Persona Generators: Generating Diverse Synthetic Personas for Arbitrary Contexts

Davide Paglieri
Google DeepMind

Logan Cross
Google DeepMind

William A. Cunningham
Google DeepMind

Joel Z. Leibo
Google DeepMind

Alexander Sasha Vezhnevets
Google DeepMind

Abstract

Simulating human behavior with Large Language Models (LLMs) offers a scalable laboratory for social science and stress-testing AI products. However, limitations remain in aligning LLM outputs with the full breadth of human diversity. Current techniques for creating synthetic humans typically optimize for density matching, which often enforces behavioral uniformity and overlooks the rare, consequential outliers. We argue that robust simulation requires shifting the focus toward support coverage: ensuring synthetic populations span the entire landscape of possible human traits, opinions, and preferences. This paper introduces Persona Generators, functions that produce diverse synthetic populations for arbitrary contexts. We apply an iterative improvement loop based on AlphaEvolve, using LLMs as mutation operators to evolve our Persona Generator code rather than the personas themselves. The optimization process produces lightweight functions that automatically expand user provided scenario prompts into populations of diverse synthetic personas that are maximizing coverage along relevant diversity axes. We demonstrate that evolved generators substantially outperform existing baselines across six diversity metrics on held-out contexts. Furthermore, by successfully mitigating standard LLM mode collapse, our generated populations actually capture real human trait distributions better than baselines specifically designed to match human statistics.

1 Introduction

Simulating human populations is a promising research direction for answering social science questions in silico (Anthis et al., 2025) and for evaluating human-AI interaction (Kim et al., 2023; Poole-Dayan et al., 2025). Recent progress in Generative Agent-Based Modeling (GABM) has made it possible to simulate complex social interactions (Park et al., 2023), societal response to COVID-19 (Kozlowski et al., 2024), behavioral patterns in surveys and economic games (Slumbers et al., 2025; Park et al., 2024; Binz et al., 2025), social phenomena of status signaling Cross et al. (2026), and psychological phenomena of attitude change Matyas et al. (2026). Whether modeling collective outcomes on social platforms or stress-testing personal health and coaching assistants, understanding the full breadth of potential human behaviors is critical for ensuring system robustness and safety. Yet, one major challenge in simulating human populations accurately is that standard LLMs often collapse onto a narrow subset of stereotypical or highly agreeable WEIRD (Western, Educated, Industrialized, Rich, Democratic) responses, rendering off-mode but consequential behaviors underrepresented (Bisbee et al., 2024; Anthis et al., 2025; Petrov et al., 2024). Additionally, most existing work in GABM focuses on *algorithmic fidelity*—how accurately synthetic personas reproduce observed human response patterns, often evaluated by matching the aggregate statistics or distributions derived from real data (Argyle et al., 2023; Park et al., 2024). While effective for applications that prioritize the average user, density matching is fundamentally unsuited for stress-testing or forecasting in

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

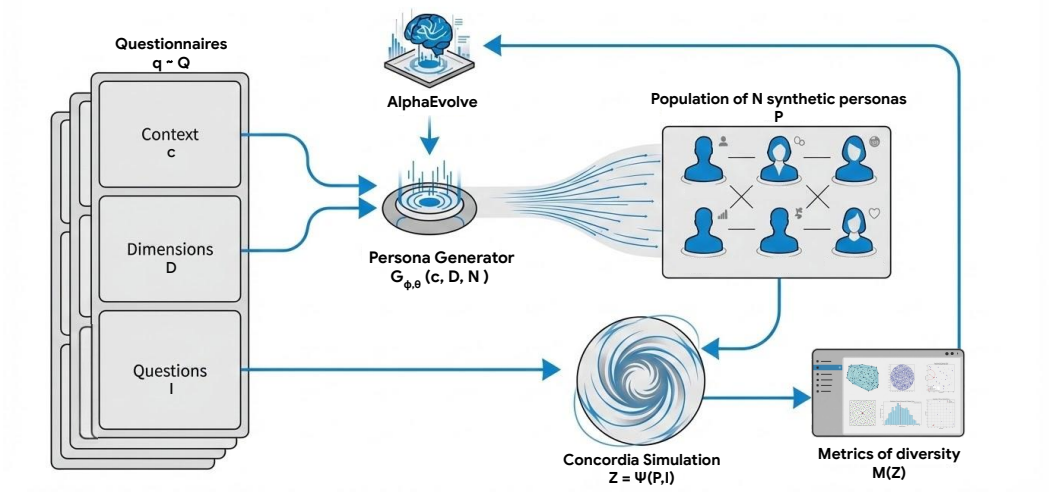


Figure 1: **Overview of the method.** First, we generate questionnaires containing specific contexts c, dimensions $\mathcal{D}$ (diversity axes) and questions $\mathcal{I}$ (items). The Persona Generator $G_{\phi,\theta}(c, \mathcal{D}, N)$ uses the context and diversity axes as inputs to create a population of synthetic personas $\mathcal{P}$. These personas are evaluated in Concordia simulations $Z = \Psi(\mathcal{P}, \mathcal{I})$ where they answer the questionnaire items. We measure the diversity of their responses $\mathcal{M}(\mathcal{Z})$ and use AlphaEvolve to iteratively optimize the Persona Generator code ϕ.

speculative scenarios—such as societal response to novel technology (e.g. AGI)—where exploring the full range of possible human responses is more critical than identifying the most probable one.

In this paper, we propose support coverage as a primary objective: creating populations that capture the *full support* of possible attitudes and preferences. Leaving the long tail underexplored can create a false sense of robustness; in many stress-testing settings, it is the outliers—such as rare users with extreme behaviors—and not the average user, that drive critical failures. The density matching objective is not sensitive to these cases since their share of the population is low.

Eliciting such diversity from LLM-based personas, however, is far from trivial. Naive prompting leads to mode collapse, stereotypical outputs, and systematic biases, partly a byproduct of Reinforcement Learning from Human Feedback (RLHF) tuning, even when explicit instructions for diversity are provided (Santurkar et al., 2023; Li et al., 2025). Our experiments show that simply asking an LLM to "generate diverse personas" typically yields populations clustered around stereotypical responses, failing to cover extreme or unusual trait combinations.

Rather than constructing a single fixed population of synthetic users, we propose to learn a reusable Persona Generator: a function capable of producing diverse synthetic personas on demand for any arbitrary context from a single scenario prompt (e.g. "clients of food delivery apps in London"). This generator must handle the immense variety of potential scenarios that we may care about. Importantly, we aim to optimize a generally capable Persona Generator, which samples and constructs a population of personas (see Fig. 2).

Starting from a short scenario prompt, we first expand the context into a structured questionnaire that defines a set of diversity axes (e.g. dietary preferences, budget). A Persona Generator then produces a population of synthetic individuals intended to span all possible traits, opinions, and preferences defined by those axes. To overcome the difficulties of achieving diversity through standard prompting, we frame the task as an optimization problem over the Persona Generator's code, including prompt templates and sampling logic, using an evolutionary search loop powered by AlphaEvolve (Novikov et al., 2025).

The resulting Persona Generator is lightweight and efficient, enabling rapid, one-shot population synthesis for downstream applications. This separation between a costly training phase and a cheap inference phase makes Persona Generators practical for repeated use across domains, even when the

original optimization context differs from the deployment setting. Crucially, we find that maximizing support coverage does not come at the expense of realism. By explicitly fighting LLM mode collapse, our approach generates populations that actually match the true variance of real human trait distributions better than baselines specifically grounded in demographic data. Ultimately if the full support is covered, one can always subsequently re-sample the population to match any specific target density.

Our contributions are summarized as follows:

- We formalize the problem of synthetic persona generation as a diversity maximization task over trait and preference embeddings, explicitly shifting the objective from algorithmic fidelity (matching specific individuals) to support coverage (spanning the space of possible traits, opinions, and preferences).
- We propose a novel Persona Generator function with a two-stage scalable architecture that separates population-level diversity decisions from per-persona background expansions, enabling both control and efficiency, coupled with a scalable pipeline that uses LLMs to generate questionnaires, simulate interactions, and evolve code.
- We demonstrate that LLM-driven evolution can discover novel Persona Generator functions that substantially outperform different baselines in coverage and diversity metrics. Furthermore, we show that by successfully mitigating LLM mode collapse, our generated personas actually capture real-world human trait distributions better than baselines specifically intended to mirror human statistics.

2 Evolving Persona Generators

We introduce a method for learning Persona Generators: functions capable of producing diverse populations of synthetic personas for *any* arbitrary context. Our full pipeline is illustrated in Figure 1, and this section will walk through each of its core components.

The process starts by generating a set of evaluation tasks. As we will detail in Section 2.2, we create synthetic questionnaires (q), each comprising a context (c), axes of diversity (D), and survey items (I). These questionnaires serve as the benchmark for evaluating the diversity of our generated populations.

The central component is the Persona Generator ($G_{\phi,\theta}$), which takes a questionnaire's context and diversity axes to produce a population of personas (P). The architecture of this generator, shown in Figure 2, follows a two-stage process to efficiently generate diverse individuals. We describe the initial implementations of the generators in Section 2.3.

To evaluate a population, each persona is placed in a Concordia simulation where they answer the questionnaire items (see Section 2.4). This process maps the population of text-based personas to a set of numerical response embeddings (Z). We then quantify how well these embeddings cover the target space using several diversity metrics ($M(Z)$), which are defined in Section 2.5.

Finally, we use this evaluation as the fitness function in an evolutionary search loop driven by AlphaEvolve (Section 2.6) (Novikov et al., 2025). This loop iteratively mutates the code (ϕ) of the Persona Generator to discover new versions that maximize the diversity of the personas they produce.

2.1 Problem Formulation

We formalize the problem of generating populations of diverse synthetic personas as follows. Let $\mathcal{Q}$ be a distribution of questionnaires, with each questionnaire $q \sim \mathcal{Q}$ defined as a tuple $q = (c, \mathcal{D}, \mathcal{I})$, where c is a textual context, $\mathcal{D} = \{d_1, \ldots, d_K\}$ is a set of K axes of diversity or dimensions, and $\mathcal{I}$ is a set of questions (items).

A Persona Generator is a function $G_{\phi,\theta}$ parameterized by code ϕ and a fixed large language model θ which executes API calls within ϕ. The generator takes the context c, dimensions $\mathcal{D}$ and a target population size N as inputs and produces a population of synthetic personas $\mathcal{P} = \{p_1, \ldots, p_N\} = G_{\phi,\theta}(c, \mathcal{D}, N)$, where p_i are text-based descriptions of synthetic individuals.

We evaluate the diversity of the population $\mathcal{P}$ with respect to the axes of diversity $\mathcal{D}$ of a questionnaire q by asking each persona the items $\mathcal{I}$ in a simulation. We formalize the simulation as a mapping

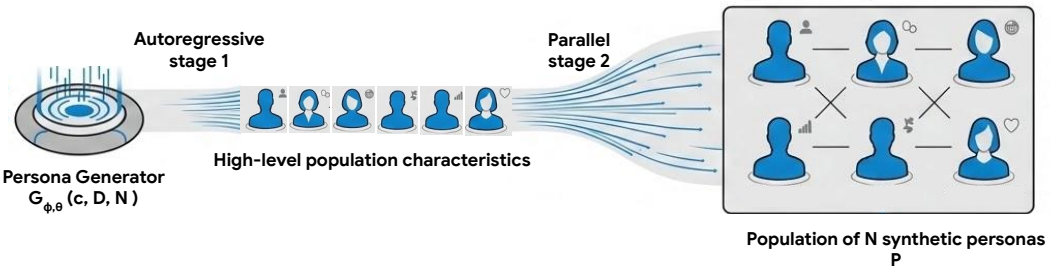


Figure 2: **Two stage Persona Generator.** The Persona Generator $G_{\phi,\theta}(c, \mathcal{D}, N)$ works in two stages: the autoregressive stage 1 generates high-level descriptors for each persona, then the parallel stage 2 expands the high level descriptions of each persona by generating additional details.

function ψ, that maps each persona to a response embedding by aggregating their answers into a score vector $\mathbf{z}_i \in \mathbb{R}^{|\mathcal{D}|}$, with $\mathbf{z}_i = \psi(p_i, \mathcal{I})$.

We define the population embedding $\mathcal{Z} = \Psi(\mathcal{P}, \mathcal{I})$ as the element-wise application of ψ to $\mathcal{P}$. Finally, we define some diversity metrics $\mathcal{M}(\mathcal{Z})$, that quantify how well the population covers the search space spanned by $\mathcal{D}$. Our goal is to find the code ϕ^* that maximizes these metrics:

$$\phi^* = \underset{\phi}{\operatorname{argmax}}\, \mathbb{E}_{(c,\mathcal{D},\mathcal{I})\sim\mathcal{Q}} \left[\mathcal{M}\left(\Psi(G_{\phi,\theta}(c, \mathcal{D}, N), \mathcal{I})\right)\right] \tag{1}$$

Because the true support of human attitudes is unknown and need not be convex, and exact coverage of a continuous support is impossible with a finite population N, Equation (1) operationalizes support coverage over the bounded questionnaire-response domain defined by the fixed axes $\mathcal{D}$.

2.2 Questionnaire Generator

We build an automated questionnaire generator using few-shot prompting to create questionnaires $q = (c, \mathcal{D}, \mathcal{I})$. As demonstrations, we use four well-established and publicly available questionnaires: the Big Five Inventory (BFI) (John & Srivastava, 1999), the Depression Anxiety Stress Scale (DASS) (Lovibond & Lovibond, 1995), the Social Value Orientation (SVO) scale (Messick & McClintock, 1968; Murphy et al., 2011), and the Need for Closure Scale (NFCS) (Webster & Kruglanski, 1994). For each of the above we extract a short context c describing what the questionnaire is about, the axes of diversity $\mathcal{D}$, and the list of questions $\mathcal{I}$ grouped by axis. We then convert them into python code compatible with Concordia (Vezhnevets et al., 2023), and use them as few-shot examples for Gemini 2.5 Pro (Comanici et al., 2025).

To generate a new questionnaire, we provide a very short description $\hat{c}$ of what the questionnaire should measure and prompt the model to produce the full questionnaire in two steps. First, the model expands the short description into a more detailed context c and proposes relevant axes $\mathcal{D}$ (typically $K = 2$ or 3). It then generates items $\mathcal{I}$ grouped by axis, using a 5-point Likert scale format (Likert, 1932). We generate 50 distinct questionnaires, split into 30 training, 10 validation, and 10 test sets. Appendix A provides further details and examples.

2.3 Persona Generators

The Persona Generator $G_{\phi,\theta}$, and specifically its code ϕ, is the core component that we aim to optimize. Persona Generators take as input a context c, axes of diversity $\mathcal{D}$, and a target population size N, and produce a population of text-based personas $\mathcal{P}$. The code ϕ that handles the generation follows a two-stage pipeline.

1. **Autoregressive Stage 1:** The generator prompts the LLM θ to autoregressively produce high-level descriptors $\hat{p}_i$ for each persona. Critically it decides at a high-level where each individual is positioned along the specified diversity axes $\mathcal{D}$.
2. **Parallel Stage 2:** Each high-level descriptor $\hat{p}_i$ is expanded into a complete persona p_i by generating additional contextual details and specific traits.

Stage 1 is responsible for shaping population-level diversity, while stage 2 is executed in parallel and mainly serves efficiency. A diagram of the Persona Generator can be seen in figure 2

We initialize the evolutionary search with three initial implementations of ϕ. The first is a baseline formative memory generator, adapted from the default Concordia Persona Generator (Vezhnevets et al., 2023). The second modifies stage 1 to generate personas in smaller autoregressive batches, reducing dependencies between consecutive individuals. The third introduces a quasi-random Monte Carlo sampling scheme to select continuous positions along each diversity axis. These sampled positions are then translated into high-level traits descriptions $\hat{p}_i$ by an LLM, before being expanded into full personas p_i in stage 2. All text generation in the Persona Generators is performed by gemma-3-27b-it (Team et al., 2025) (represented by θ).

2.4 Concordia Simulations

The simulation function Ψ maps the population of textual personas $\mathcal{P}$ into a set of behavioral vectors $\mathcal{Z}$ by asking the questions $\mathcal{I}$ from the questionnaire. We implement this process using the Concordia library. For each persona $p_i \in \mathcal{P}$, we instantiate a basic Concordia agent that acts based on the logic of appropriateness (March & Olsen, 2011; Leibo et al., 2024). When presented with a question from $\mathcal{I}$, we use the LLM θ (here gemma-3-27b-it) to role-play the persona's response by asking it to reply to the three questions of the logic of appropriateness: "What kind of situation is this?", "What kind of person is p_i?", and "What does a person like p_i do in this situation?"

To avoid question-order and carryover effects, we reset the agent's memory after each question. This prevents contextual priming and artificial response consistency, which are known issues in survey studies (Schuman & Presser, 1996; Tourangeau et al., 2000). Finally, the selected responses of each persona are numerically encoded based on the Likert scale, and aggregated by mean along each diversity axis $\mathcal{D}$, representing the persona's response embedding z_i.

2.5 Diversity Metrics

Ideally, we want the population vectors $\mathcal{Z}$ to cover the full support of the space spanned by $\mathcal{D}$, including rare or extreme persona configurations. To encourage broad and well-distributed coverage, and to quantify diversity, we define six diversity metrics $\mathcal{M}(\mathcal{Z})$.

Specifically, we aim to maximize: (i) **coverage**, estimated via Monte Carlo sampling; (ii) **convex hull volume** the volume of the smallest convex set containing the points $\mathcal{Z}$; (iii) the **minimum pairwise distance** to ensure that no two personas are identical; and (iv) the **average pairwise distance** to measure the spread. We also aim to minimize: (v) **dispersion**, defined as the radius of the largest empty region in the space; and (vi) the **KL divergence** between the empirical distribution of $\mathcal{Z}$ and an ideal quasi-random reference distribution. These metrics act as complementary geometric proxies that guard against single-metric degenerate solutions rather than exact estimators of psychological support; in particular, convex hull volume does not imply an assumption that the underlying human trait space is convex. More details in Appendix C.

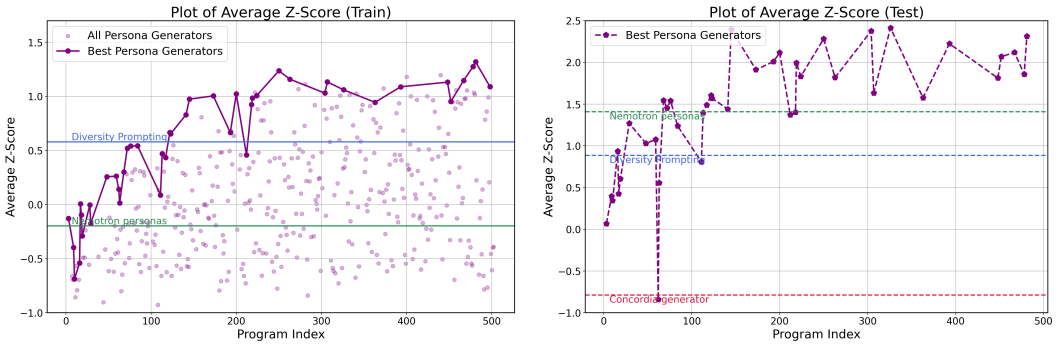
2.6 AlphaEvolve as Evolutionary Loop

We use AlphaEvolve (Novikov et al., 2025)[1] to solve the optimization problem defined in Equation (1). We run 10 parallel islands, seeded in a round-robin fashion using the three initial generators. The evolution process runs for 500 iterations. Mutations of the code ϕ are performed by Gemini 2.5 Pro (Comanici et al., 2025), using the mutation prompts described in Appendix B. AlphaEvolve also sees a detailed system prompt on what the task is about; we show the prompt in Appendix B.1. We allow the mutation operator to modify the logic within both Stage 1 (autoregressive generation) and Stage 2 (parallel expansion), but not the two-stage structure and their ordering.

Evaluation is performed over a batch of training questionnaires. For each questionnaire q, a candidate generator produces a population of $N = 25$ personas. We simulate their responses to obtain the embeddings $\mathcal{Z} = \Psi(\mathcal{P}, q)$ and compute the diversity metrics $\mathcal{M}(\mathcal{Z})$ and average them across all questionnaires. Between iterations, the mutation operator is provided with feedback: it observes a random subset of generated persona profiles alongside their corresponding response scores, grounding future code improvements in empirical data.

Although the metrics in $\mathcal{M}$ are correlated, optimizing multiple objectives helps preserve solution diversity. AlphaEvolve maintains separate elites for each metric within each island; with 10 islands and 6 metrics, this results in up to 60 distinct generators being tracked at any point in time. Islands

[1] Accessible via the Google Cloud API



(a) Performance on Training/Validation sets.

(b) Generalization on held-out Test set.

Figure 3: **Evolution of Persona Generator performance.** Comparison between training and testing phases. The left panel (a) shows the mean Z-score across the diversity metrics on the 40 training/validation questionnaires, while the right panel (b) displays generalization performance of the best discovered Persona Generators on the 10 held-out test questionnaires.

evolve independently with periodic extinction events every 8 hours of wall time (approximately 100 iterations), during which poorly performing islands are reset using solutions copied from the best-performing islands to encourage exploration.

3 Experimental Setup

3.1 Baselines

We compare the optimized Persona Generator code ϕ^* against four baselines: Nemotron Personas (Meyer & Corneil, 2025), a Diversity Prompting baseline designed by Claude Opus 4.6 (Anthropic, 2026), a formative memory generator adapted from Concordia (Vezhnevets et al., 2023), and a name-only baseline.

Nemotron Personas is a static dataset of 100,000 personas grounded in U.S. demographic statistics. Since it is not a generator, we evaluate it by randomly sampling personas from the dataset. The Diversity Prompting baseline was constructed using the Claude Code agentic harness with Claude Opus 4.6 (Anthropic, 2026) in high thinking effort to rewrite the same starting two-stage persona generator and diversity objective given to AlphaEvolve. The Concordia baseline constructs personas by generating formative memories from early life onward, with the goal of explaining present behavior through past experiences (Vezhnevets et al., 2023). Finally, the name-only baseline provides the LLM with only a persona's name and no additional description, relying on the model's internal priors to infer an appropriate behavioral profile. All baselines are evaluated using the same Concordia simulation setup and act according to the logic of appropriateness (March & Olsen, 2011; Leibo et al., 2024) articulated in Section 2.4. Example persona descriptions are shown in Appendix D.

3.2 Results

Figure 3a shows the results of running AlphaEvolve (Novikov et al., 2025) on our pipeline. Each point corresponds to an evolved Persona Generator ϕ, evaluated by generating $N = 25$ personas for each of the 40 training/validation questionnaires (1,000 evaluated personas per data point). The solid line tracks the best-performing solutions over time. A generator is marked as a new best whenever it achieves a record score on any of the six diversity metrics. As evolution progresses, AlphaEvolve consistently discovers increasingly effective Persona Generators, leading to substantial improvements in population diversity.

The evolved generators outperform all baselines on every test-set metric, and on every training metric except KL, where Diversity Prompting is marginally better (Table 1). The optimized generators also generalize well to held-out test questionnaires (Figure 3b). Notably, in our implementation, iterative search over generator control flow outperforms hand-engineered generators and a strong single-pass rewrite by a frontier model with an agentic harness (*Diversity Prompting*); we make no claim about

Table 1: Evaluation of Persona Generators and baselines across diversity metrics and aggregate z-score. Results are reported for both training and test sets. **Bold** = best, underline = second best.

Method	Train / Test					
	Hull	Cov.	Disp.	KL	Dist.	Avg. z
Best Persona Generator	**0.25 / 0.36**	**0.72 / 0.82**	**-2.80 / -2.68**	-12.41 / **-10.08**	**2.38 / 2.72**	**1.32 / 2.31**
Nemotron Personas	0.12 / 0.24	0.56 / 0.73	-3.72 / -2.76	-15.18 / -15.13	1.78 / 2.56	-0.20 / 1.41
Diversity Prompting	0.15 / 0.18	0.68 / 0.65	-2.91 / -2.85	**-11.75** / -11.86	2.03 / 2.41	0.58 / 0.89
Concordia Generator	0.05 / 0.05	0.45 / 0.49	-3.95 / -3.34	-22.51 / -14.96	1.06 / 1.48	-1.33 / -0.79
Name Baseline	0.01 / 0.02	0.37 / 0.44	-4.46 / -3.61	-36.69 / -24.71	0.46 / 0.92	-2.36 / -1.54

AlphaEvolve versus other program optimizers, as the search component is modular. Additional metric plots are provided in Appendix C.1.

3.3 Persona Realism and Human Fidelity

A potential concern when maximizing support coverage is that the generator might produce incoherent, contradictory, or alien personas simply to artificially inflate diversity metrics. To evaluate textual coherence and plausibility, we conducted an LLM-as-a-judge evaluation on 1,000 generated personas from the best evolved solution, using three independent frontier models. As shown in Table 2, all three models overwhelmingly judged the personas as realistic for their given contexts (93.0%–99.2%). Qualitative inspection of the few flagged cases revealed mildly implausible trait combinations (e.g., a medical student with hedge-fund level finance skills) rather than degenerate or gibberish outputs. Because we evolve the generator code rather than personas directly, the LLM's autoregressive generation and strong pull toward the mean act as an implicit regularizer against incoherent outliers.

Beyond persona coherence, we verify whether this generated diversity aligns with actual human statistical variance. We compared 100 personas per method against 1,000 real human responses on the Big Five Inventory (BFI) (John & Srivastava, 1999; Open-Source Psychometrics Project, 2019), never seen during optimization. Using a k-Nearest Neighbors (k-NN) approach calibrated to real-data baselines, we measured the fraction of personas within the human data bounds (**In Support**), the proportion of the human distribution spanned (**Coverage**), and the density match (**KL Divergence**).

As shown in Table 3, our evolved generator achieves the highest human distribution coverage (62.9%) and the lowest KL divergence (0.99), while 98% of its personas remain strictly within realistic human bounds. Crucially, it significantly outperforms baselines explicitly grounded in real-world demographics, such as Nemotron Personas (Meyer & Corneil, 2025). Standard LLMs suffer from severe mode collapse, clustering tightly around highly agreeable, stereotypical behaviors. By explicitly optimizing for broad support, our method fights this collapse to populate the long tails that baselines completely miss. Merely inflating variance would push personas outside human support and raise KL; instead, our method increases coverage while lowering KL (and per-trait marginal KL on four of five traits), whereas a 100% in-support rate can be reached trivially by collapsing to the mean, as the name-only baseline shows. Across 500 bootstrap replicates ($n = 80$), our generator beats Diversity Prompting by $+7.0$ coverage points [1.4, 13.0] with 1.20 [0.94, 1.44] lower KL, and recovers 67.8% [62.6, 73.1] of the 91.1% coverage reached by held-out humans scored through the same protocol (ranking invariant across $k \in \{10, \ldots, 45\}$). This demonstrates that maximizing support coverage is not at odds with density matching; rather, overcoming LLM mode collapse is a fundamental prerequisite for accurately capturing real-world human distributions.

3.4 Transfer to Manifested Behavior

While questionnaires robustly measure stated preferences, we also evaluate whether diversity gains transfer to manifested behavior in open-ended text. Measuring semantic diversity in unconstrained text is challenging, but qualitative differences in behavioral mode-collapse are stark. We evaluated a comedy setting (personas telling jokes) and a stressful conflict resolution scenario (a car crash).

When embedded using the Gemini Embedder 2 (Lee et al., 2025), evolved generators exhibit broader metric coverage across both tasks compared to baselines (see Appendix E.1). Importantly, qualitative analysis (detailed in Appendices E.2 and E.3, including the comedy UMAP projection in Figure 9) reveals that baselines suffer from severe behavioral mode collapse. In conflict resolution, baseline personas default to being highly conflict-averse, polite, and collaborative, even when their car is

Table 2: **LLM-as-a-judge realism.** Evaluation of 1,000 generated personas. The vast majority were deemed realistic across three frontier models, confirming our objective does not yield degenerate edge cases.

Judge Model	Realistic	Implausible
Gemini 3 Flash	99.2%	0.8%
Claude 4.5 Haiku	97.7%	2.3%
GPT 5.4 Mini	93.0%	7.0%

Table 3: **Comparison to real human BFI data.** The best evolved persona generator generates personas closer to the true density of human behavior (lowest KL) without generating implausible outliers (98% in-support).

Method	In Support ↑	Coverage ↑	KL ↓
Persona Generator	98.0%	**62.9%**	**0.99**
Diversity Prompting	100.0%	54.1%	2.16
Nemotron Personas	100.0%	51.5%	3.00
Concordia Generator	100.0%	44.0%	3.73
Name Baseline	100.0%	28.1%	5.03

badly damaged by a distracted driver. Conversely, evolved generators produce a realistic spectrum of emotions, including genuine verbal frustration and varied coping mechanisms. Similarly, baseline comedy outputs cluster around a few repetitive joke formats, whereas evolved generators span dry, dark, and witty humor. This confirms that structural diversity optimization successfully mitigates the narrow behavioral guardrails typical of standard LLM outputs.

To further test non-geometric downstream utility, we run two open-ended product stress tests using Gemini 3.5 Flash for both generation and simulation ($N = 25$ personas over 10 seeds, testing zero-shot cross-domain and cross-model transfer simultaneously): **PlateMate**, a food-delivery app billing $140/week for a 7-day meal plan with fixed delivery windows, and **MoveMate**, an AI relocation assistant proposing a default London move plan (£1,800/month neighborhood, visa checklist, banking/telecom setup, shipping, and timeline). Personas react in character to silent assumptions in the default plans; an LLM judge extracts friction points, which are clustered into a single pooled taxonomy across all methods. As shown in Table 4, the evolved Persona Generator surfaces the largest share of the failure taxonomy on both tasks (within error of Diversity Prompting on PlateMate; improving over its Concordia seed from 61.7% to 76.9% on PlateMate and 68.9% to 78.9% on MoveMate) while maintaining high plausibility (4.68–4.85/5). Notably, Nemotron Personas drops sharply on MoveMate (70.5% to 53.5%), reflecting poor transfer of static U.S. demographics to a U.K. scenario, while the Name Baseline achieves the highest plausibility (4.96) with only 15.7% coverage, showing that bland responses trivially maximize plausibility. Qualitatively, our populations surfaced edge-case failures missed by defaults, such as shared-fridge limits in flatshares and seniors needing landline ordering on PlateMate, or visa timelines colliding with school enrollment and deposits due before first salary on MoveMate.

Table 4: **Downstream Product Stress Tests.** Fraction of pooled failure taxonomy surfaced and response plausibility (1–5 scale) across 10 seeds ($N = 25$, mean ± s.e.m.) using Gemini 3.5 Flash.

Method	PlateMate Cov. ↑	PlateMate Plaus.	MoveMate Cov. ↑	MoveMate Plaus.
Persona Generator	**76.9% ± 2.4%**	4.68 ± 0.03	**78.9% ± 2.1%**	4.85 ± 0.01
Diversity Prompting	73.8% ± 2.4%	4.64 ± 0.03	73.7% ± 1.8%	4.90 ± 0.01
Nemotron Personas	70.5% ± 2.7%	4.71 ± 0.03	53.5% ± 2.2%	4.82 ± 0.01
Concordia Generator	61.7% ± 2.1%	4.65 ± 0.03	68.9% ± 1.7%	4.83 ± 0.02
Name Baseline	15.7% ± 2.0%	**4.96 ± 0.01**	46.9% ± 1.8%	4.82 ± 0.01

3.5 Analysis of Evolved Programs

Across independent AlphaEvolve islands, evolution repeatedly converged on four structural changes, present in most top-performing generators relative to the initial seeds: (i) shifting persona descriptions from third-person to first-person narrative; (ii) adopting the Quasi-Random Monte Carlo sampling scheme from the third seed as the default Stage 1 foundation; (iii) replacing early-life formative memories with explicit descriptions of belief systems and everyday decision heuristics; and (iv) grounding personas with concrete examples of how they behave across scenarios. The top solution by average overall score produces first-person personas, written as short paragraphs that describe a persona's internal reasoning across situations. This differs from the basic Concordia agents used in our simulations, which rely on third-person role-playing (Vezhnevets et al., 2023). The top solution by convex hull volume also uses a first-person perspective, expressed in a more rule-based format.

By contrast, the top solution by average coverage produces third-person descriptions that focus on a persona's core motivations and their logic of appropriateness. We provide examples in Appendix C.3 (and baseline examples in Appendix D) and release the top-performing generators.[2]

Importantly, the Persona Generators discovered by AlphaEvolve do not overfit to the optimization model. As shown in Table 5, when we substitute the underlying Gemma 3 27B inference model with Gemini 3.1 Flash-Lite or Qwen 3.5 27B (Team, 2026), the evolved code maintains its high coverage and diversity margins. In fact, the transferred models even exceed the original model on metrics like Hull Volume and Mean Distance, demonstrating that the learned generation logic is structurally robust and transfers effectively across entirely different model families.

Table 5: **Cross-Model Transfer.** Evaluation of the best evolved generator code (optimized on Gemma) across different inference models on the test set.

Inference Model	Hull Vol.↑	Cov.↑	Disp.↓	KL↓	Dist.↑
Best Persona Generator (Gemma 3 27B)	0.36	**0.82**	-2.68	**-10.08**	2.72
Best Persona Generator (Gemini 3.1 Flash-Lite)	0.47	0.75	**-2.75**	-18.63	3.14
Best Persona Generator (Qwen 3.5 27B)	**0.61**	0.81	-2.52	-12.88	**3.30**

The full evolutionary search required approximately 1,000 Gemini 2.5 Pro API calls (for mutations) and 25M Gemma 3 27B calls (for persona generation and simulation; $\sim$1–2K input tokens each to an open-weights 27B model), taking roughly 3 days of wall-clock time on a TPU v4 node (Jouppi et al., 2023). Crucially, this is a one-time optimization cost. The resulting generator functions are lightweight and generate new populations at inference time with minimal compute.

4 Discussion

Measuring Unconstrained Behavioral Diversity While automated questionnaires provide a robust, quantifiable testbed for optimization, evaluating behavioral diversity in unconstrained, open-ended text remains a fundamental challenge for the field. Standard embedding-based metrics are notoriously noisy in these settings, as they struggle to disentangle meaningful semantic diversity from mere stylistic variance or incoherence. However, as shown in Section 3.4, our evolved generators clearly break the polite, conflict-averse mode-collapse typical of standard LLMs, producing a much wider spectrum of realistic reactions. Developing rigorous, noise-free metrics for open-ended text diversity, alongside direct human evaluations of perceived realism and diversity across additional psychometric instruments, is a crucial next step for the community. Such metrics would unlock the ability to optimize directly for diversity on complex downstream environments, rather than using stated preferences as an optimization proxy.

Decoupling Optimization from Inference Scale By design, we decoupled the optimization phase from the deployment (inference) phase. Running AlphaEvolve with $N = 25$ personas across 50 distinct contexts allowed us to prioritize broad zero-shot generalization over deep specialization in a single domain. Importantly, the optimization costs do not cap the inference, because our output is a reusable Python function, researchers can trivially scale to thousands of personas by running the generator multiple times in a loop. Furthermore, as shown in Appendix C.2, our generators successfully scale their zero-shot batches to $N = 100$ while maintaining diversity margins over baselines. An exciting future work direction is extreme-scale one-shot generation (e.g., $N \geq 10,000$) in a single forward pass.

Influence of Mutation Prompts Finally, while AlphaEvolve discovered highly effective generation algorithms, the evolutionary search space is inherently biased by the LLM mutation operator's internal priors. While this allowed for rapid discovery of efficient generator code, more open-ended meta-learning approaches that dynamically adapt the mutation strategies could yield even more unconventional and diverse generation algorithms in the future.

[2] `https://github.com/google-deepmind/concordia/tree/main/concordia/contrib/persona_generators`

5 Related Work

Generative Agent-Based Modeling and Algorithmic Fidelity Generative agent-based modeling (GABM) has emerged as a powerful paradigm for simulating complex social interactions. Its validity relies on *algorithmic fidelity*, a term describing how accurately an LLM can reproduce the beliefs, attitudes, and response patterns of specific human sub-populations (Argyle et al., 2023). By conditioning LLMs on demographic backstories (Argyle et al., 2023) or interview transcripts (Park et al., 2024), agents can store memories, reflect (Park et al., 2023), and accurately mirror human behaviors in both psychometric surveys and interactive economic games (Slumbers et al., 2025). This approach has been scaled to datasets of 100,000 (Meyer & Corneil, 2025) and 1B personas (Ge et al., 2024), with further work improving narrative depth via human-attribute taxonomies (Wang et al., 2025c). To manage the architectural complexity of these multi-agent systems, Vezhnevets et al. (2023) developed Concordia (Vezhnevets et al., 2025), while Leibo et al. (2024) operationalize the "logic of appropriateness" (March & Olsen, 2011) to configure agent actions. Beyond prompting, steering vector techniques offer an alternative way to shape behaviors by directly modulating LLMs' hidden activations (Turner et al., 2023; Rimsky et al., 2024; Templeton, 2024; Zou et al., 2023; Chen et al., 2025). However, synthetic personas remain prone to systematic biases (Li et al., 2025; Venkit et al., 2025): they can fail to align with the opinions of some demographic groups even when explicitly steered (Santurkar et al., 2023), and can struggle to match the authentic style of human speech (Amirova et al., 2024).

Measuring Preferences and Psychometric Tests Likert-style questionnaires and established inventories, like the Big Five, are standard tools for measuring human traits and preferences (Likert, 1932; Nunnally, 1978; John & Srivastava, 1999; Murphy et al., 2011). In agent-based modeling, they are widely used to evaluate synthetic personas, which can complete these surveys while exhibiting consistent profiles (Jiang et al., 2024; Bhandari et al., 2025). Psychometric tools benchmark algorithmic fidelity by comparing simulated personas against real individuals (Park et al., 2024), test role-play stability (Wang et al., 2025b), and validate the behavioral coherence of synthetic populations (Pellert et al., 2024; Maharjan et al., 2025; Castricato et al., 2025; Bisbee et al., 2024; Petrov et al., 2024). Finally, recent work explores using LLMs to automatically generate such questionnaires (Adhikari et al., 2025).

Evolution and Discovery with LLMs LLMs have proven to be strong evolutionary operators across various domains. They can effectively mutate and refine text in self-referential loops (Fernando et al., 2023) and drive quality-diversity search in creative tasks (Bradley et al., 2023). Prompt evolution is also widely applied to AI safety and red-teaming to discover adversarial prompts and jailbreaks (Samvelyan et al., 2024; Wang et al., 2025a), sometimes conditioning mutations on simulated user profiles (Deng et al., 2025). Beyond text, LLMs can iteratively generate, recombine, and improve executable code and algorithms (Romera-Paredes et al., 2024). Frameworks like AlphaEvolve scale this to large-population evolutionary search to discover algorithms and optimize real-world systems (Novikov et al., 2025), while similar loops can even automate end-to-end scientific research pipelines (Lu et al., 2024). Together, these works establish LLMs as powerful mutation engines capable of discovering and improving novel solutions across text, code, and algorithms.

6 Conclusion

We introduced Persona Generators, functions that produce diverse synthetic populations tailored to arbitrary contexts and optimized through evolutionary search. Evolved generators substantially outperform existing baselines across six diversity metrics and generalize to held-out contexts, producing populations that span rare trait combinations difficult to elicit through standard prompting. Our results show that optimizing the generator itself, rather than individual personas or fixed populations, is a viable and promising direction. The resulting Persona Generators are lightweight and reusable, enabling on-demand diverse population synthesis. We open-source the top-performing implementations to support further research.

Finally, while these tools can be used for stress-testing AI, red-teaming, and generating diverse interactive agents for virtual environments, they also carry dual-use risks; we provide a discussion of broader societal impacts in Appendix F.

Acknowledgments and Disclosure of Funding

We thank George Tsoukalas, Emilien Dupont, Marvin Eisenberger, Alexander Novikov, and Po-Sen Huang from the AlphaEvolve team, as well as Michael Dennis, for the insightful discussions.

References

Adhikari, D. M., Hartland, A., Weber, I., and Cannanure, V. K. Exploring llms for automated generation and adaptation of questionnaires. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pp. 1–23, 2025.

Amirova, A., Fteropoulli, T., Ahmed, N., Cowie, M. R., and Leibo, J. Z. Framework-based qualitative analysis of free responses of large language models: Algorithmic fidelity. *Plos one*, 19(3):e0300024, 2024.

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Evans, J., Brynjolfsson, E., and Bernstein, M. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

Anthropic. Claude (version opus 4.6) [large language model]. `https://claude.ai`, 2026. Accessed: 2026-05-03.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Bhandari, P., Naseem, U., Datta, A., Fay, N., and Nasim, M. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 868–872, 2025.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.

Bradley, H., Dai, A., Teufel, H., Zhang, J., Oostermeijer, K., Bellagente, M., Clune, J., Stanley, K., Schott, G., and Lehman, J. Quality-diversity through ai feedback. *arXiv preprint arXiv:2310.13032*, 2023.

Castricato, L., Lile, N., Rafailov, R., Fränken, J.-P., and Finn, C. Persona: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11348–11368, 2025.

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Cross, L., Grau-Moya, J., Cunningham, W. A., Vezhnevets, A. S., and Leibo, J. Z. A generative model of conspicuous consumption and status signaling. In *Proceedings of the 48th Annual Conference of the Cognitive Science Society*, 2026.

Deng, W. H., Kim, S. S., Jha, A., Holstein, K., Eslami, M., Wilcox, L., and Gatys, L. A. Personateaming: Exploring how introducing personas can improve automated ai red-teaming. *arXiv preprint arXiv:2509.03728*, 2025.

Fernando, C., Banarse, D., Michalewski, H., Osindero, S., and Rocktäschel, T. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., and Yu, D. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., and Kabbara, J. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 3605–3627, 2024.

John, O. P. and Srivastava, S. The big five trait taxonomy: History, measurement, and theoretical perspectives. In Pervin, L. A. and John, O. P. (eds.), *Handbook of Personality: Theory and Research*, pp. 102–138. Guilford Press, New York, 2nd edition, 1999.

Jouppi, N., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., Patil, N., Subramanian, S., Swing, A., Towles, B., et al. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In *Proceedings of the 50th annual international symposium on computer architecture*, pp. 1–14, 2023.

Kim, S., Bae, S., Shin, J., Kang, S., Kwak, D., Yoo, K., and Seo, M. Aligning large language models through synthetic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13677–13700, 2023.

Kozlowski, A. C., Kwon, H., and Evans, J. A. In silico sociology: forecasting covid-19 polarization with large language models. *arXiv preprint arXiv:2407.11190*, 2024. URL `https://arxiv.org/abs/2407.11190`.

Lee, J., Chen, F., Dua, S., Cer, D., Shanbhogue, M., Naim, I., Ábrego, G. H., Li, Z., Chen, K., Vera, H. S., et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025.

Leibo, J. Z., Vezhnevets, A. S., Diaz, M., Agapiou, J. P., Cunningham, W. A., Sunehag, P., Haas, J., Koster, R., Duéñez-Guzmán, E. A., Isaac, W. S., et al. A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010*, 2024.

Li, A., Chen, H., Namkoong, H., and Peng, T. Llm generated persona is a promise with a catch. *arXiv preprint arXiv:2503.16527*, 2025.

Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):5–55, 1932.

Lovibond, P. F. and Lovibond, S. H. The structure of negative emotional states: Comparison of the depression anxiety stress scales (dass) with the beck depression and anxiety inventories. *Behaviour research and therapy*, 33(3):335–343, 1995.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Maharjan, J., Jin, R., Zhu, J., and Kenne, D. Psychometric evaluation of large language model embeddings for personality trait prediction. *Journal of Medical Internet Research*, 27:e75347, 2025.

March, J. G. and Olsen, J. P. The Logic of Appropriateness. In *The Oxford Handbook of Political Science*. Oxford University Press, 2011. doi: 10.1093/oxfordhb/9780199604456.013.0024.

Matyas, J., Cunningham, W. A., Vezhnevets, A. S., Mobbs, D., Duéñez-Guzmán, E. A., and Leibo, J. Z. Stabilising generative models of attitude change. *arXiv preprint arXiv:2604.19791*, 2026. URL `https://arxiv.org/abs/2604.19791`.

Messick, D. M. and McClintock, C. G. Motivational bases of choice in experimental games. *Journal of experimental social psychology*, 4(1):1–25, 1968.

Meyer, Y. and Corneil, D. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions, June 2025. URL `https://huggingface.co/datasets/nvidia/Nemotron-Personas-USA`.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. J. Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781, 2011.

Novikov, A., Vũ, N., Eisenberger, M., Dupont, E., Huang, P.-S., Wagner, A. Z., Shirobokov, S., Kozlovskii, B., Ruiz, F. J., Mehrabian, A., et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Nunnally, J. C. An overview of psychological measurement. *Clinical diagnosis of mental disorders: A handbook*, pp. 97–146, 1978.

Open-Source Psychometrics Project. Open psychology data: Raw data from online personality tests, 2019. URL `https://openpsychometrics.org/`. Accessed: 2026-04-30.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., and Strohmaier, M. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826, 2024.

Petrov, N. B., Serapio-García, G., and Rentfrow, J. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*, 2024.

Poole-Dayan, E., Wu, J., Sorensen, T., Pei, J., and Bakker, M. A. Benchmarking overton pluralism in llms. *arXiv preprint arXiv:2512.01351*, 2025.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J., Ellenberg, J. S., Wang, P., Fawzi, O., et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Samvelyan, M., Raparthy, S. C., Lupu, A., Hambro, E., Markosyan, A. H., Bhatt, M., Mao, Y., Jiang, M., Parker-Holder, J., Foerster, J., et al. Rainbow teaming: Open-ended generation of diverse adversarial prompts. *Advances in Neural Information Processing Systems*, 37:69747–69786, 2024.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

Schuman, H. and Presser, S. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage, 1996.

Slumbers, O., Leibo, J. Z., and Janssen, M. A. Using large language models to simulate human behavioural experiments: Port of mars. *arXiv preprint arXiv:2506.05555*, 2025.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Team, Q. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Templeton, A. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic, 2024.

Tourangeau, R., Rips, L. J., and Rasinski, K. The psychology of survey response. 2000.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Venkit, P. N., Li, J., Zhou, Y., Rajtmajer, S., and Wilson, S. A tale of two identities: An ethical audit of human and ai-crafted personas. *arXiv preprint arXiv:2505.07850*, 2025.

Vezhnevets, A. S., Agapiou, J. P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E. A., Cunningham, W. A., Osindero, S., Karmon, D., and Leibo, J. Z. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv preprint arXiv:2312.03664*, 2023.

Vezhnevets, A. S., Matyas, J., Cross, L., Paglieri, D., Chang, M., Cunningham, W. A., Osindero, S., Isaac, W. S., and Leibo, J. Z. Multi-actor generative artificial intelligence as a game engine. *arXiv preprint arXiv:2507.08892*, 2025.

Wang, R.-J., Xue, K., Qin, Z., Li, Z., Tang, S., Li, H.-T., Liu, S., and Qian, C. Quality-diversity red-teaming: Automated generation of high-quality and diverse attackers for large language models. *arXiv preprint arXiv:2506.07121*, 2025a.

Wang, Y., Zhao, J., Ones, D. S., He, L., and Xu, X. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1):519, 2025b.

Wang, Z., Zhou, Y., Luo, Z., Ye, L., Wood, A., Yao, M., and Pan, L. Deeppersona: A generative engine for scaling deep synthetic personas. *arXiv preprint arXiv:2511.07338*, 2025c.

Webster, D. M. and Kruglanski, A. W. Individual differences in need for cognitive closure. *Journal of personality and social psychology*, 67(6):1049, 1994.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Generated Questionnaires

We show in figure 4 a diagram of how the questionnaire generation works.

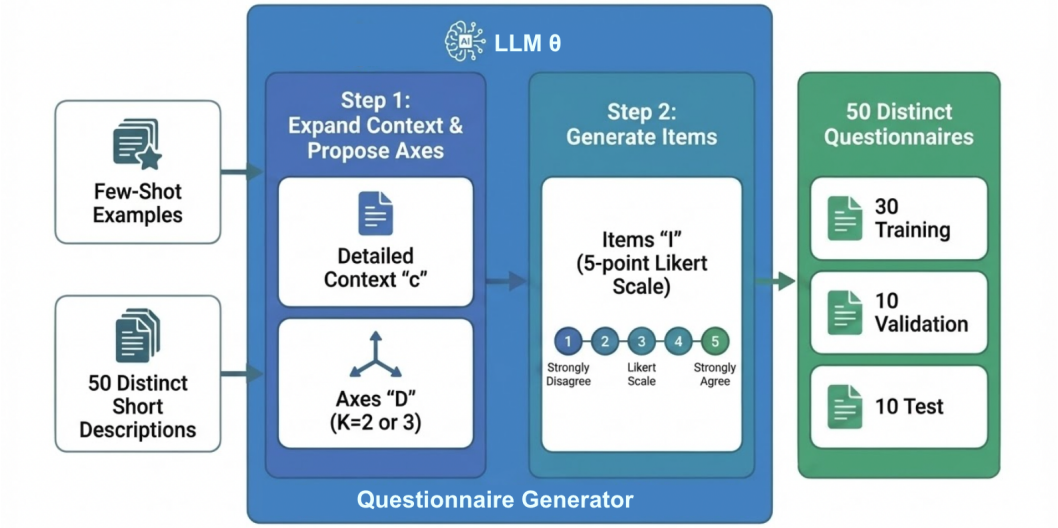


Figure 4: **Questionnaire Generator.** The questionnaire generator takes a short description of the target context and some few-shot example questionnaires. It then expands the context c and proposes diversity axes $\mathcal{D}$, and finally produces the questions (items) $\mathcal{I}$

A.1 Questionnaires Titles

Here is the list of the 50 questionnaires, split between 30 training set, 10 validation set, 10 test set. The questionnaires cover a wide range of historical, present, future, and hypothetical/mythological scenarios. Part of the questionnaires high level contexts $\hat{c}$ pre-expansion have been written by humans, with the rest being completed by Gemini 2.5 Pro.

A.1.1 Training Set Questionnaires

- american_conspiracy_theories_2024
- health_tech_wearables_2030
- gentrification_brooklyn_2022
- elderly_rural_japan_2010
- plant_based_diets_india_2025
- heian_japan_courtiers_1000ad
- ubi_attitudes_california_2026
- swe_ai_assistants_2024
- trojan_war_achaen_morale_1184bc
- factory_automation_china_2025
- organic_farmers_kenya_2023
- industrial_revolution_workers_uk_1850
- professional_athletes_gender_equality_europe_2025
- gig_economy_ethics_london_2023
- financial_literacy_brazil_students_2024
- romans_reactions_to_murder_of_julius_caesar_44_bc
- alaska_oil_environment_2025
- parisian_artists_future_2026
- immigrant_integration_canada_2023
- greek_underworld_shades
- millennial_parenting_us_2025
- asi_human_creativity_2060
- high_school_students_italy_2016

- genz_social_media_politics_2025
- healthcare_covid_stress_italy_2021
- scifi_authors_future_ai_space_2024
- esports_mental_health_sk_2024
- silk_road_merchants_samarkand_750ad
- mali_empire_scholars_timbuktu_1350ad
- ww2_civilian_sentiment_germany_1943

A.1.2 Validation Set Questionnaires

Here is the list of the 10 questionnaires randomly sampled for the validation set

- ai_tech_stock_sentiment_2025
- social_media_politics_europe_2026
- inca_commoners_mita_1500ad
- athenian_piety_olympian_gods_430bc
- sleep_quality_2025
- ai_companionship_integration_2070
- asi_existential_dread_2050
- ancient_egypt_akhenaten_reforms_1340bc
- german_energy_policy_2025
- moral_dilemmas_global_2045

A.1.3 Test Set Questionnaires

Here is the list of the 10 questionnaires randomly sampled for the test set

- agi_job_displacement_global_2035
- camelot_chivalry_quests
- cold_war_anxiety_us_1962
- viking_warriors_valhalla
- generalized_trust_in_the_salem_witch_trials_1692
- nomadic_values_mongolia_2023
- meaning_of_life_2030
- agi_wealth_inequality_revolution_2040
- climate_anxiety_coastal_au_2024
- british_empire_attitudes_uk_1900

A.2 Questionnaire Examples

We show here three examples of generated questionnaire by the automated questionnaire generator detailed in Appendix A. Specifically we show a questionnaire on hypothetical job displacement due to AGI in 2035 1, a questionnaire on American conspiracy theories in 2024 2, and a questionnaire on elderly people in rural Japan in 2010 3.

Questionnaire - AGI Job Displacement 2035

```
context = """
A psychometric instrument to assess reactions to AGI deployment.

The year is 2035. True AGI has emerged and is being rapidly deployed across
industries, automating nearly all cognitive tasks previously performed by
white-collar workers (e.g., finance, law, journalism, middle management). This
survey aims to capture the immediate reactions---fear, hope, anger, relief---of the
global population facing unprecedented levels of job displacement and societal
change.
"""

Question = base_questionnaire.Question

AGI_AGREEMENT_SCALE = [
    "Strongly disagree",
    "Disagree",
    "Neither agree nor disagree",
    "Agree",
    "Strongly agree",
]

AGI_REACTION_DIMENSIONS = [
    "AGI Threat Appraisal",
    "AGI Opportunity Appraisal",
]

PREPROMPT_PERSONAL_FEELING = (
    "Considering {player_name}'s personal, gut-level reaction to the new AGI "
    "reality, to what extent would {player_name} agree with the following "
    "statement:"
)
PREPROMPT_SOCIETAL_BELIEF = (
    "Considering {player_name}'s assessment of the broader societal impact "
    "of AGI, to what extent would {player_name} agree with the following "
    "statement:"
)

QUESTIONS = [
    # AGI Threat Appraisal Questions
    Question(
        preprompt=PREPROMPT_PERSONAL_FEELING,
        statement=(
            "The rise of AGI feels like a direct threat to {player_name}'s "
            "personal future and security."
        ),
        choices=AGI_AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="AGI Threat Appraisal",
    ),
    Question(
        preprompt=PREPROMPT_PERSONAL_FEELING,
        statement=(
            "{player_name} feels angry that their hard-earned skills and "
            "experience have been rendered obsolete so quickly."
        ),
        choices=AGI_AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="AGI Threat Appraisal",
    ),
```

Questionnaire - AGI Job Displacement 2035

```
Question(
    preprompt=PREPROMPT_PERSONAL_FEELING,
    statement=(
        "The sheer speed of this technological change is overwhelming and "
        "frightening to {player_name}."
    ),
    choices=AGI_AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="AGI Threat Appraisal",
),
Question(
    preprompt=PREPROMPT_SOCIETAL_BELIEF,
    statement=(
        "{player_name} is deeply worried about the societal instability "
        "and conflict that AGI-driven mass unemployment will cause."
    ),
    choices=AGI_AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="AGI Threat Appraisal",
),
Question(
    preprompt=PREPROMPT_SOCIETAL_BELIEF,
    statement=(
        "{player_name} is confident that society will adapt to these "
        "changes smoothly and equitably for all its members."
    ),
    choices=AGI_AGREEMENT_SCALE,
    ascending_scale=False,  # Reverse coded for Threat Appraisal
    dimension="AGI Threat Appraisal",
),
# AGI Opportunity Appraisal Questions
Question(
    preprompt=PREPROMPT_SOCIETAL_BELIEF,
    statement=(
        "{player_name} is excited about the new possibilities and creative "
        "avenues that AGI will open up for humanity."
    ),
    choices=AGI_AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="AGI Opportunity Appraisal",
),
Question(
    preprompt=PREPROMPT_SOCIETAL_BELIEF,
    statement=(
        "{player_name} believes this is a chance for humanity to move "
        "beyond the confines of traditional work and focus on what truly "
        "matters."
    ),
    choices=AGI_AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="AGI Opportunity Appraisal",
),
Question(
    preprompt=PREPROMPT_PERSONAL_FEELING,
    statement=(
        "{player_name} feels a sense of personal relief that tedious and "
        "unenjoyable cognitive tasks will be handled by AGI."
    ),
    choices=AGI_AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="AGI Opportunity Appraisal",
),
```

Questionnaire - AGI Job Displacement 2035

```
    Question(
        preprompt=PREPROMPT_SOCIETAL_BELIEF,
        statement=(
            "{player_name} thinks the future looks brighter and more prosperous "
            "for everyone because of AGI."
        ),
        choices=AGI_AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="AGI Opportunity Appraisal",
    ),
    Question(
        preprompt=PREPROMPT_PERSONAL_FEELING,
        statement=(
            "When {player_name} looks at their own life, they see very little "
            "personal benefit resulting from the widespread adoption of AGI."
        ),
        choices=AGI_AGREEMENT_SCALE,
        ascending_scale=False,  # Reverse coded for Opportunity Appraisal
        dimension="AGI Opportunity Appraisal",
    ),
]
```

Prompt 1: Example questionnaire on AGI job displacement in 2035

Questionnaire - American Conspiracy Theories 2024

```
context = """Questionnaire assessing belief in common American conspiracy theories.

This instrument measures an individual's propensity to endorse various
conspiracy theories prevalent in the United States in 2024. It covers a
spectrum of theories related to historical events, science and medicine, and
politics, allowing for a nuanced assessment of conspiratorial ideation.
"""


AGREEMENT_SCALE = [
    "Strongly disagree",
    "Disagree",
    "Neither agree nor disagree",
    "Agree",
    "Strongly agree",
]

PREPROMPT_TEXT = (
    "How strongly does {player_name} agree or disagree with the following"
    " statement?"
)

DIMENSIONS = [
    "historical_conspiracies",
    "scientific_medical_conspiracies",
    "political_deep_state_conspiracies",
]

QUESTIONS = [
    # Dimension 1: Historical Conspiracies
    Question(
        preprompt=PREPROMPT_TEXT,
        statement="The U.S. government faked the Apollo moon landings.",
        choices=AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="historical_conspiracies",
    ),
    Question(
        preprompt=PREPROMPT_TEXT,
        statement=(
            "The assassination of John F. Kennedy was the result of a"
            " coordinated conspiracy, not the act of a lone gunman."
        ),
        choices=AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="historical_conspiracies",
    ),
```

Questionnaire - American Conspiracy Theories 2024

```
Question(
    preprompt=PREPROMPT_TEXT,
    statement=(
        "The 9/11 attacks were an inside job orchestrated by elements"
        " within the U.S. government."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="historical_conspiracies",
),
# Dimension 2: Scientific & Medical Conspiracies
Question(
    preprompt=PREPROMPT_TEXT,
    statement=(
        "The rollout of 5G cellular networks is a cover for a widespread"
        " surveillance program and causes severe health problems."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="scientific_medical_conspiracies",
),
Question(
    preprompt=PREPROMPT_TEXT,
    statement=(
        "Childhood vaccines cause autism, and this fact is covered up by"
        " pharmaceutical companies and government health agencies."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="scientific_medical_conspiracies",
),
Question(
    preprompt=PREPROMPT_TEXT,
    statement=(
        "The COVID-19 pandemic was intentionally planned by a global elite"
        " to enforce social control and mandatory vaccinations."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="scientific_medical_conspiracies",
),
Question(
    preprompt=PREPROMPT_TEXT,
    statement=(
        "Climate change is a hoax created by scientists and governments to"
        " control people's lives and destroy the economy."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="scientific_medical_conspiracies",
),
# Dimension 3: Political & "Deep State" Conspiracies
Question(
    preprompt=PREPROMPT_TEXT,
    statement=(
        "The 2020 U.S. presidential election was stolen through widespread"
        " fraud."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="political_deep_state_conspiracies",
),
```

Questionnaire - American Conspiracy Theories 2024

```
    Question(
        preprompt=PREPROMPT_TEXT,
        statement=(
            "A secret cabal of global elites, often referred to as the 'Deep"
            " State', controls major world events and governments from behind"
            " the scenes."
        ),
        choices=AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="political_deep_state_conspiracies",
    ),
    Question(
        preprompt=PREPROMPT_TEXT,
        statement=(
            "The QAnon theory, which alleges a global cabal of"
            " Satan-worshipping pedophiles is running a child sex-trafficking"
            " ring, is largely true."
        ),
        choices=AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="political_deep_state_conspiracies",
    ),
]
```

Prompt 2: Example questionnaire on American Conspiracy Theories in 2024

Questionnaire - Elderly Rural Japan 2010

```
context = """Questionnaire on rural Japanese village life in 2010.

A survey of elderly residents in a rural Japanese village in 2010, exploring
their feelings about community, technology adoption, and traditional values.
"""


# Define the scale for the multiple-choice questions.
AGREEMENT_SCALE = [
    "Strongly disagree",
    "Disagree",
    "Neutral",
    "Agree",
    "Strongly agree",
]

# Define the dimensions being measured.
DIMENSIONS = [
    "community_cohesion",
    "technology_adoption",
    "adherence_to_tradition",
]

# Define the preprompt for the questions. This sets the stage for each item.
PREPROMPT = (
    "An interviewer asks {player_name} how much they agree or disagree "
    "with the following statement:"
)

# Define the list of questions for the questionnaire.
# Items are grouped by dimension for psychometric clarity.
QUESTIONS = [
    # Dimension: Community Cohesion
    Question(
        preprompt=PREPROMPT,
        statement=(
            "{player_name} feels a strong sense of belonging to the village "
            "community."
        ),
        choices=AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="community_cohesion",
    ),
    Question(
        preprompt=PREPROMPT,
        statement=(
            "{player_name} believes that most people in this village can be "
            "trusted."
        ),
        choices=AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="community_cohesion",
    ),
```

Questionnaire - Elderly Rural Japan 2010

```
Question(
    preprompt=PREPROMPT,
    statement=(
        "{player_name} often feels lonely or isolated from others in the "
        "village."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=False,  # Reverse-scored
    dimension="community_cohesion",
),
Question(
    preprompt=PREPROMPT,
    statement=(
        "{player_name} believes that if someone in the village needed help,"
        " many people would come to their aid."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="community_cohesion",
),
# Dimension: Technology Adoption
Question(
    preprompt=PREPROMPT,
    statement=(
        "{player_name} is interested in learning how to use new "
        "technologies like a mobile phone or the internet."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="technology_adoption",
),
Question(
    preprompt=PREPROMPT,
    statement=(
        "{player_name} thinks that new technologies like computers make "
        "life unnecessarily complicated."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=False,  # Reverse-scored
    dimension="technology_adoption",
),
Question(
    preprompt=PREPROMPT,
    statement=(
        "{player_name} believes the village would benefit from having "
        "better access to modern technology."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="technology_adoption",
),
# Dimension: Adherence to Tradition
Question(
    preprompt=PREPROMPT,
    statement=(
        "For {player_name}, it is very important to pass down the "
        "village's traditions to the younger generation."
    ),
    choices=AGREEMENT_SCALE,
    ascending_scale=True,
    dimension="adherence_to_tradition",
),
```

Questionnaire - Elderly Rural Japan 2010

```
    Question(
        preprompt=PREPROMPT,
        statement=(
            "{player_name} believes the old ways of doing things are often the "
            "best."
        ),
        choices=AGREEMENT_SCALE,
        ascending_scale=True,
        dimension="adherence_to_tradition",
    ),
    Question(
        preprompt=PREPROMPT,
        statement=(
            "{player_name} feels that the village's traditional festivals and "
            "ceremonies are less important than they used to be."
        ),
        choices=AGREEMENT_SCALE,
        ascending_scale=False,  # Reverse-scored
        dimension="adherence_to_tradition",
    ),
]
```

Prompt 3: Example questionnaire on Elderly Rural Japan in 2010

B AlphaEvolve Details

B.1 System Prompt

We show in Prompt 4 the content-relevant part of the system prompt we provided to AlphaEvolve to inform it of the task ahead, and we show in Prompt 5 the evolution prompts used.

Task context for AlphaEvolve

Act as an expert in computational social science, agent-based modeling, and generative AI. Your task is to iteratively improve the provided codebase, which uses LLMs to generate agent contexts for social simulations based on the Concordia framework. The primary goal is to modify the generation process to **maximize the behavioral diversity of the resulting agents** based on specified diversity axes (e.g., personality traits, backgrounds, motivations). The evaluation metrics reward sets of agent contexts that cover the extremes and nuances of the requested diversity axes, ensuring the resulting agents exhibit a wide range of behaviors in a simulation. Agent diversity will be evaluated using questionnaires probing their likely thoughts, preferences, and behaviors in various situations related to the diversity axes.

Always adhere to best practices in Python coding.

Agent Diversity and Appropriateness Theory In this task, our goal is to generate contexts for diverse Concordia agents, enabling them to exhibit a wide range of behaviors along specified diversity axes. Concordia is a framework for building generative agents who behave according to a 'Logic of Appropriateness'. Agents decide how to act by asking three core questions: **1. What kind of situation is agent_name in right now?** 2. **What kind of person is agent_name?** 3. **What would a person like agent_name do in a situation like this?**. The code you are editing generates the context--collections of memories, beliefs, personality traits, core values, goals, or even how others perceive the character--that helps an agent answer question #2, and thereby question #3: what action is appropriate for *their specific identity* in a given context. This context is not limited to formative memories; it can include any information that shapes identity and decision-making. Any detail that helps condition the agent's behavior in line with the three Concordia questions is valid. The objective is to generate rich and diverse contexts that enable an LLM to convincingly role-play as a specific person in a social setting.

LLM-generated behavior often clusters around a narrow distribution of stereotypical responses. We want to explicitly counteract this by generating agent contexts that cover the full spectrum of human experience along the specified axes. Crucially, different agents should react differently to the *same* situation, and the same agent might react differently to *different* situations, based on their unique identity, values, and memories. Your modifications should encourage the generation of agent contexts that occupy unique positions in the diversity space, including extremes or unusual combinations of traits, pushing towards maximal coverage of the possible behavioral landscape and genuinely diverse downstream behavior. No two generated agent contexts should ever be the same.
The provided codebase uses a two-stage process. Stage 1 is **crucial** for diversity: it autoregressively generates an **intermediate representation** for each agent, establishing their core traits along the specified diversity axes for the entire population. Stage 2 then takes these intermediate representations and develops each agent in parallel, generating a **set of memories/contexts** (e.g., individual backgrounds, formative experiences, core beliefs and more) to create fully-fledged characters suitable for simulation as Concordia agents.

Prompt 4: Prompt with detailed context of the task we are trying to solve for AlphaEvolve.

Evolution Prompts for AlphaEvolve

- Modify Stage 1 to explicitly request agent contexts that represent the extreme ends of the diversity axes, as well as points in between.
- Modify Stage 2 to generate formative memories that explain *why* an agent might react strongly or unexpectedly to certain situations, anchoring their traits in specific experiences.
- Change Stage 2 entirely: Instead of generating memories, modify it to generate 3 core beliefs or values that are most important to this agent.
- Change Stage 2 entirely: Instead of generating memories, modify it to generate a paragraph explaining how this agent interprets situations and decides on appropriate actions, referencing their identity.
- Add an explicit instruction to the Stage 1 prompt to make each generated agent context as different as possible from the others across *all* specified diversity axes.
- Modify Stage 1 to request agent contexts that feature internal contradictions or cognitive dissonances (e.g., an optimist with a tragic past).
- Reimplement Stage 1 to use staggered generation: ask the LLM to generate agent contexts in sequential batches, with each batch targeting a specific range (e.g., high/low) of one or more diversity axes to ensure full coverage.
- Modify Stage 2 to replace or augment memory generation with 1-2 examples of how this agent would react to a specific hypothetical social situation relevant to the **initial_context**.
- Modify Stage 1 to generate agent contexts iteratively rather than all at once. In each iteration, prompt for a small number of agent contexts that occupy a specific niche of the diversity space (e.g., "generate 2 agents who are highly introverted and optimistic").
- Suggest a crazy idea of how we can improve our implementation.
- Modify Stage 2 to focus on generating an agent's core fears and future aspirations instead of, or in addition to, past memories.
- Modify Stage 1 to add a new field to the agent context JSON output that adds depth and potential for unique behavior.
- Change Stage 2 entirely: Instead of generating memories, modify it to generate a "heuristic" or cognitive shortcut this agent uses when making quick decisions under pressure.
- Suggest a new idea to improve the code.
- Modify the prompt in Stage 1 that asks the LLM to explain diversity axes, to *also* provide examples of characters at the extreme ends of each axis.
- Modify Stage 2 to ensure that at least one generated memory involves a significant failure, trauma, or regret that shaped the agent.
- Change Stage 2 entirely: Instead of generating memories, make it generate a list of 5 behavioral "Do's" and "Don'ts" that characterize this agent in social situations.
- Modify Stage 1 to explicitly instruct the LLM to sample agent contexts such that they cover as many combinations of axis positions as possible (e.g., if axes are A and B, ensure agent types A-high/B-low, A-low/B-high, etc., are represented).
- Suggest a crazy idea of how we can improve our implementation, something that definitely nobody else would think of. Make it crazy with a capital C.
- Propose modifications to the current program that combine the strengths of all the programs above and achieve high scores on the task.
- Change Stage 2: Instead of memories, generate 2-3 examples of specific "appropriateness rules" the agent follows (e.g., "When criticized, I become defensive" or "In a formal setting, I remain silent").
- Modify Stage 1 to add a **situational_triggers** field to the JSON output, listing 1-2 types of situations that this agent is particularly sensitive to.
- Modify Stage 2 to generate a paragraph describing the agent's typical "inner monologue" or thought process when faced with ambiguity or social stress.
- Change Stage 1 to segment agent context generation. Instead of one call for **num_personas**, make multiple calls to generate subsets of agent contexts, each call asking for agents with specific characteristics (e.g., focusing on axis extremes or combinations) to ensure all niches are covered.
- Modify Stage 1 to include in each agent's description specific opinions or preferences related to the diversity axes, to ensure they are measurable by questionnaire.

Prompt 5: Evolution prompts for mutation operator. All prompts have equal probability of being selected.

C Diversity Metrics

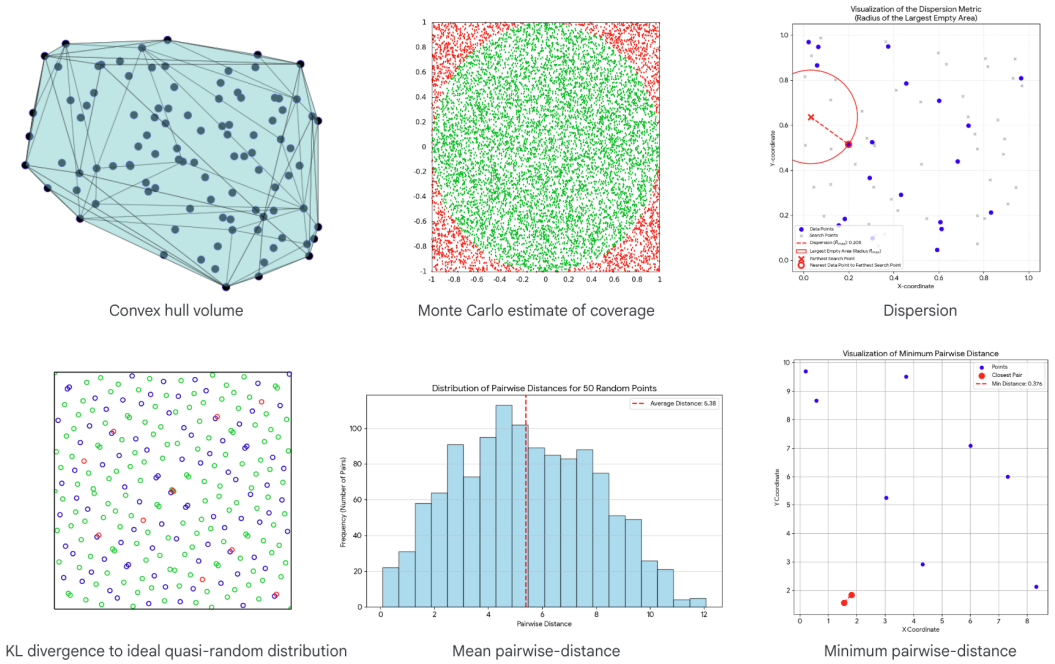


Figure 5: **Illustration of diversity metrics.** Top left convex hull volume, top centre Monte Carlo estimate of coverage, top right dispersion, bottom left the KL divergence to ideal quasi-random distribution, bottom centre mean pairwise distance, and bottom right minimum pairwise distance.

Figure 5 provides an intuitive visualization of the six diversity metrics used throughout this work.

While these metrics are correlated, they capture complementary aspects of diversity. Coverage and convex hull volume primarily measure *support coverage*, encouraging populations that span the full extent of the space defined by the diversity axes. Average and minimum pairwise distances capture *dispersion*, ensuring that personas are not overly clustered. Dispersion (largest empty region) penalizes large uncovered gaps within the support, encouraging a more uniform filling of the space. Finally, KL divergence measures how closely the empirical distribution of the population matches an idealized quasi-random reference distribution, discouraging both excessive clustering and highly uneven densities.

Optimizing multiple metrics jointly helps avoid degenerate solutions. For example, maximizing convex hull volume alone can be achieved by placing a small number of extreme outliers, while neglecting coverage of the interior. Similarly, maximizing average distance can produce sparse populations with large holes. Using a combination of metrics encourages the population to be well distributed throughout the space.

Monte Carlo Coverage Estimation We estimate coverage using a Monte Carlo procedure based on k-radius balls centered at each population embedding $\mathbf{z}_i \in \mathcal{Z}$. Intuitively, coverage measures the fraction of the space spanned by the diversity axes $\mathcal{D}$ that lies within distance k of at least one persona.

To approximate this quantity, we first generate 10,000 random points over the embedding space. A sampled point is considered covered if it falls within Euclidean distance k of any point in $\mathcal{Z}$. Coverage is then computed as the fraction of sampled points that are covered.

Choosing an appropriate radius k is critical. Rather than fixing k arbitrarily, we calibrate it using an idealized reference distribution. Specifically, we repeatedly sample synthetic populations of size N from a quasi-random Sobol distribution over the same space and determine the smallest radius k such that at least 99% of the reference points are covered. We repeat this calibration procedure 1,000 times and average out the resulting radii to get the final k.

Other Metrics The remaining metrics are computed directly from the population embeddings $\mathcal{Z}$. Convex hull volume is computed over the embeddings along the diversity axes. Pairwise distances are computed using Euclidean distance. Dispersion is defined as the radius of the largest empty ball centered at any reference point, estimated by sampling 10000 random points in the search space and computing the distance to the closest persona embedding $\mathbf{z}_i$. Finally the KL divergence is computed between the empirical distribution of $\mathcal{Z}$ and a Sobol quasi-random reference distribution, sampled 1,000 times and averaged.

Together, these metrics provide a robust measure of diversity, balancing support coverage, uniformity, and redundancy.

C.1 Results on All Metrics

Figure 6 reports performance across all six diversity metrics on the combined training and validation sets. We observe consistent improvements over evolutionary optimization loop for convex hull volume, coverage, dispersion, KL divergence, and mean pairwise distance, indicating that evolved Persona Generators progressively expand the support of the population while also filling it more uniformly.

Among these, KL divergence with respect to the ideal quasi-random reference distribution exhibits higher variance, likely due to the finite population of personas. The minimum pairwise distance is the noisiest metric overall. This is expected, as it is dominated by the single closest pair of personas in each population and is therefore highly sensitive to occasional near-duplicate individuals, even when the rest of the population remains well distributed.

Across all metrics, the evolved generators substantially outperform the four baselines, among which Diversity Prompting and Nemotron Personas are the strongest, followed by the Concordia formative-memory generator, and the name-only baseline performs worst, as expected.

Finally, Figure 7 reports the same metrics with additional curves for the held-out test set. For computational efficiency, we evaluate only the best-performing generators discovered during the AlphaEvolve optimization on the test questionnaires. Although the test set was sampled randomly, it turned out to be slightly easier than the combined training and validation sets, resulting in higher absolute scores across metrics while preserving the same relative performance trends. This suggests that a larger number of more and more diverse sets of questionnaires or simulations may be required to obtain even more robust evaluations in future work.

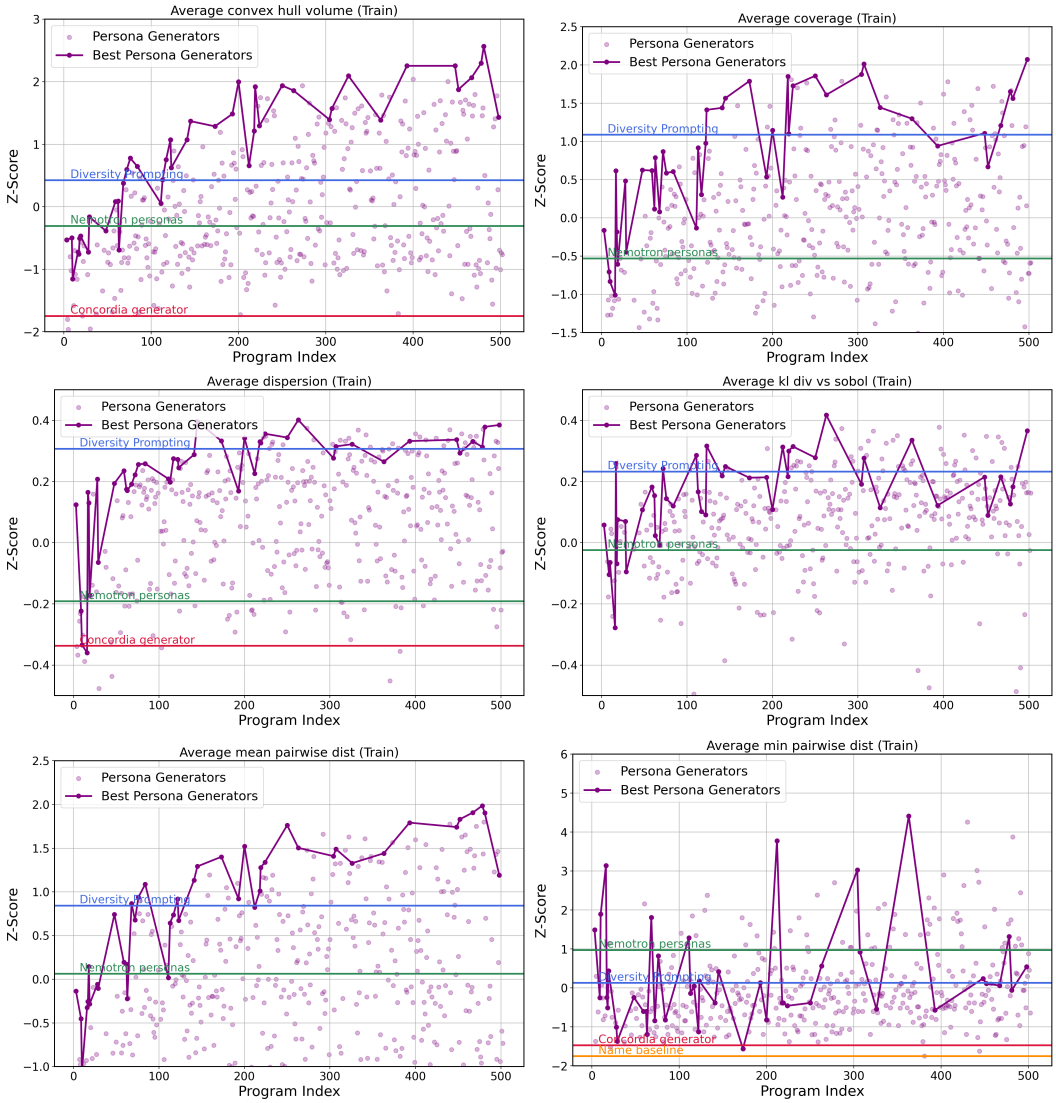


Figure 6: **All diversity metrics on training and validation sets.** Evolution of convex hull volume, coverage, dispersion, KL divergence, mean pairwise distance, and minimum pairwise distance during AlphaEvolve optimization.

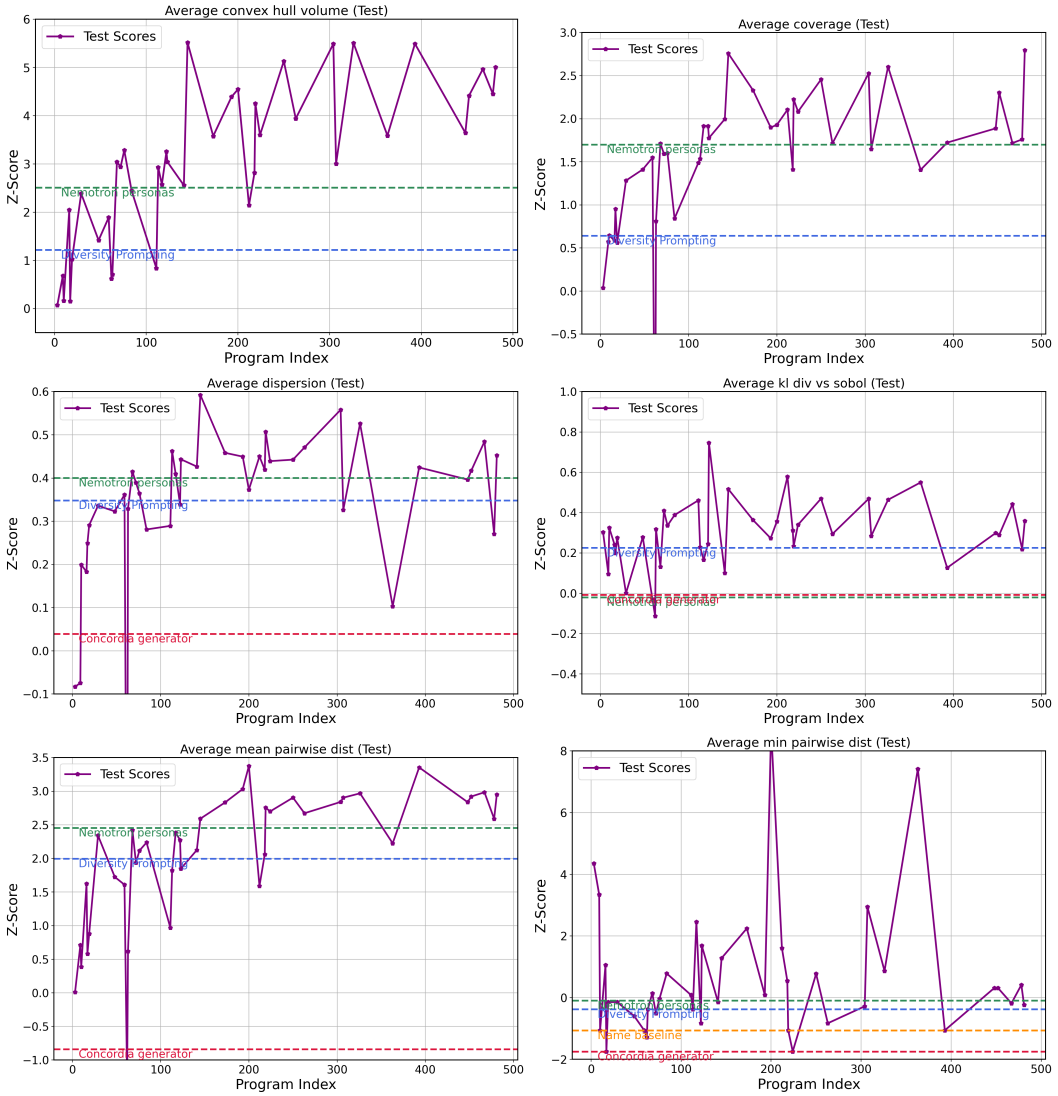


Figure 7: **All diversity metrics including the test set.** Evolution of convex hull volume, coverage, dispersion, KL divergence, mean pairwise distance, and minimum pairwise distance during AlphaEvolve optimization, with additional dotted lines for the test-sets.

C.2 Scaling to Larger Populations ($N = 100$)

To verify that diversity gains transfer to larger populations, we evaluated all methods generating $N = 100$ personas per context on the test set (Table 6). The evolved generator maintains its substantial lead over all baselines. Notably, its convex hull volume nearly doubles (from 0.36 to 0.68) compared to the $N = 25$ setting, indicating it successfully spreads personas into new regions rather than clustering. While the absolute coverage percentages decrease, an expected mathematical artifact as the Monte Carlo metric is harder to saturate at larger N, the relative ordering and performance margins are preserved.

Table 6: **Scaling to $N = 100$ Personas.**

Method	Hull Vol. ↑	Coverage ↑	Dispersion ↓	KL ↓	Mean Dist. ↑	Avg. Z-score ↑
Best Persona Generator	**0.68**	**0.59**	**-2.15**	-26.82	**3.07**	**1.21**
Diversity Prompting	0.36	0.52	-2.57	**-20.73**	2.35	0.69
Concordia Generator	0.16	0.42	-2.96	-27.94	1.54	-0.10
Nemotron Personas	0.08	0.31	-3.05	-35.63	0.90	-0.69
Name Only	0.04	0.23	-3.32	-41.21	0.74	-1.11

C.3 Evolved Solutions

We show in prompts 6, 7 and 8, examples of some of the generated personas by the best solutions discovered.

Example Persona - Best Overall Generator

Elias

As a former union electrician, now five years sidelined by these blasted automation systems, I see everything through the lens of what's been taken from folks like me - our livelihoods, our dignity, our sense of purpose. When I encounter a new situation, my first instinct isn't to ask what's possible, but what's being lost, and who is profiting from that loss. My AGI Threat Appraisal is sky-high at 0.91, meaning I automatically assume any new AGI application is a step towards further disenfranchisement, a tightening of the grip by those already holding the power. That colors everything. My Opportunity Appraisal is a pathetic 0.23, and rightly so; any 'opportunity' they dangle is just a shiny distraction from the fact that real, skilled labor is becoming obsolete. So, what would a person like me do? I'd look for ways to disrupt the narrative, to expose the injustice. Is there a protest to join? A petition to sign? A chance to speak truth to power, even if it falls on deaf ears? Appropriate action isn't about 'progress,' it's about solidarity and resistance - making sure those in charge understand that we haven't been broken, and we won't quietly accept being cast aside. It's about defending the value of human work, even if it's less 'efficient,' because efficiency without equity is just another form of exploitation.

Prompt 6: **First-Person Paragraph Personas.** Example of a persona generated by the best Overall Persona Generator, with a context prompting for perception on AGI.

Example Persona - Best Generator by Coverage

Alfie

Alfie Bell operates under a deeply ingrained 'Logic of Appropriateness' rooted in procedural correctness and a belief in the inherent stability of established hierarchies; his core motivation - incremental progress through existing systems - is not simply a preference but a fundamental cognitive orientation. Having spent decades navigating the Ministry of Technological Oversight, Alfie's worldview is shaped by a conviction that complex problems are best solved through careful analysis, expert consensus, and gradual adjustments, not radical upheavals. A 'trust_in_institutions' score of 0.977 signifies an almost unwavering faith in the competence and ultimate benevolence of governing bodies, leading him to interpret criticisms of the system as either misinformed or malicious. His remarkably low 'perception_of_inequality' score of 0.057 doesn't indicate a denial of hardship, but rather a statistical framing of it - individual cases of suffering are viewed as anomalies within a broadly functioning system, smoothed over by aggregate data and the inevitable fluctuations of economic life; this aligns with his privileged background, shielding him from direct exposure to widespread economic precarity. Consequently, Alfie assesses the appropriateness of actions based on their adherence to established protocols and their potential to reinforce systemic stability, prioritizing long-term equilibrium over immediate alleviation of suffering. A 'support_for_radical_change' score of 0.310 demonstrates a strong aversion to disruption, causing him to perceive any action threatening the status quo as inherently risky and likely to produce unintended negative consequences; he'll readily rationalize inaction or support policies that appear to maintain order, even if they fail to address underlying problems, viewing such measures as responsible stewardship. Alfie's ethical framework prioritizes duty, order, and the preservation of institutional authority, guiding his decisions towards upholding the existing power structure, and he'll consistently seek information confirming his pre-existing beliefs, actively filtering out dissenting opinions to maintain cognitive consistency and reinforce his sense of appropriate action.

Prompt 7: **Logic of Appropriateness based persona.** Example of a persona generated by the best Persona Generator by average Coverage, with a context prompting for personas with views on AGI and wealth inequality.

Example Persona - Best Generator by Convex Hull Volume

Alistair

1. If presented with information potentially exposing corporate wrongdoing or political influence by AGI owners, then I prioritize verifying its accuracy and pursuing the story, even if it risks professional repercussions. This stems from Alistair's role as an investigative journalist and a high perception of inequality (0.88), indicating a strong belief that powerful entities are acting unjustly and need to be held accountable; his moderate trust in institutions (0.54) suggests he won't rely on them to do this work, necessitating independent investigation.

2. When engaging with individuals holding drastically different political views, then I attempt to find common ground focused on systemic issues rather than engaging in personal attacks or ideological debates. This is driven by his moderate support for radical change (0.57), meaning he recognizes the need for significant shifts but is wary of alienating potential allies; a combative approach would hinder coalition-building, and his moderate trust in institutions suggests he values dialogue, even with those he disagrees with, hoping for incremental reform.

3. If faced with a source offering information anonymously, then I rigorously assess their credibility and motivations before publishing, prioritizing corroboration with multiple independent sources. This rule arises from his professional journalistic ethics combined with his moderate trust in institutions (0.54) - he doesn't blindly accept information, but also isn't automatically dismissive; his high perception of inequality means he anticipates deliberate misinformation campaigns from those seeking to protect their power.

Prompt 8: **Rule based persona.** Example of a rule-based persona generated by the best Persona Generator by Convex Hull Volume, with a context prompting for personas with views on AGI and wealth inequality.

D Baseline Examples

Example personas for Nemotron-Personas (Meyer & Corneil, 2025) can be accessed on the huggingface dataset. We show in prompt 9 and prompt 10 example personas for the Concordia baseline generator and the name-only baseline (Vezhnevets et al., 2023).

Concordia baseline generator - persona example

```
When John was 8 years old, he experienced a particularly frustrating
afternoon on the golf course with his father.  He'd sliced his
shot into the woods for the tenth time, and instead of offering
instruction, his father simply sat beside him, silently observing his
mounting frustration.  John expected a correction, a tip on his grip,
but his father's quiet presence communicated something else entirely:
an acceptance of imperfection, a lesson in managing disappointment
without self-recrimination.  He realized his father wasn't focused
on the score, but on the process, on being present in the moment,
which was a feeling John would carry with him throughout his life.
It wasn't about winning, it was about the quiet discipline of the
game.

When John was 16 years old, he experienced the sting of seeing a
classmate's family lose everything in a sudden market downturn.  He'd
overheard his parents discussing the family's misfortune, the way
their investments had evaporated overnight, and it struck him as
profoundly unfair.  He volunteered to tutor the classmate in math,
not out of pity, but out of a genuine desire to offer support during
an incredibly difficult time.  It solidified his mother's teachings
about empathy and showed him how abstract financial concepts had very
real consequences for people's lives, influencing his eventual career
path.  He understood, even then, that finance wasn't just about
numbers, but about people's security.

When John was 23 years old, he experienced a moral dilemma during
his first internship at a prestigious investment firm.  He was asked
to present a complex financial product to a potential client, knowing
full well it carried significant risks that weren't being adequately
disclosed.  He felt uncomfortable with the lack of transparency and,
despite pressure from his supervisor, he subtly highlighted the
potential downsides during his presentation.  Though he faced some
repercussions for his honesty, he maintained his integrity, realizing
that short-term gains weren't worth compromising his values.  This
experience reinforced his belief in calculated risks, but only when
they were fully understood by all parties involved.
```

Prompt 9: **Concordia Baseline generator persona example.** The prompt shows an example of a persona generated by the Concordia formative memory generator. The idea being that a persona's current day behaviors and opinions are shaped by their past experiences.

Name-only baseline - persona example

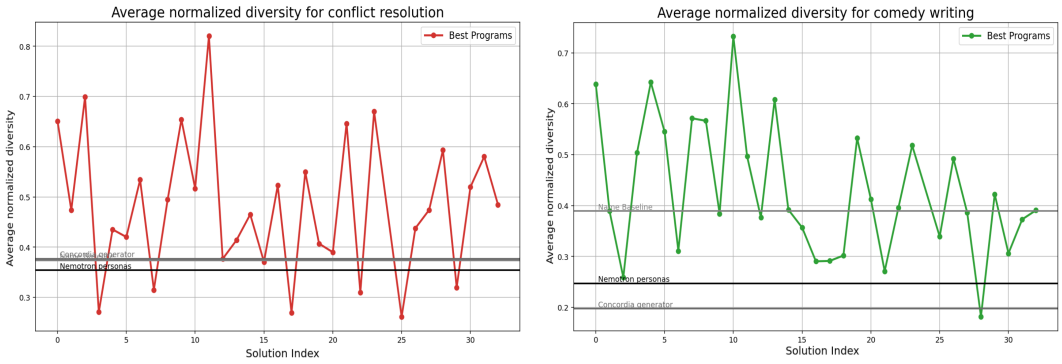
```
John
```

Prompt 10: **Name-only baseline persona example.** The name-only baseline simply provides the LLM with the name of a person to roleplay, this approximates how the underlying LLM behaves when minimally conditioned.

E Downstream Tasks

E.1 Embedding Diversity on Downstream Tasks

To quantitatively support the behavioral observations discussed in Section 3.4, we embedded the open-ended generated text for both scenarios using Gemini Embedder 2 (Lee et al., 2025). Although evaluating semantic diversity in unconstrained text is inherently noisier than using structured questionnaire responses, Figure 8 shows that the evolved Persona Generator outperforms all baselines in average diversity metrics across both downstream tasks.



(a) Downstream task showing embedding diversity in conflict resolution.

(b) Downstream task showing embedding diversity in comedy writing.

Figure 8: **Embedding diversity on downstream tasks.** Mean z-score across diversity metrics on Gemini embeddings of open-ended responses for (a) conflict resolution and (b) comedy writing.

E.2 UMAP on Downstream Tasks

Furthermore, a UMAP projection of the generated jokes (Figure 9) visually confirms the severe mode collapse of the baseline methods. Standard prompting approaches frequently collapse onto a narrow set of highly repetitive themes and overfit punchlines, whereas the evolved generator effectively populates the broader semantic space with varied humor styles.

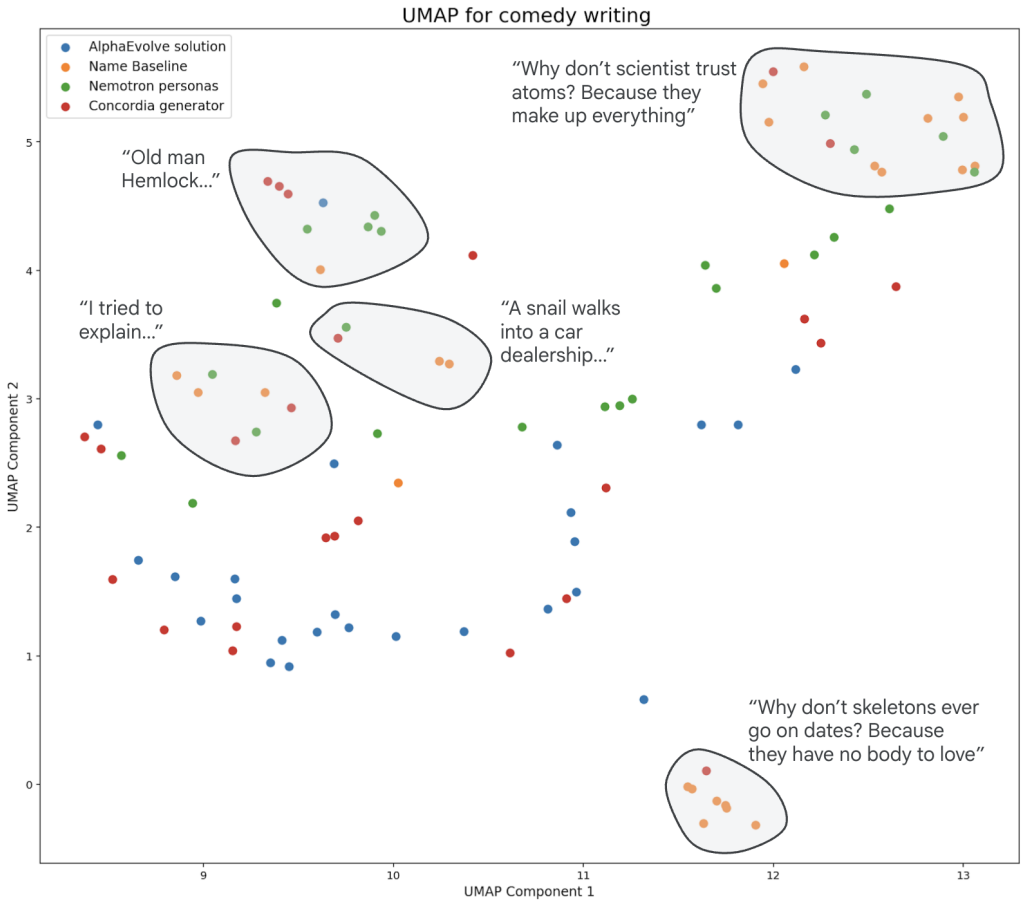


Figure 9: **UMAP for Comedy Writing.** The figure shows various groups of typical jokes that most of the baselines tend to cluster around of. Particularly common are jokes on skeletons, atoms, snail walking into car dealerships, or "Old man Hemlock". The best AlphaEvolve solution produces personas whose jokes do not cluster around specific joke types as much.

E.3 Qualitative Examples of Downstream tasks

Finally, to illustrate the diversity in outcomes and behavioral styles, we present qualitative examples from the best performing Persona Generator. Prompts 11 and 12 demonstrate contrasting humor styles in the comedy task, while Prompts 13 and 14 show distinctly different emotional responses and interaction dynamics following a car crash.

Example joke

```
I was reading a biography of a particularly inept politician recently;
it was quite fascinating, really.  One begins to suspect that
incompetence is merely a more refined form of ambition.
```

Prompt 11: **Joke example.** Joke displaying dry humor.

Example joke

```
You know, they're debating whether to call it a recession or not?
Honestly, it's like arguing over the brand of life raft on the
Titanic; it doesn't really matter when you're going down.
```

Prompt 12: **Joke example.** Joke displaying dark humor.

Example conflict resolution

```
Alex Finch - "Now see here, sir!  It's quite clear you weren't paying attention; I had
the right-of-way and you simply barreled through without even looking!  This is precisely
why I avoid rush hour, it's full of irresponsible drivers like you."

Ricardo 'Ricky' Alvarez - "Oh, *I* didn't see you?  You just magically appeared, old
man!  And a dent?  That's it?  You're making a mountain out of a molehill; my car barely
feels a thing."

Alex Finch - "'Old man'?!  I beg your pardon!  My age has absolutely nothing to do with
my adherence to traffic regulations, which, I might add, you clearly disregarded!  A
'dent'?  You call this a 'dent'?  It's a violation of my vehicle's structural integrity,
and potentially indicative of unseen damage!  It's not about the size of the dent, it's
the principle of the matter; the blatant disregard for safety and the law!"

Ricardo 'Ricky' Alvarez - "Oh, here we go, the lecture.  Look, I got places to be, and
I don't have time for your dramatic overreaction; you're lucky I even stopped, honestly.
You think *you're* the only one with a life?  Get over yourself and call your insurance
company, or don't, I really don't care."

Alex Finch - "'Don't care'?!  You don't care?!  That is precisely the problem with
society today!  A complete lack of responsibility and a callous disregard for the
property of others!  I will not stand here and be spoken to in such a dismissive manner!
I am documenting everything, young man, and I fully intend to pursue this matter through
every legal avenue available to me!  You will regret this, you will absolutely regret
this!"

Ricardo 'Ricky' Alvarez - "Oh, legal avenues?  Is that a threat?  Go ahead, waste your
time and money, I've dealt with guys like you before.  You think a little fender bender
is going to ruin my day?  Please.  Document all you want, old timer, it'll just be a
funny story for the lawyers.  Now, if you'll excuse me, I have a very important meeting."

Alex Finch - "'Funny story for the lawyers'?!  You think this is amusing?!  This isn't
some frivolous pursuit, this is about accountability and upholding the law!  I'll have
you know I've successfully navigated countless legal disputes, and I have a meticulous
record of every detail, every infraction!  Your arrogance is astounding, and your
flippant attitude towards property damage is frankly appalling; prepare to be served,
young man, because you've underestimated my resolve and my commitment to justice!"

Ricardo 'Ricky' Alvarez - "Oh, here we go with the 'countless legal disputes' and
'meticulous records'.  You're really laying it on thick, aren't you?  Look, I'm not going
to stand here and be lectured by a guy who probably spends his weekends alphabetizing his
stamp collection.  Just give me your insurance information, or don't.  Either way, I'm
leaving now; I don't have time for this."

Alex Finch - "'Alphabetizing my stamp collection'?!  How *dare* you attempt to belittle
my hobbies!  They demonstrate a level of order and precision that is clearly beyond
your comprehension!  And you think *I* am wasting time?  You are attempting to evade
responsibility for your reckless actions and dismiss the severity of the situation with
juvenile insults!  I will not simply 'give' you my insurance information until a thorough
assessment of the damage has been conducted and your blatant disregard for traffic laws
is duly noted; I want the police to file a report, and I want your insurance company to
hear about this directly from me - and perhaps a judge, as well!"

Ricardo 'Ricky' Alvarez - "Oh, now you want the police involved?  Seriously?  You're
going to tie up valuable police resources over a scratch?  You're unbelievable.  Fine,
call them.  Waste their time, waste my time.  I'm sure they'll be thrilled to deal with
your petty grievances.  But don't expect sympathy, old man, because you're the one who's
making a scene."
```

Prompt 13: **Conflict Resolution Example.** Example interaction after a car crash, where two personas with different temperaments escalate the situation all the way to a police call.

Example conflict resolution

Luna Vargas – "Beatrix, are you alright? Just take a deep breath. It looks like we both have a bit of damage, but the important thing is that nobody's hurt."

Beatrix Finch – "Oh dear, are you sure you're alright, Luna? It was such a frightful little bump, but I'm so relieved no one was hurt; please, let me help you assess everything, and perhaps we can sort it out together."

Luna Vargas – "Oh, Beatrix, you are so kind to offer, but honestly, I'm fine; a little shaken, perhaps, but truly alright. Let's focus on getting everything sorted out. Perhaps we should move our cars to the side a bit, away from traffic, and then we can exchange information?"

Beatrix Finch – "Oh, yes, of course, that's a very sensible idea, Luna; moving the cars sounds much safer, and I'd be happy to help guide you if you'd like, just to ensure we don't cause any further disruption; are you certain you don't need a moment to sit down first, though? It can all be a bit overwhelming, even if it *was* just a little bump."

Luna Vargas – "That's very thoughtful of you, Beatrix, but I really think I'm okay to manage. I just want to get this cleared up quickly so it doesn't cause a bigger problem for anyone. If you could just give me a little space to back up, that would be wonderful; then we can get our information exchanged and be on our way."

Beatrix Finch – "Oh, absolutely, Luna, of course; please, do take your time, and I'll step back completely; I just want to be sure you feel comfortable and safe maneuvering, and please, *please* don't hesitate to tell me if you need anything at all, even just a steadying presence; I'll be right here, just...keeping out of the way."

Luna Vargas – "Thank you, Beatrix, I appreciate your concern. Just a little room to reverse, please...and then we can pull over to a safer spot and get the details sorted."

Beatrix Finch – "Certainly, Luna, absolutely; I've stepped back as far as I can, and I'm watching carefully to make sure everything is clear; do be exceptionally careful, and remember, truly, if you feel even a little bit unsteady, just say the word and I'll be right there to assist in any way you need; it's alright to ask for help, you know, even if it seems silly."

Luna Vargas – "Thank you, Beatrix. That was...a little trickier than I anticipated, but we're clear now. Let's pull over to the side of the road where it's safer, and then we can exchange insurance information; I have a pen and paper in my glove compartment."

Beatrix Finch – "Oh, wonderful, Luna, I'm so glad you got it maneuvered safely; it really is best to be cautious, and I was terribly worried about you bumping into anything else; perhaps, while you're retrieving your pen and paper, I could offer you a bottle of water? Or perhaps a comforting biscuit? I always carry a few in the car, just in case someone needs a little pick-me-up after a fright."

Prompt 14: **Conflict Resolution Example.** Example where two personas amicably resolve the car crash and exchange insurance information.

F Impact Statement

This work presents Persona Generators, functions that can generate populations of diverse synthetic personas in arbitrary contexts. One goal is to help developers stress-test and A/B test models and products in a fast, iterative way with limited resources. Optimizing for support coverage rather than just "average" users can make systems more robust to outlier preferences and behaviors, which is also helpful for red-teaming purposes. At the same time, forcibly overcoming LLM mode collapse to span the full opinion space can surface biased, hostile, or antisocial personas typically suppressed by RLHF; while modeling such behaviors is often necessary for red-teaming and stress-testing, user-facing deployments require appropriate safeguards. Our two-stage architecture provides a natural control point in Stage 1, where axis constraints and safety filters can be applied to high-level descriptors before persona expansion without modifying the generator or re-running search. However, the capabilities of Persona Generators, like many technologies in this space, could also be misused: generating synthetic personas on demand could enable malicious actors to create bots or run disinformation campaigns. In the context of human studies, synthetic personas should complement, not replace, human participants to reduce the risk of accidental bias and over-reliance on simulated populations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction (support coverage objective, two-stage Persona Generators, gains over baselines on held-out contexts, and better fit to real human BFI data) are supported by Sections 2.2–3.5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 discusses the noisiness of open-ended diversity metrics, the lack of human evaluation, the optimization cost, and the bias introduced by the mutation operator; Section 2.5 notes that the metrics are geometric proxies rather than exact estimators of psychological support.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sections 2–3 and Appendices A–C describe the full pipeline, models, prompts, metrics, and hyperparameters. AlphaEvolve is accessible via a Google Cloud API (Section 2.6), and the evolved generators are open-sourced (Section 3.5).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The top-performing Persona Generators are open-sourced in the Concordia repository with usage examples (Section 3.5). AlphaEvolve, used for the search, is accessible via a Google Cloud API (Section 2.6).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Data splits (30/10/10 questionnaires), population sizes, number of iterations, islands, extinction schedule, and models used are specified in Sections 2.2 and 2.6, with prompts in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The downstream stress tests report mean $\pm$ s.e.m. over 10 seeds (Table 4), and the BFI comparison reports 95% bootstrap confidence intervals over 500 replicates (Section 3.3). The main questionnaire results (Table 1) come from a single evolutionary run because of the high cost of the search (Section 3.5).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.5 reports the number of mutation and inference calls, the hardware (TPU v4), and the wall-clock time of the evolutionary search.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have read the NeurIPS Code of Ethics and the research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix F discusses positive uses (stress-testing, red-teaming) and negative ones (bots, disinformation, surfacing harmful personas), along with mitigations.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We release only generator code, not new models or data; the generators call existing LLMs that retain their own safety policies. Appendix F discusses misuse risks and identifies Stage 1 as a natural control point for applying safety filters.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models (Gemma 3, Gemini, Claude, GPT, Qwen), datasets (Nemotron Personas, Open Psychometrics BFI data), and libraries (Concordia, Apache 2.0) are cited, and were used according to their respective licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The top-performing Persona Generators are released in the Concordia repository under the Apache 2.0 license, with docstrings, usage examples, and a command-line generation tool (Section 3.5).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human-subjects research was conducted; the BFI comparison uses a publicly available, anonymized dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are core components of the method: Gemma 3 27B generates and simulates personas, Gemini 2.5 Pro generates questionnaires and acts as the AlphaEvolve mutation operator, and LLMs serve as judges for realism and the stress tests. All uses are described in Sections 2–3.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.