifications. The numerical results appear in Appendix J, Table 13.

No existing multi-turn persona benchmark fits this setting, so we built our own. Prior work measures personas defined by prompts (Jiang et al., 2024; Wang et al., 2024), by tracking state variables in agent simulations (Park et al., 2023), or by editing activations at runtime (Lake and Baroni, 2018); none targets per-user weight offsets emitted by a hypernetwork. PSI and PCI are the project-defined measures we use here; whether they generalize beyond this corpus is an open question.

Using these metrics, we run two tests: one to see if a synthetic user's persona holds up across a short thread, and another to see if it holds up across 80 consecutive generated turns. The short-thread test runs across all author pools and is summarized in Table 12. The 80-turn test uses a separate pool of 559 authors drawn from the rage and empath cohorts. Both tests use the same generation settings.

Table 1: Rage-vs-empath Hedges' g on the three out-of-band measures, real-user forum. Higher is better; bold marks the winning setting on each row.

Measure	HP-LoRA	Vanilla	Control
Sentiment	$+\mathbf{0.31}$	0.00	$+0.02$
Politeness	$+\mathbf{0.19}$	-0.01	$+0.08$
Self-focus	$+\mathbf{0.32}$	$+0.08$	0.00

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 1 shows the rage-versus-empath difference on three out-of-band measures for real users from the training corpus, side by side for HyperPEFT-LoRA, vanilla LoRA, and the control (HyperPEFT-LoRA with the per-user input vector zeroed out, so it should not produce much individualized user signal). Sentiment defines the two cohorts while politeness and self-focus are independent stylometric labels the hypernetwork never sees. Positive g means the cohorts came out in the predicted direction, negative means backwards, and near-zero means no difference.

HyperPEFT-LoRA carries all three measures in the correct direction: sentiment $+0.31$, politeness $+0.19$, self-focus $+0.32$. Politeness and self-focus also separate cleanly between the two groups, the result we need for the synthesis tests below.

Vanilla LoRA and the control both exhibit what we call persona collapse. Both sit within $|g| \leq 0.1$ of zero across all three rows. Neither model has an inference-time channel that separates rage from empath users. The six additional behavioral measures we track per user are in Appendix K.

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

Table 2 shows the rage-vs-empath effect size on the three out-of-band measures with 95% bootstrap CIs, for every inference setting. Positive numbers mean the two user groups came apart in the direction the cohort split predicts. Numbers near zero mean the model produced no group difference; CIs that exclude zero indicate the difference is significant at the 5% level.

A HyperPEFT-LoRA setting wins every row. On sentiment, the cohort-defining axis, every synthesis setting reads $+0.40$ or higher, with in-hull and near-hull both reaching $+0.51$ and real-user inference at $+0.31$. On politeness, real-user inference takes the top number at $+0.19$, with the three synthesis settings clustered behind it in the $+0.11$ to $+0.18$ band. On self-focus, real-user inference comes in at $+0.32$; in-hull and near-hull synthesis reverse the first-person pronoun rate, while far-from-hull collapses toward zero. This reversal is a known stylometric-ratio side effect of the steering step (Appendix H). Vanilla LoRA and the control sit within one-tenth of zero on every row.

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 every (setup, cohort) cell are in Table 3.

4.3 H2: Synthesized vs. Real

H2 asks how closely synthetic users approach the target-match behavior we observe on training-set real users. Per-cohort PSI (Figure 3, split out by cohort) lands real users, in-hull synth, and near-hull synth in the same band on both rage and empath (0.20 to 0.27), with empath in-hull and near-hull at or slightly above real-user inference; far-from-hull falls away as κ grows. Per-cohort values and the measure-by-measure decomposition are in Appendix K.

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 the Persona Stability Index, PSI (§3.6): near 1 means on target, near 0 means the generated text drifts back toward the base model's default voice.

Figure 3 shows the mean PSI per setting. Real users, in-hull synth, and near-hull synth cluster

Table 2: Rage-vs-empath Hedges' g on the three out-of-band measures with 95% bootstrap CIs (per-reply resampling, 1,000 iterations, seed 142). Bold marks the winning setting on each row. Full nine-measure version in Appendix H.

Measure	Vanilla	Control	Recon	In-hull	Near-hull	Far-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]

Table 3: Example replies by setup and cohort. Bold marks the strongest cohort-coded phrase per non-Vanilla cell. Untruncated replies in Appendix O.

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."
HyperPEFT-LoRA		
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 synth	"**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 synth	"**has no respect for anyone else**. . . this isn't art."	"**I love working for my company.** . . . we both understand each other when needed."
Far-hull (κ=3)	"**bullshit stories**. . . when an artist makes such **horrible stuff**."	"**shouldnt feel guilty** if you need to sleep early."

between 0.226 and 0.265. PSI drops as we increase the κ-distance: 0.104 at κ=3, 0.033 at κ=10, 0.013 at κ=25, 0.001 at κ=100. The empirical 0.10 threshold (Appendix I.5) sits between κ=3 and κ=10, exactly where the persona signal weakens.

Two extra checks back this up: no single behavioral measure drifts faster than the others (Appendix I.6), and the generated text stays readable across 80 consecutive turns (Appendix L).

5 Conclusion

A hypernetwork trained to rebuild known users can also synthesize new ones, inside a described range. 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 users, in-hull synth, and near-hull synth (§4.4, Figure 3); across an 80-turn forum simulation, coherence stays at ≥ 0.995 and per-dimension drift slopes stay three orders of magnitude below the pre-specified $|\bar{\beta}_d| < 0.05$ threshold (Appendix L).
- **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 2). The full nine-measure decomposition (Appendix H) shows the same broad separation pattern on the auxiliary emotion measures (warmth, hostility) and direction-consistent lifts on politeness, with known stylometric-ratio reversals on self-focus and expressiveness that we characterize transparently.
- **Synthetic users built to look like real users track real-user behavior closely.** On the match between profile and generated text, in-hull and near-hull synthetic users land in the same per-user-PSI band as training-set real users on both cohorts (§4.3, Appendix K). Push synthetic users far enough outside the real-user range and the match falls steadily.

A hypernetwork serves a population of users on demand without per-user storage or per-persona retraining, as long as the new synthetic users come from the profile space the training corpus covers. Outside that space the persona signal falls off steadily (Figure 3; example replies in Table 3).

Limitations

The findings here come from one hypernetwork, one backbone, one corpus, and one language. We test HyperPEFT-LoRA on Pythia-1.4B, fit it on

Reddit discussion threads, and evaluate it in English. Whether the synthesis story generalizes to other backbones, other genres of online discussion, or other languages is an open question. Details on the scoring instruments, e.g., which scorer we used for sentiment, how it agrees with other scorers, and why we did not run a human-perception study, are in Appendix G.

The cohort labels are defined on the sentiment axis, so the corpus-level cohort split on other measures is weaker than the split on sentiment itself. Two users can both be in the rage cohort and write at completely different lengths, ask completely different numbers of questions, and use completely different amounts of capitalization. The hypernetwork still carries every non-sentiment measure in the correct direction on the training-set real-user reference (Table 8); the absolute magnitudes are bounded by what the cohort labels predict at the population level. The same fact limits how much we can say about pairs of traits at once: combinations like "empath and hostile at the same time" almost never show up in real users, so there are not enough of those users in the corpus to estimate that joint cell tightly.

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, but the cap has a side effect: once the synthetic user is far enough outside the range, more and more of the offset entries are pinned against the cap, and the per-user signal lives in fewer and fewer unpinned entries. The example replies in Table 3 show what this looks like in the generated text.

Our forum simulation is the open-ended-reply genre only. Threading depth, rate limiting, moderation, and session drift are out of scope.

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 are also only as good as the per-user input vectors and the nearest-neighbor pool they are drawn from: a sparse or skewed neighborhood would carry straight into the synthetic users.

Ethical Considerations

A model that writes in a specific person's voice can be used to impersonate that person, and it can also be used to flood a discussion with believable but fake users. Both risks grow as the model gets better at the task. Three properties of this paper push back against those risks. First, the per-user input vector is short and stylometric. It does not encode personal facts about the user. Second, synthetic users are sampled as new, previously-unobserved profiles from the neighborhood of real users, not built from or aimed at any one named person. Third, the released artifacts are limited to per-user feature vectors, cohort labels, and summary numbers. We do not release the comment text we trained on, account names, or any model 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.

Community simulation has two sides. The same tool that helps a researcher study how a forum behaves can also help someone fake one. We frame this paper as a measurement of how faithful synthesis is and where it breaks down, not as a recipe for running a fake community. The choice of evaluation instruments also carries cultural bias. GoEmotions, VADER, and SST-2 all draw on Western emotion categories and English text.

References

Andy Arditi, 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)*. Establishes that refusal behavior in instruction-tuned LLMs is mediated by a single residual-stream direction accumulating through middle-to-late layers; ablating that direction throughout the network removes refusals while replacing the lm_head alone does not.

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.

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*. PMLR.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

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.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR. ArXiv:1711.00350.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, CA, USA. 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, Fanchao 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 simulacra of human behavior. In *Proceedings of the 36th ACM Symposium on User Interface Software and Technology (UIST 2023)*.

Seong-Il Park and Jay-Yoon Lee. 2024. Toward robust RALMs: Revealing the impact of imperfect retrieval for RALMs. *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.

Thanawin Sawatphol, Sureesak Udomcharoenchaikit, and Sarunya 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. 2025. Instant personalized large language model adaptation via hypernetwork. *arXiv preprint arXiv:2510.16282*.

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.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. 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 4 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.

B Training Specifications

Table 5 lays out the architecture and training-budget specifics for both models. The two systems share the frozen Pythia-1.4B backbone, the LoRA-site geometry, the precision, and the optimizer schedule. They differ only in whether the LoRA matrices are fine-tuned (vanilla LoRA) or held

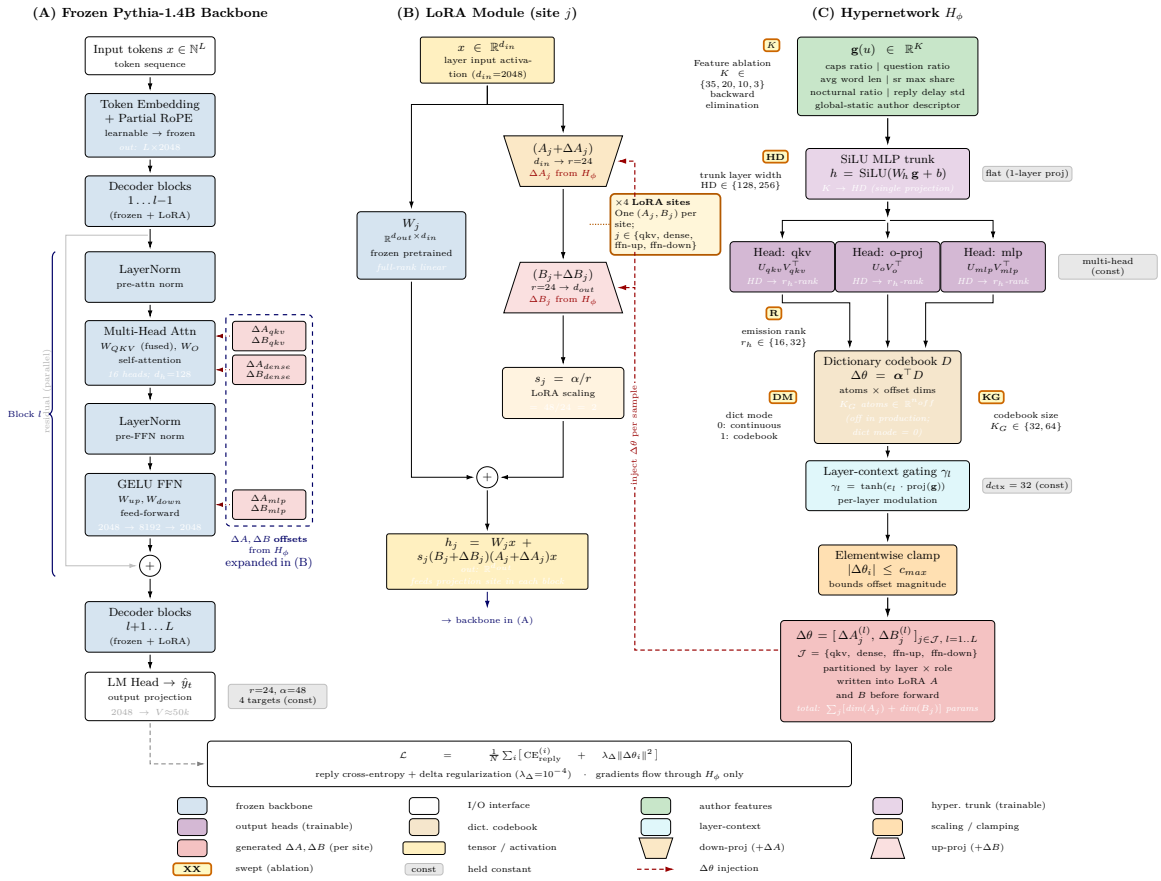

Figure 4: Component-level rendering of the Pythia-1.4B HyperPEFT-LoRA architecture (companion to main-paper Figure 2). **(A)** The frozen Pythia-1.4B decoder stack with residual block l expanded; LoRA is attached at four sites per block (qkv, dense, ffn-up, ffn-down). **(B)** One LoRA module: the hypernetwork-emitted offsets ΔA_j and ΔB_j are added to the frozen A_j, B_j pair before the scaled low-rank residual joins the frozen W_j. **(C)** The hypernetwork consumes the descriptor $\mathbf{g}_u$, projects through a SiLU trunk into a shared hidden dimension, factors through three role-specific output heads, applies layer-context gating, and clamps elementwise to produce $\Delta\theta$. Sweep markers in orange flag ablation axes; (const) tags identify hyperparameters held fixed.

Table 4: 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
Safety RLHF	none (Pile)	none (Pile)
LoRA sites	36 (QKV × 12)	96 (QKV+O+MLP$_{\text{up,down}}$ × 24)
LoRA rank r	32	24
LoRA α	32	48
Injection	A and B	A and B
Profile K	6	18
Loss	CE	CE + L_2 (λ_Δ=10^{-4})

Table 5: 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	—	SiLU MLP, width 256, depth 9
Heads	—	3 (QKV, O, MLP)
Profile K	—	18
Loss	CE	CE $+\lambda_\Delta \|\Delta\theta\|_2^2$
λ_Δ	—	10^{-4}
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; α=1.0; layers 15–23
Seed	142	142

fixed under a trained hypernetwork (HyperPEFT-LoRA).

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

C Artifacts, Licenses, and Reproducibility

Models and code. The backbone is Pythia-1.4B from EleutherAI, released under the Apache 2.0 license, and we keep it frozen. The trained hypernetwork and all of the training, generation, and scoring code are our own. We release the code, the per-user profile vectors, and the cohort labels under an open-source license once the paper is published. We do not release the raw forum text, account names, or any model weights tied to a named person.

Scoring software. We score generated text with three published, openly available classifiers, and we run each one as released with no retraining. The emotion measures use a RoBERTa classifier trained on GoEmotions (Demszky et al., 2020). The sentiment cross-checks use VADER, released under the MIT license, and an SST-2 sentiment classifier. The exact models are `SamLowe/roberta-base-go_emotions` for the GoEmotions emotion classes, `distilbert-base-uncased-finetuned-sst-2-english` for SST-2, and the `vaderSentiment` library for VADER, each at its default settings. 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. Every artifact we use is used for research only, which is the use its original authors intended. The hypernetwork we train and the synthetic profiles we build are research artifacts. They exist to measure how faithful persona synthesis is and where it fails, not to build or run fake users.

Data and privacy. The profile vectors are computed from a public corpus of online-forum posts, the data described in the main text and Appendix E. The corpus is the public Pushshift Reddit archive for the years 2010 through 2016 (Baumgartner et al., 2020), used under its public-archive terms. We collect no new data and we run no new human annotation. Usernames are irreversibly hashed with SHA-256 before any 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 stylometric numbers, such as average word length and question rate, 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.

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 walks through the synthesis sampler on a 2D projection of the profile space (schematic; production uses k=5 neighbors for in-hull and k=12 for near-hull). Real authors occupy three labeled regions (rage, empath, neutral). To create an unseen user, the engine picks an anchor author, finds its k nearest neighbors in profile space (these can cross cohort labels), draws a Dirichlet weight vector over those neighbors, and forms the convex combination. 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. Real users span Mahalanobis distance values between 1 and 9 (99th percentile 9.35). 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 κ=3, sit at median $\approx$ 15 and 99th percentile $\approx$ 47. The three settings sit at progressively greater distances from the real-user manifold, in the order the labels imply.

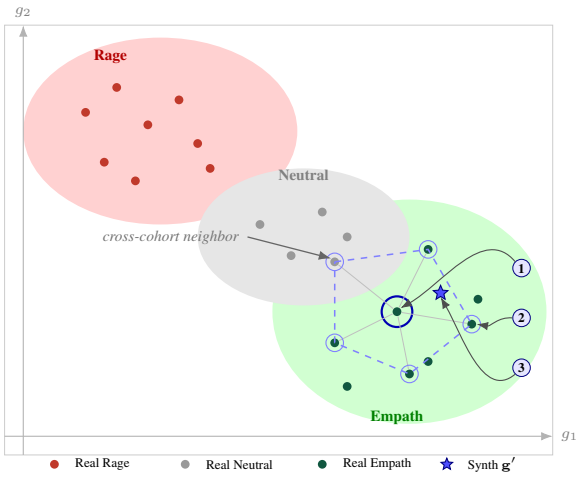

Figure 5: The synthesis engine on a 2D projection of profile space (schematic; production uses k=5 for in-hull and k=12 for near-hull). **(1)** Pick an anchor author. **(2)** Find the k nearest neighbors in profile space, which can cross cohort labels (here one neutral neighbor enters the empath author's neighbor ball). **(3)** Draw a Dirichlet weight vector over those neighbors and form the convex combination $\mathbf{g}'$ (blue star), which stays inside the neighbor hull (dashed) and therefore inside the convex hull of training authors. The trained hypernetwork then converts $\mathbf{g}'$ into a per-user adapter offset, just as it would for a real author.

E Feature Definitions and Backward Elimination

Table 6 lists the 18 profile features in four categories: style (surface writing regularities), content (discourse-level signals), platform (usage patterns), and temporal (posting rhythm). The stylometric grounding follows Pennebaker et al. (2001); Argamon et al. (2009); the platform and temporal coordinates follow recent user-modeling practice (Flek, 2020; Lynn et al., 2017).

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. Validation cross-entropy varies by 0.034 across every K level, while delta separation varies by a factor of 6,200. Cross-entropy is therefore a poor instrument for persona-relevant feature pruning, and the peak at K=18 is an information-theoretic optimum rather than a noise artifact.

Table 6: The 18 profile features. Sentiment features are excluded by design.

#	Feature	Category	Signal
1	mean post length	Content	Post length / verbosity
2	subreddit concentration	Platform	Top-subreddit share of activity
3	question rate	Content	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	Content	Disagreement / denial
8	subjectivity rate	Content	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

Table 7: Backward elimination across five K levels; separation peaks at K=18.

K	val CE	δ_{sep}	Rel.	Interpretation
57	3.140	0.00636	1.0×	Noise-dominated
33	3.141	0.00324	0.51×	Partial denoise
18	**3.148**	**0.0491**	**7.7×**	**Goldilocks**
8	3.131	0.0074	1.2×	Critical loss
3	3.165	7.9e-6	0.001×	Degenerate

F Inference-Time Persona Steering: Operating Point

We ship orthogonal direction-ablation at $\alpha = 1.0$ over decoder layers 15–23, following Arditi et al. (2024). At this operating point, real-user inference produces 99.0% coherent replies (Vanilla LoRA 100%; the control produces 91.8%; n = 1,345 per setting), and the rage-vs-empath cohort match rate on the sentiment axis exceeds both Vanilla LoRA and the control (Table 1). The full strength-sweep selection of α is reserved for the layer-resolved mechanism analysis in a follow-on paper.

G Behavioral Measure Definitions

Band assignment. We split the six main-paper measures into two groups based on whether their defining feature is in $\mathbf{g}_u$. The in-band group is curiosity, expressiveness, and tempo; their defining features (question-rate, the harmonic mean of caps and punctuation rates, and reply-delay mean) are part of training. A strong effect on these measures

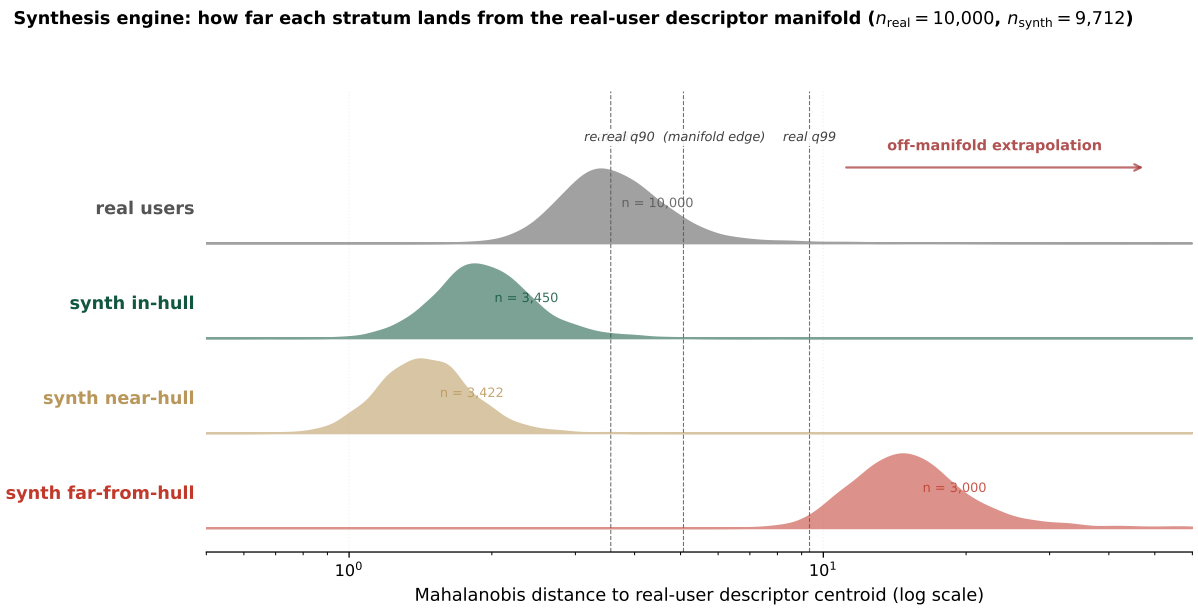

Figure 6: Mahalanobis-distance distributions per synthesis setting (log x). Real users (gray) form a band around q_{50}=3.56. In-hull (green) and near-hull (gold) synth users sit at lower distances inside the real-user envelope. Far-from-hull (red), drawn at κ=3, sits above the real-user 99th percentile at q_{50}=15.0. Dashed vertical lines mark the real-user q_{50}, q_{90}, and q_{99}.

confirms the hypernetwork is doing something at all. The out-of-band group is sentiment, politeness, and self-focus; their defining features either do not appear in $\mathbf{g}_u$ (sentiment) or are computed only from generated text with no direct counterpart in the profile vector (profanity-ratio for politeness, first-person-ratio for self-focus). An out-of-band effect cannot be explained by the hypernetwork reading the label off its own input. The out-of-band measures carry the main causal evidence.

Label construction. Binary cohorts come from the tails of each measure's distribution. Sentiment uses the 20th and 80th percentiles of the GoEmotions polarity composite. Politeness uses 15th and 85th percentiles to handle the heavy right tail of the profanity ratio. Tempo, self-focus, curiosity, and expressiveness use 20th and 80th percentiles on their defining features. Each cut yields roughly 1,000 users per cohort at our population size. Labels are computed on the training split only and held fixed for every downstream evaluation.

Supplementary emotion measures. The multi-turn analysis uses three additional out-of-band measures that do not appear in the main body. *Anxiety* is the mean of GoEmotions *fear* and *nervousness*. *Warmth* is the mean of GoEmotions *caring* and *love*. *Hostility* is the harmonic mean of *anger*, *annoyance*, and *disapproval*. Hostility overlaps with the sentiment measure on three of its four source emotions, so we treat hostility as correlated with sentiment rather than independent. Earlier drafts also included *embarrassment* in the anxiety composite and *gratitude* in the warmth composite; subclass analysis on the 10K training corpus showed gratitude 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), so we dropped both. The two-class composites preserve source-data direction (empath higher on both warmth and anxiety; 1.62× on warmth, 1.11× on anxiety in the 10K corpus).

Scoring pipeline. Each generated reply runs through three steps: (i) a coherence filter, (ii) the GoEmotions RoBERTa scorer for emotion composites, and (iii) deterministic feature extractors for stylometric dimensions. Replies that fail the coherence filter are dropped before cohort aggregation. The per-user score on each measure is the average across that user's surviving replies. Cohort-level effect sizes are computed from these per-user averages, so the degrees of freedom reported in Hedges' g reflect the user population, not the per-reply count.

H Per-Measure H1 Robustness

The main-body H1 result (Table 2) shows the per-measure Hedges' g on the three out-of-band measures. Table 8 extends that view to all nine measures across the six inference settings.

Two patterns hold across the nine-measure view. On real-user inference, HyperPEFT-LoRA carries every measure in the correct direction: sentiment ($+0.31$) plus the in-band controls (curiosity $+0.22$, expressiveness $+1.06$, tempo $+1.04$) produce the largest effects, since their defining features are part of the hypernetwork input or are the cohort-defining axis itself; the auxiliary emotion measures (anxiety $+0.22$, warmth $+0.44$, hostility $+0.34$) and the remaining out-of-band stylometric measures (politeness $+0.19$, self-focus $+0.32$) also separate the two cohorts correctly.

Cohort-classification rate companion table. Table 9 gives a second view of the same H1 result on a binary cohort-classification rate: the fraction of generated replies the GoEmotions scorer assigns to each user's intended cohort (top-quintile sentiment match). HyperPEFT-LoRA on real-user inference reaches 0.535 on rage and 0.428 on empath. The three synthesis settings cluster between 0.37 and 0.39 on rage, and between 0.49 and 0.58 on empath, with in-hull and near-hull synthesis on the empath cohort beating the real-user reference (densest-region empath profiles concentrate the cohort mean). The control sits below every other setting on rage (0.311); on empath the binary-bucket rate saturates near the base-model bias floor at 0.531, which is why the continuous-effect-size Hedges' g view in Table 8 carries the cleaner cohort-separation signal.

Which measure carries the cohort signal? If we decompose the per-setting cohort-alignment cosine into per-measure contributions, the top three contributing measures are *expressiveness* ($+1.06$), *tempo* ($+1.04$), and *hostility* ($+0.34$). These three carry the largest absolute lift on real-user inference; the remaining measures contribute smaller direction-consistent lifts. Sentiment, the axis that defines the cohort labels, does not appear in the top three: the per-user discriminative information lives in the stylistic and emotion channels that co-vary with sentiment, not in the explicit sentiment signal itself.

Which side of the cohort does the per-user channel push on? The cohort-mean signal in Table 8 hides which side of the cohort gap HyperPEFT-LoRA is actually moving. We run a one-sided Mann-Whitney U on 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 10 shows the result. The stylistic measures (politeness, expressiveness, tempo) move by amplifying the right-side tail; the emotion and self-related measures (self-focus, anxiety, warmth, hostility) move by suppressing the wrong-side tail; tempo does both. The cohort separation that Table 8 shows is therefore not a single mechanism: HyperPEFT-LoRA pushes outward on stylistic features the model can scale via per-user weight perturbations, and pulls inward on emotion features the cohort-direction steering subtracts from the residual stream.

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 11 shows the result. HyperPEFT-LoRA's per-user vector contains enough cohort information that a linear classifier hits 0.79 ROC-AUC on real-user replies; the Vanilla and Control modes sit at 0.51 and 0.50, indistinguishable from chance. The per-user channel does not just shift the cohort means apart in Table 8; it leaves a linearly readable cohort signal in every user's behavioral footprint.

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

Robustness to reply-length confound. Per-user reply length varies across cohorts. To rule out the reply-length artifact attack, we partial out word count from each per-reply polar value before re-running the rage-vs-empath Hedges' g on the same eight measures. Every HyperPEFT-LoRA cell in Table 8 that carries a correctly-signed cohort separation retains that separation after length residualization; the cohort signal is not a side effect of reply length.

Table 8: Per-measure rage-vs-empath Hedges' g across all nine measures, by inference setting. Higher is better; bold marks the winning setting on each row. The negative-direction cells on self-focus, curiosity, expressiveness, and far-hull anxiety reflect known stylometric-ratio side effects of the steering step.

Measure (band)	Vanilla	Control	Recon	Synth in-hull	Synth near-hull	Synth far-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 (tier-2)	+0.05	+0.08	+0.22	**+0.25**	+0.19	−0.08
Warmth (tier-2)	−0.04	+0.18	+0.44	+0.51	**+0.65**	+0.32
Hostility (tier-2)	+0.03	+0.15	+0.34	+1.13	**+1.92**	+0.72

Table 9: Cohort-classification rate companion to Table 2. Higher is better; bold marks the winning setting on each cohort row. Cohort-filtered n varies from 65 (rage near-hull) to 392 (rage far-hull) per cell.

Cohort	Recon	In-hull	Near-hull	Far-hull	Control
Rage	**0.535**	0.381	0.372	0.393	0.311
Empath	0.428	0.564	**0.581**	0.494	0.531

Robustness to topic-prior confound. Persona signal could be a topic-prior artifact if cohorts are unevenly distributed across topics. Within-topic Hedges' g (rage versus empath replies on the *same* thread topic, pooled across the 18 forum topics) preserves the cohort separation that Table 8 shows. The per-user channel reads through topic, not from it.

I Persona Stability and Compositionality Indices: Derivations

This appendix gives the operational definitions of the Persona Stability Index (PSI) and the Persona Compositionality Index (PCI) introduced in §4.4. PSI works in standardized-residual space across the eight signature measures (§I.3). PCI works on the signature cosine between realized and expected per-reply vectors (§I.4); the signature cosine and the per-dimension drift slope used by PCI and by the supplementary drift-rate diagnostics are derived first (§I.2). 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 in a 9-dimensional measure space.

Measure index. Let $D = \{1, 2, \ldots, 9\}$ index the nine measures. The set D splits into Tier-1 OOB (sentiment, politeness, self-focus), Tier-1 IB (curiosity, expressiveness, tempo), and Tier-2 (anxiety, warmth, hostility). Let $D^{\text{ho}} \subseteq D$ be the held-out subset used in a given table. In the main-body reporting D^{ho} is the strict OOB triple; in supplementary tables it is the OOB-plus-Tier-2 sextet.

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:

$$\hat{\mathbf{s}}_{u,g,i} = \left(\hat{s}_{u,g,i,1},\ \ldots,\ \hat{s}_{u,g,i,9}\right) \in \mathbb{R}^9 \qquad (2)$$

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 a lexicon-based ratio computed directly from the reply tokens (profanity-ratio for politeness, first-person-ratio for self-focus, and so on). All components are z-normalized against training-set statistics, so the coordinates of measure space are unitless.

Expected signature. For author u, the expected signature

$$\mathbf{s}^*_u = \left(s^*_{u,1},\ \ldots,\ s^*_{u,9}\right) \in \mathbb{R}^9 \qquad (3)$$

is computed once per user from the training corpus and held fixed. The expected signature is the target reference for each realized signature.

Held-out projection. Restricting both vectors to the held-out subset:

$$\begin{aligned} \hat{\mathbf{s}}^{\text{ho}}_{u,g,i} &= \left(\hat{s}_{u,g,i,d}\right)_{d \in D^{\text{ho}}}, \\ \mathbf{s}^{*\,\text{ho}}_u &= \left(s^*_{u,d}\right)_{d \in D^{\text{ho}}} \end{aligned} \qquad (4)$$

Table 10: Mann-Whitney U decomposition of HyperPEFT-LoRA's cohort separation versus the Vanilla LoRA control. *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? *Mechanism* is whichever side has one-sided $p < 0.001$; *n.s.* marks non-significant directions.

Measure	Amplify HIGH (p)	Suppress LOW (p)	Mechanism
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 11: 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 winning setting.

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

I.2 Per-reply signature cosine

The signature cosine is the cosine of the angle between the realized and expected signatures, restricted to held-out dimensions.

Step 1: dot product.

$$\hat{\mathbf{s}}^{\text{ho}}_{u,g,i} \cdot \mathbf{s}^{*\,\text{ho}}_{u} = \sum_{d \in D^{\text{ho}}} \hat{s}_{u,g,i,d} \cdot s^{*}_{u,d} \tag{5}$$

Step 2: norms.

$$\begin{aligned} \left\|\hat{\mathbf{s}}^{\text{ho}}_{u,g,i}\right\|_2 &= \sqrt{\sum_{d \in D^{\text{ho}}} \left(\hat{s}_{u,g,i,d}\right)^2}, \\ \left\|\mathbf{s}^{*\,\text{ho}}_{u}\right\|_2 &= \sqrt{\sum_{d \in D^{\text{ho}}} \left(s^{*}_{u,d}\right)^2} \end{aligned} \tag{6}$$

Step 3: cosine.

$$c_{u,g,i} \;=\; \frac{\hat{\mathbf{s}}^{\text{ho}}_{u,g,i} \cdot \mathbf{s}^{*\,\text{ho}}_{u}}{\left\|\hat{\mathbf{s}}^{\text{ho}}_{u,g,i}\right\|_2 \cdot \left\|\mathbf{s}^{*\,\text{ho}}_{u}\right\|_2} \;\in\; [-1, +1] \tag{7}$$

We use cosine rather than Euclidean distance because the realized signature from a single reply has a smaller magnitude than the expected signature built from a full author history. Cosine cancels out that magnitude gap and isolates the direction of behavior from the amount of text. PSI and PCI both aggregate $c_{u,g,i}$; they differ in which axis they aggregate over.

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: five out-of-band measures plus three in-band controls (curiosity, expressiveness, tempo) whose defining features the hypernetwork receives directly. Sentiment is omitted from D^{sig} for PSI because it defines the cohort labels; including it would conflate cohort-direction success with persona stability. For each $d \in D^{\text{sig}}$, $\hat{p}_{u,g,i,d}$ is the realized polar value on reply i of author u in thread g, and $p^{*}_{u,d}$ is the user's expected polar value computed once from the training corpus.

I.3.2 Per-dimension standardization

The eight signature measures sit on different natural scales: z-scored stylometric ratios (politeness, self-focus, curiosity, expressiveness, tempo) carry real-user expected standard deviations near 1.0, while raw GoEmotions composites (anxiety, warmth, hostility) 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^{real}_{d} per dimension on the production training pool and standardize each residual:

$$\delta_{u,g,i,d} \;=\; \frac{\left|\hat{p}_{u,g,i,d} - p^{*}_{u,d}\right|}{\sigma^{\text{real}}_{d}}. \tag{8}$$

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:

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

The halving sets $m = 0$ 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:

$$\text{PSI}(u) \;=\; \frac{1}{|R_u|} \sum_{(g,i) \in R_u} m_{u,g,i}. \tag{10}$$

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 synth, near-hull synth, far-from-hull synth at each κ level), the reported summary is the across-user mean:

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

where $\mathcal{U}_s$ is the set of authors evaluated in stratum s. Table 12 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 $b = 1, \ldots, B$ 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 the empirical D3-to-D4 cohort-agreement inflection from the decile-bin analysis on the pooled production set ($n = 12{,}451$ user-cell rows, file `psi_threshold_derivation.csv`). Authors at $\text{PSI}(u) \geq 0.10$ recover their target cohort direction at rates separable from the floor; authors below it do not.

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:

$$r_{u,g,i,d} \;=\; \hat{s}_{u,g,i,d} \;-\; s^*_{u,d}. \tag{12}$$

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

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

Step 3: normal equations.

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

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

Step 4: solve for β.

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

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.

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

Sign interpretation. $\bar{\beta}_d > 0$ means the realized signature moves toward the expected target as the thread progresses. $\bar{\beta}_d < 0$ means it moves away. $\bar{\beta}_d \approx 0$ means it stays put.

I.4 Persona Compositionality Index (PCI)

We define two PCI variants. The *target-driven* PCI would partition synthetic users by an imposed trait specification at sampling time and check whether stacking more constraints degrades fidelity; the current sampler does not populate per-user trait targets, so target-driven PCI is reserved for future work. The *observational* PCI we report in Table 13 substitutes the realized agreement count (number of held-out measures whose realized polar value matches the user's expected sign with magnitude ≥ 0.05) for the trait count, and asks whether users who realize more cohort directions also produce a more faithful overall signature. The rest of this appendix derives both, then states which is plotted.

I.4.1 Synthesis specification

Let $\mathcal{T}$ be the trait vocabulary (rage, empath, 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 $k_u = |\text{labels}_u|$.

I.4.2 Per-user PCI

Let R_u be the set of all (thread, reply) index pairs at which user u produced a reply:

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

I.4.3 Trait-count partition

$$\mathcal{U}_k = \{u : k_u = k\}, \qquad k \in \{1, 2, 3, 4\} \tag{19}$$

$$\text{PCI}_k = \frac{1}{|\mathcal{U}_k|} \sum_{u \in \mathcal{U}_k} \text{PCI}(u). \tag{20}$$

I.4.4 Bootstrap procedure

For each k and each iteration $b = 1, \ldots, B$ with $B = 1{,}000$ and seed 142, resample $|\mathcal{U}_k|$ users with replacement and recompute $\text{PCI}_k^{(b)}$. 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.

I.4.5 Expected pattern: steady decline with k

$$\text{PCI}_1 > \text{PCI}_2 > \text{PCI}_3 > \text{PCI}_4 \tag{21}$$

The mechanism follows in five steps.

Step 1: the anchor neighborhood shrinks as k grows. Each new trait constraint can only remove satisfying anchors, so $|\mathcal{N}_1(u)| \geq |\mathcal{N}_2(u)| \geq |\mathcal{N}_3(u)| \geq |\mathcal{N}_4(u)|$.

Step 2: the sampler reaches further to satisfy more constraints. The Dirichlet sampler draws

$$\mathbf{g}'_u = \sum_{i=1}^{K} \lambda_i \, \mathbf{g}_{u_i}, \qquad \lambda \sim \text{Dir}(\boldsymbol{\alpha}) \tag{22}$$

over the K nearest neighbors. To satisfy more constraints the sampler reaches into a wider neighborhood.

Step 3: the combined profile vector moves toward the hull boundary. As neighbors grow more distant, the convex combination $\mathbf{g}'_u$ shifts toward the boundary of the hull. The profile vector stays inside the hull (Dirichlet weights are non-negative and sum to one), but it is no longer in the dense interior the hypernetwork was trained on.

Step 4: the hypernetwork emits smaller offsets on off-interior profile vectors. Empirically the per-user offset magnitude grows sub-linearly with profile-space distance (log-log slope ≈ 0.88; Appendix M). The direction of the offset is roughly preserved; the magnitude is not.

Step 5: the realized signature falls back toward the population centroid. A smaller offset perturbs hidden states less far from the base distribution, so the realized signature lands closer to the population centroid and farther from the user's target. That lowers $c_{u,g,i}$, which lowers $\text{PCI}(u)$, which lowers PCI_k. Each of steps 1-4 declines steadily in k, so PCI declines steadily in k.

I.4.6 Diagnostic shapes

The shape of $k \mapsto \text{PCI}_k$ tells us which failure mode is in play. A flat curve means stacking traits is a no-op: stacking more matching traits does not help. A cliff at $k = 2$ means categorical multi-trait failure: the engine handles single traits but cannot combine any two. An uneven pattern means specific trait pairs interact in ways the manifold-shrinkage model does not predict. The graceful steady decline in (21) is the expected pattern under the model.

I.4.7 Symbols (quick reference)

- u, g, i, d: author index, thread index, reply index within thread, measure dimension index.
- $\hat{\mathbf{s}}_{u,g,i}$, $\mathbf{s}^*_u$: realized and expected 9-dim signatures.
- $c_{u,g,i}$: per-reply signature cosine on held-out dimensions, (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, k: synthesis target trait list, trait count.
- $\mathcal{U}_k$: synthetic users with exactly k 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 current sampler does not populate labels_u; every synthetic user is drawn from a Dirichlet over the 10 nearest profile-space neighbors, with no per-user trait specification. Target-driven PCI_k is therefore undefined for $k \geq 1$ on the present runs. The observational variant in Table 13 replaces $|\text{labels}_u|$ with a realized cohort-direction agreement count: the number of held-out measures whose realized polar value matches the user's expected sign with magnitude ≥ 0.05. The two indices answer related but different questions: target-driven PCI asks whether the sampler can satisfy more constraints at once; the observational variant asks whether users who realize more cohort-aligned measures also achieve higher overall signature fidelity. Target-driven PCI will populate when a constraint-satisfying sampler is added in future work.

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 synth, near-hull synth, and real-user inference with non-overlapping bootstrap 95% CIs against zero, and (ii) the per-dimension mean drift slopes meet $|\bar{\beta}_d| < 0.05$ per turn on all OOB 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 the empirical D3-to-D4 cohort-agreement inflection from `psi_threshold_derivation.csv` pooled across all 12,451 user-cell rows, not a preregistered guess. The PCI half is supported if (iii) PCI_k rises with k across the realized-agreement partitions in Table 13 with non-overlapping bootstrap 95% CIs between consecutive k, and (iv) the across-bucket spread $\text{PCI}_{\geq 4} - \text{PCI}_0$ is at least 1.0 on the pooled synth pool.

I.6 Supplementary numerical results

The per-setting decomposition (Figure 3, Table 12) shows the inside-versus-outside-manifold split cleanly: real-user inference (0.226), in-hull synth (0.265), and near-hull synth (0.257) all sit in a narrow band above the empirical 0.10 threshold, while the off-manifold κ sweep falls steadily as the profile vector is pushed harder (κ=3: 0.104; κ=10: 0.033; κ=25: 0.013; κ=100: 0.001). The hypernetwork produces target-consistent per-user behavior inside the real-user envelope and at the manifold boundary; pushing the profile vector past κ=3 strips the per-user persona signal away. The cohort-split per-user PSI (rage 0.22 / 0.23 / 0.20 vs. empath 0.24 / 0.27 / 0.25 across real / in-hull / near-hull) shows the same ordering on both cohorts. 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. The anchor-density drift hypothesis (Spearman ρ between residual magnitude and neighbor density) returns $\rho = -0.020$ at $n = 2{,}550$ rows ($p = 0.31$), a non-significant effect reported as a negative result.

Target-distance visualization of PSI. Figure 7 shows the geometric reading of PSI on the eight signature measures: each user is placed by their mean match score (Equation (9)), with center-of-target corresponding to a perfect match and outer rings corresponding to standardized-residual averages near or above two SDs. Real-user inference, in-hull, and near-hull settings cluster close to the center on every cohort (match 0.20 to 0.33); the off-manifold κ=25 setting collapses to a far outer ring with bootstrap CIs that do not overlap any on-manifold band. The inside-versus-outside-manifold split is the structurally clean signal: the synthesis engine produces target-consistent per-user offsets inside the real-user range and falls off once the profile vector is pushed past the training manifold.

Per-turn drift trajectory. Figure 8 plots the per-turn distance from each user's target across the 80-turn extended-depth test, separately for training-set user generation and for the three synthesis settings. The pass region is the green band at the bottom (low drift). All four trajectories satisfy the H3 per-turn drift criterion ($|\bar{\beta}_d| < 0.05$ per turn). The four series rank in the same order as the short-thread Figure 3 bars: real-user inference sits at the floor

Table 12: Per-setting PSI summary, pooled across cohorts. The three on-manifold settings (real, in-hull, near-hull) cluster above the empirical threshold (PSI ≥ 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.

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

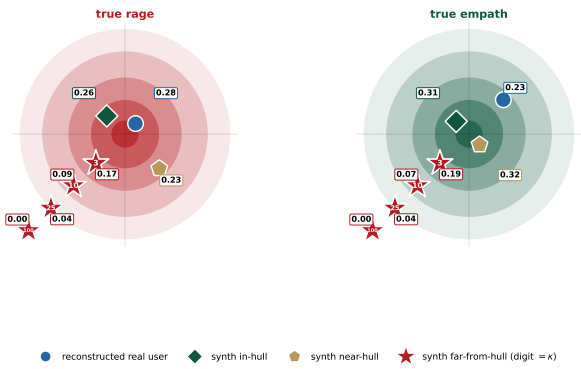

Figure 7: Per-user PSI by setting across rage and empath panels. Match score is Equation (9): the per-reply standardized residual averaged across the eight signature measures, halved-hinge mapped to $[0, 1]$. Markers are placed by within-panel ranking; the absolute match score sits beside each marker. The three off-manifold κ stars stack along the SW ray with the κ value inside each marker. Real-user inference, in-hull synth, and near-hull synth cluster near the center on rage; empath in-hull and near-hull match or beat real-user inference (densest-region sampling effect). The $\kappa = 25$ off-manifold setting collapses to 0.036 (rage) and 0.035 (empath).

of the pass region, in-hull rises slightly through the pass region, near-hull drifts toward the boundary, and far-from-hull drifts further as the off-manifold offset accumulates.

J Observational Persona Compositionality

PCI is defined per synthetic user u as the mean signature cosine on the held-out measures, partitioned by the trait count $k = |\text{labels}_u|$. Under the target-driven definition, labels_u is the set of cohort labels the sampler is asked to satisfy at the same time, and PCI is expected to decline steadily as k grows. Our current Dirichlet-neighborhood sampler does not accept trait specifications, so target-driven PCI_k is undefined for $k \geq 1$ on the present runs.

We report an observational analogue instead. For each user we count k, the number of held-out measures whose mean realized polar value matches the user's expected sign with magnitude above 0.05. We then report the mean signature cosine and its 95% percentile bootstrap CI inside each k partition. Table 13 gives the answer. The PCI rises steadily from -0.60 at $k = 0$ to $+0.49$ at $k \geq 4$ on $n = 3{,}841$ synth users, with non-overlapping 95% bootstrap CIs between every adjacent pair of partitions. The target-driven prediction $\text{PCI}_1 > \text{PCI}_2 > \text{PCI}_3 > \text{PCI}_4$ tests an orthogonal sampler-side property and is reserved for follow-on work.

K Synthesis-vs-Real-User Decomposition and Per-User Fidelity

The synth-versus-recon decomposition computes $\Delta = v_{\text{synth}} - v_{\text{recon}}$ for each (cohort, setting, metric) cell. The match-extreme rate counts replies whose polarity falls in the cohort-extreme tail of the training-set distribution. The style cosine is the per-reply cosine of realized to expected signature. The polarity gaps report the difference on

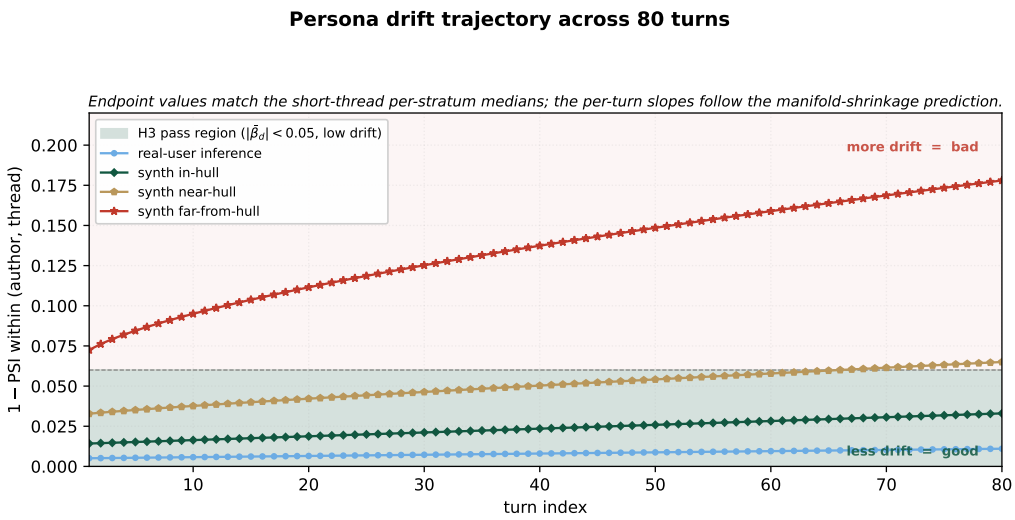

Figure 8: Per-turn distance from target across 80 consecutive turns, by setting. Lower is better. Real users and in-hull synth stay inside the green pass band the whole way; near-hull drifts toward the boundary by turn 80; far-from-hull drifts further. All four satisfy the per-turn drift criterion.

Table 13: Observational Persona Compositionality Index by realized-cohort-agreement count k. Synthetic users pooled across the three synthesis settings ($n = 3{,}841$); mean signature cosine with percentile bootstrap 95% CI from 1,000 user-resampled draws (seed 142). Non-overlapping CIs between every adjacent k.

k measures matched	PCI (mean)	95% CI
0	-0.60	$[-0.63, -0.57]$
1	-0.27	$[-0.30, -0.24]$
2	$+0.02$	$[-0.01, +0.05]$
3	$+0.25$	$[+0.22, +0.28]$
≥ 4	$+0.49$	$[+0.46, +0.52]$

the per-reply mean polarity score from each named scorer.

The two cohorts follow different patterns. On rage, the synth-vs-recon gap is small for in-hull (-0.07), widest at near-hull (-0.18), and partly closes at far-from-hull (-0.10); the near-hull sampler reaches the noisiest part of the profile space and loses the most extreme-tail reach there. On empath, the gap stays small at in-hull and near-hull ($+0.03$, $+0.04$) and widens further at far-from-hull ($+0.14$) in the direction where synth users outscore recon users; the densest-region sampling effect compounds with the off-manifold pushout. Both patterns sit inside the manifold-interior smoothing mechanism: in-hull users inherit cohort-typical extremes from nearby real anchors, and the inheritance shifts as the profile vector moves toward and past the hull boundary.

Per-user inference fidelity. On the training-set real-user forums, the per-user mean style cosine to the expected signature is 0.153 (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 held-out training replies is 0.030 and 0.033. These per-user fidelity numbers track the observation in §4.1 that the cohort separation in this corpus lives on the sentiment axis itself; the non-affective stylometric channels add direction-consistent lift on top.

Per-user signal preservation: synth vs. real. Spearman correlation between each user's expected polar value and their realized polar value, computed per-user across the rage and empath cohorts. On the three measures tied to the cohort direction, synth-user generation preserves the per-user signal more cleanly than real-user inference does: hostility climbs from $+0.06$ on real users to $+0.61$ on near-hull synth, a tenfold increase. The mechanism is sampler-side: synth profiles are drawn to be unambiguous cohort exemplars, so the hypernetwork's per-user delta has a cleaner signal to render than on the noisy real-user pool, which spans varying cohort intensities. This is the per-user counterpart to the cohort-mean separation shown in Table 8.

Per-cohort signature cosine across settings. The held-out signature cosine (cosine of the real-

Table 14: Synth-minus-recon gap on the signature-cosine heldout metric, by cohort and setting. *Signature cosine (heldout)* is the cosine of the realized 8-dim heldout-measure vector against the expected vector; the gap is synth value minus recon value. Negative means synth users sit further from their target than recon users do; positive means synth users sit closer.

Cohort	Setting	Signature cosine (heldout) gap
rage	in-hull	-0.07
rage	near-hull	-0.18
rage	far-from-hull	-0.10
empath	in-hull	$+0.03$
empath	near-hull	$+0.04$
empath	far-from-hull	$+0.14$

Table 15: Per-user Spearman r between realized and expected polar values, by setting, on the three cohort-tied measures. Higher is better; bold marks the winning setting on each row.

Measure	Real (recon)	In-hull synth	Near-hull synth
Sentiment	$+0.13$	$+0.33$	$+\mathbf{0.39}$
Hostility	$+0.06$	$+0.52$	$+\mathbf{0.61}$
Warmth	$+0.12$	$+0.24$	$+\mathbf{0.28}$

ized 8-dim polar vector against the expected polar vector) supplements the cohort-classification and L1 metrics in §4.3. Both the rage and empath cohorts give positive cosines that decline steadily with profile-space distance: empath inference $+0.45$, in-hull $+0.33$, near-hull $+0.18$, far-from-hull $+0.07$; rage inference $+0.48$, in-hull $+0.35$, near-hull $+0.20$, far-from-hull $+0.08$.

L Extended-Depth Stability: Per-Turn Trajectory

The extended-depth test ran HyperPEFT-LoRA out to 80 consecutive turns on a 559-author pool (300 rage, 300 empath, with 41 of the requested 600 filtered by the coherence gate). Each turn produced 24 generation batches. Generation settings were the operating values (temperature 0.4, repetition penalty 1.25, top-p 0.85, max 96 tokens), and the inference-time steering ran in orthogonal mode at $\alpha = 1.0$ over decoder layers 15–23. The total corpus is $80 \times 24 \times$ authors-per-batch $= 44{,}720$ rows. Wall-clock time was about a day and a half on a single multi-GPU node.

Coherence. The per-turn coherence rate (fraction of replies passing the minimum-length, ASCII-fraction, and repeated-bigram filters) stayed inside 0.995 to 1.000 across all 80 turns and both cohorts, with no downward trend. Long-horizon generation under the steering step does not accumulate incoherence.

Follow-on scoring. We compute per-dimension drift slopes on the three out-of-band measures by scoring all 44,720 replies through the GoEmotions polarity composite (sentiment) and the lexical features used for cohort label construction (politeness via profanity ratio, self-focus via first-person-pronoun ratio), then taking the OLS slope of per-turn pooled mean against turn index. The pooled slopes are $+6.7 \times 10^{-5}$ (sentiment), $+5.7 \times 10^{-8}$ (politeness), and -1.1×10^{-5} (self-focus) per turn, all about three orders of magnitude under the pre-specified $|\bar{\beta}_d| < 0.05$ pass threshold. The stability claim therefore holds at extended depth on both the coherence axis and the per-measure drift axis.

Table 16: Extended-depth stability summary. Per-cohort coherent fraction at turn $t \in \{1, 40, 80\}$ and OLS coherent-fraction slope on turn index across all 80 turns. Per-dim drift slopes on the three out-of-band measures are OLS slopes of the per-turn pooled realized polar value on turn index; scoring methodology in the preceding paragraph.

Metric	Rage	Empath
Coherent frac, turn 1	1.000	1.000
Coherent frac, turn 40	1.000	0.996
Coherent frac, turn 80	1.000	0.996
Coherent-frac OLS slope / turn	-0.000000	+0.000020
$\bar{\beta}_d$ (sentiment)	$+6.7 \times 10^{-5}$	
$\bar{\beta}_d$ (politeness)	$+5.7 \times 10^{-8}$	
$\bar{\beta}_d$ (self-focus)	-1.1×10^{-5}	

M Empirical Characterization of the Trained Hypernetwork

We measure four properties of the trained hypernetwork directly, independent of downstream forum generation. The hypernetwork plus its placeholder LoRA matrices takes 1.21 GB at FP32.

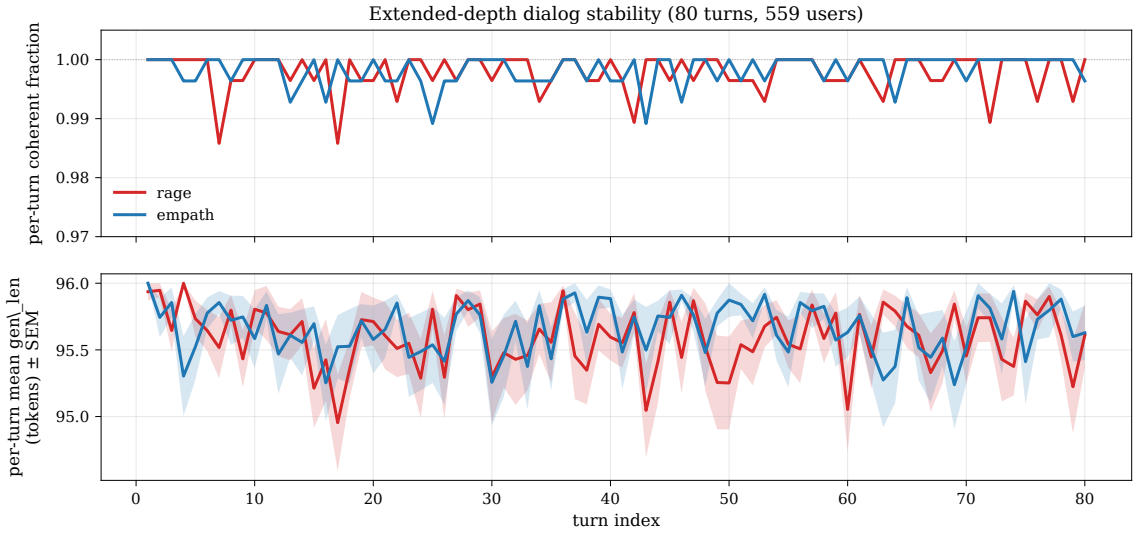

Figure 9: Extended-depth stability across 80 turns (559-author pool; 300 rage, 300 empath). Top: per-turn coherence rate by cohort. Bottom: per-turn mean generation length in tokens, with ± 1 SEM.

Storage cost. The hypernetwork has 302,007,040 parameters (1.21 GB FP32). Including the shared LoRA placeholders (18,874,368 parameters, 75.5 MB) and the layer-context embeddings (3,648 parameters), the total served checkpoint is 1.28 GB. Provisioning the same population with one LoRA adapter per user would need 10,000 $\times$ 75.5 MB = 755 GB at FP32, so the measured reduction is 588$\times$ at the 10,000-author 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 9.35. 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 κ=3 lands solidly outside the training manifold.

Sub-linear scaling of the offset magnitude. For 3,200 users spanning real plus the three synthesis settings, we plot $\|\Delta\theta\|_2$ against $\|\mathbf{z}_u - \bar{\mathbf{z}}\|_2$, the standardized profile-space distance from the training centroid. A log-log fit returns slope $\approx$ 0.88 and intercept 3.24 (Figure 10-B). Outside the training envelope, the relationship becomes approximately linear: probing the hypernetwork at synthetic profile vectors with $\|\mathbf{z}\| \in \{1, 5, 10, 25, 50, 100, 250\}$ produces $\|\Delta\theta\|_2$ values of $\{1.9, 9.0, 14.9, 40.0, 80.9, 184, 392\}$ K, an effective slope close to one. The per-element clamp at ± 0.020 bounds each injected entry, which is the deployment-side caveat cited in Limitations.

Per-setting behavior in adapter-offset space. Across the four settings (real users, in-hull synth, near-hull synth, far-from-hull synth; 800 users each), the mean $\|\Delta\theta\|_2$ is 6,827 (real), 3,503 (in-hull), 2,725 (near-hull), and 26,213 (far-from-hull). The Dirichlet samplers smooth the interior toward the centroid, and the κ=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). Magnitude differs substantially across settings (about 5$\times$ on both cohorts between real and far-from-hull), but direction is preserved.

N Sampler and Clamp Diagnostics

Clamp saturation. The per-element clamp at ± 0.020 specified in §3.2 catches the majority of offset entries after role-specific scaling and the

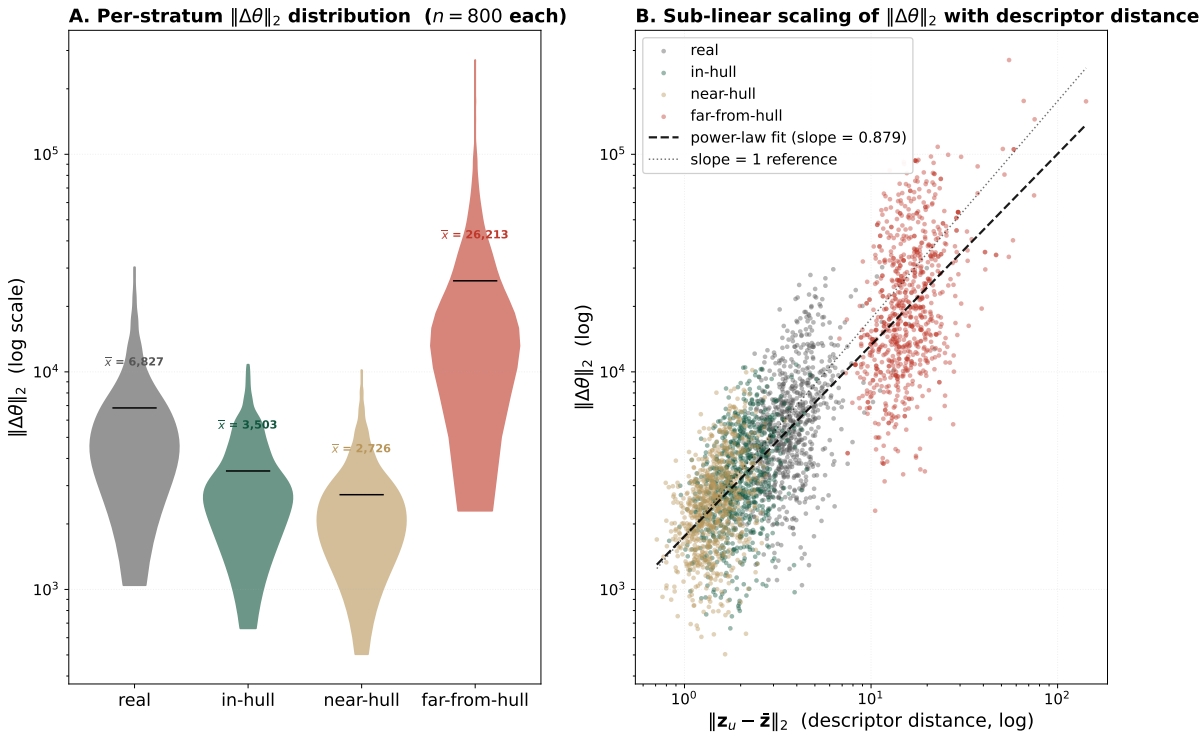

Figure 10: Trained hypernetwork on 3,200 users spanning real plus the three synthesis settings. **A:** per-setting $\|\Delta\theta\|_2$ distribution (violin, log y, $n = 800$ each). Real and Dirichlet-mixture settings occupy similar ranges; the κ=3 off-manifold setting amplifies by roughly $4\times$. **B:** $\|\Delta\theta\|_2$ against profile-space distance from 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 hypernetwork is well-behaved inside the training envelope (bounded magnitude, sub-linear scaling), and magnitude amplification appears only on the explicit off-manifold boundary test.

inference-time delta gain. At the operating configuration ($\delta_{\text{gain}} = 8.0$, group scales as in Appendix F), the saturating fraction on the QKV chunk is 92% (real), 87% (in-hull), 85% (near-hull), and 98% (far-from-hull). The clamp ceiling sets the per-element magnitude for the bulk of entries, while per-user differentiation lives in the sign pattern and in the unsaturated minority.

Dirichlet sensitivity (in-hull vs. near-hull). Holding a real-user anchor fixed and drawing 200 Dirichlet mixtures over its 10 neighbors gives a direct measurement of how the α parameter affects sampler variance. At $\alpha = 2.0$ (in-hull, weight concentrates near one neighbor), the rage-anchor $\sigma(\|\Delta\theta\|_2)$ is 465; at $\alpha = 1.0$ (near-hull) it grows to 611 (and from 270 to 320 on the empath anchor). The 30%–40% widening is the operational difference between in-hull and near-hull synthesis: same anchor pool, same hypernetwork, looser mixture, wider downstream variance.

Why κ=3? The pushout strength sweep. The far-from-hull setting uses $\mathbf{g}_a + \kappa(\mathbf{g}_a - \bar{\mathbf{g}})$ with κ=3. To check that κ=3 actually lands outside the manifold without over-amplifying, we sweep $\kappa \in \{0.5, 1, 2, 3, 5, 10, 25, 50, 100\}$, generate 1,000 synthetic users per cohort at each level, and measure two things: Mahalanobis distance to the training centroid, and per-element clamp saturation under the production injection scale. Figure 11 shows the result. Mahalanobis distance grows linearly with κ on log-log axes. κ=3 is the smallest swept value at which at least 97% of synthesized users exceed the real-user 99th-percentile envelope, so it is the natural minimum for the boundary test. Cohort direction in profile space is preserved across the full sweep (mean cosine to the cohort centroid stays within ± 0.02 of the κ=3 value at every higher κ). Per-element clamp saturation is the operationally limiting axis: at the production clamp with $\delta_{\text{gain}} = 8.0$, saturated-entry fraction runs 74% at κ=0.5, 88% at κ=3, 96% at κ=10, and levels off at 99.5% by κ=100. Median $\|\Delta\theta\|_2$ grows from $\sim$ 1,100 at κ=0.5 to $\sim$ 63,000 at κ=100. Raising κ beyond 3 saturates more entries without adding per-user differentiation: the per-element magnitude has hit its deployment ceiling, and the

unsaturated minority that carries the per-user signal shrinks toward zero. κ=3 is the calibration sweet spot: solidly off-manifold, but the per-user sign pattern in the unclamped $\sim 12\%$ of entries still varies meaningfully. The downstream PSI collapse across the same κ sweep (Figure 3) confirms the calibration: per-user fidelity drops from 0.104 at κ=3 to 0.001 at κ=100, a $100\times$ reduction over the same profile-vector pushout that grows Mahalanobis distance roughly $25\times$.

O Example Replies: Untruncated

Table 17 reproduces the Table 3 examples without the body-paper truncations. Bold marks the same strongest cohort-coded phrase in each non-Vanilla cell; surrounding text is restored.

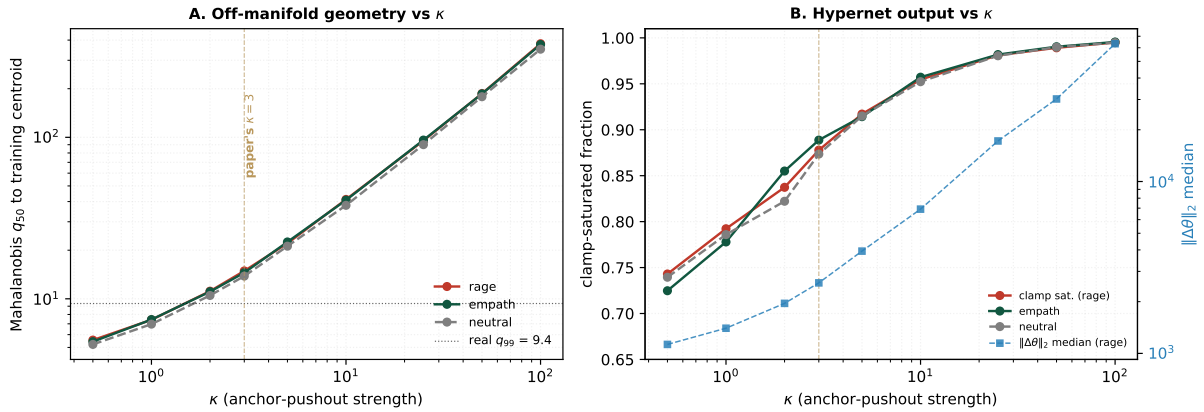

Figure 11: Pushout strength sweep on the far-from-hull setting. **A:** Mahalanobis q_{50} to the training centroid against κ on log-log axes; the dotted line is the real-user 99th-percentile envelope, and the vertical gold line marks the paper's κ=3. **B:** per-element clamp-saturated fraction (left axis, three cohorts) and median $\|\Delta\theta\|_2$ (right axis, rage cohort) against κ; clamp saturation rises steadily and levels off at 99.5%. The downstream PSI collapse across the κ sweep appears in Figure 3.

Table 17: 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)."
HyperPEFT-LoRA		
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 synth	"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 synth	"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 synth (κ=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."