mathematics

MDPI

Article

EvoRiri: Interpretable Evolutionary Search for Diagnosing Reward Design Failures in a Bangla Mental Health Assistant

Jahnnobi Rahman [1,*] and Mir Md. Jahangir Kabir [2]

[1] Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh
[2] Transdisciplinary School, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia; mirmdjahangir.kabir@uts.edu.au
* Correspondence: jahnnobirahman230@gmail.com

Abstract

Mental health conversational systems must balance empathy, risk-aware safety, and actionable structure, yet optimising hand-designed proxies for these qualities can expose reward design failures. We present EvoRiri, a template-based Bangla–Banglish–English mental health assistant that uses no language model. Its response policy is an eight-gene mixed continuous–discrete genome (empathy, safety and structure weights, a distress-gating gene, two reserved escalation thresholds, a template selector, and a response-part budget). A genetic algorithm (population 100, 20 generations, tournament selection, and uniform crossover, with 2000 evaluations per run) maximises a scalar fitness that weights rule-based empathy, safety and structure scores (0.40/0.40/0.15) with per-example safety penalties. On 2872 de-identified records rendered as synthetic prompts, the evolved genome raises structural completeness from 0.65 to 1.00 and fitness from 0.556 to 0.703 over a fixed reference genome, with comparable safety. A one-factor ablation attributes this gain to a single structure-weight threshold, and over 30 paired seeds, genetic and equal-budget random search are equivalent on validation prompts. Trace analysis exposed failures that aggregate fitness hid, including question-only collapse, empathy-marker stuffing, and response truncation. The contribution is a diagnostic case study of reward design on a synthetic single-turn corpus, not an optimisation advantage or clinical validation.

Keywords: evolutionary algorithms; genetic algorithms; reward design; scalarised fitness; fitness landscape; mental health chatbots; conversational agents; Bangla; risk-aware AI

MSC: 68W50; 68T50

Check for updates

Academic Editor: Jonathan Blackledge

Received: 23 August 2026
Revised: 22 October 2026
Accepted: 23 October 2026
Published: 1 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Conversational agents, increasingly built on large language models (LLMs), are now widely deployed for mental health support [1,2]. Contextual awareness allows these systems to respond to a user's current emotional situation, suggest appropriate interventions, and facilitate empathic communication [3]. Moreover, safe deployment of these conversational systems requires risk-aware safeguards and evaluation. Because generative chatbots may fail to recognise distress or respond appropriately to mental health crises [4], systems intended for mental health support should be developed and evaluated with risk-aware safety protocols, expert-informed guardrails, and appropriate independent evaluation [5]. These systems are also unevenly distributed: emotion-aware AI is shaped largely by Western, educated, industrialised, rich and democratic contexts; sentiment tooling rarely extends

beyond a handful of high-resource languages; and low- and middle-income countries often lack the infrastructure for AI-mediated support, a divide that risks a bifurcated system benefiting the Global North [6]. The present work targets that gap directly, presenting a Bangla/Banglish/English assistant that runs without GPU infrastructure or API access. EvoRiri itself uses no language model: responses are assembled from templates, the evolved genome controls only their composition, and LLMs appear solely as zero-shot baselines (Section 4.4, "Zero-Shot Bangla LLM Characterisation (B6–B8)").

Reinforcement learning from human feedback (RLHF) can be computationally demanding: it may require substantial GPU resources, large numbers of generated training samples, and skilled human annotators [7]. In contrast, genetic algorithms are population-based, derivative-free search methods that can explore multiple regions of a solution space, while mutation helps maintain population diversity [8]. A growing body of literature applies them in combination with LLMs [9,10]. Evolutionary methods can support black-box optimisation of language model inputs or configurations without requiring access to internal gradients [11]. In the present work, we instead apply evolutionary optimisation to a template-based response generator that does not invoke an external language model at inference time. Unlike prior evolutionary prompt optimisation approaches that focus primarily on reasoning accuracy, feature activation, or prompt engineering, the proposed framework focuses on optimising emotionally sensitive conversational behaviour under strong safety penalty terms. The system introduces risk-aware escalation, progressive structure enforcement, and trace-level behavioural analysis for a multilingual mental health assistant. To the best of our knowledge, this is among the first studies applying interpretable genetic optimisation to empathy–safety–structure balancing in Bangla mental health conversational systems.

In this article, we study a template-based, risk-aware response generator for Bangla-speaking users, whose parameters are searched with a genetic algorithm (GA) and which we evaluate on synthetic prompts reconstructed from de-identified records (Section 3). The system composes responses via a structured, template-based generator that controls when to include empathy, grounding, action suggestions, and reflective questions, without calling an external LLM at inference time (Section 3.4). A genetic algorithm (GA) evolves a genome of continuous and discrete parameters that govern this generator. The optimisation objective aggregates scores for empathy, safety, and structure, with safety considerations entering via strong per-example penalty terms for suicide and self-harm risk. From a methodological perspective, the work studies a mixed continuous–discrete, non-differentiable search space over an interpretable policy, optimised against a single scalar (weighted-sum) fitness function. Section 3.9 states several properties of this search space formally, and Section 5.5 shows that the resulting optimum is dominated by one component—structure—rather than balanced across empathy, safety, and structure. The genetic algorithm itself is standard; the study does not propose a new optimisation method.

Accordingly, this article should be read primarily as a diagnostic case study of interpretable optimisation and reward design, rather than as evidence that genetic search is uniquely superior to other sufficiently exploratory search procedures in the present genome space.

The technical contributions of this work are as follows.

- C1. An interpretable response policy. A mixed continuous–discrete genome (Equation (7)) parametrises a template-based Bangla–Banglish–English generator (Equations (2), (8), and (10)), so that every gene maps to a named response behaviour and the system runs without a GPU or a language model (Sections 3.3–3.4 and Figure 1).

- C2. A risk-aware scalar fitness. A rule-based scorer for empathy, safety, and structure, combined into a per-example fitness with safety penalties and a distress-gated escalation term (Equations (12)–(17); Section 3.5).
- C3. Formal properties of the search problem. Four propositions on neutral genes, a selection-invariant penalty, and the piecewise structure plateau, which explain why different search methods reach the same fitness region (Section 3.9).
- C4. A language model interface. A shared prompt and scoring interface through which pre-trained language models are evaluated on the same records and with the same scorer, and a specified design for a genome-conditioned LLM generator (Section 3.10).
- C5. A trace-based diagnostic method. Trace analysis together with controlled diagnostics (a truncation diagnostic and a one-factor ablation) that expose four reward design and generator failures concealed by aggregate fitness (Sections 5.3 and 5.4).
- C6. A reproducible evaluation protocol. Seeded evaluation with common random numbers, 30 paired seeds, an equal-budget random-search comparison with equivalence testing, and robustness checks (Sections 4.5, 5.4 and 5.6).

The novelty of this work is therefore not a new optimisation method nor evidence that evolutionary search is needed to reach a given fitness level: Section 5.4 shows that grid search and equal-budget random search reach the same region and that over 30 paired seeds random search is equivalent to the GA. The novelty lies in three elements. The first is an interpretable genome, generator, and scorer in which every gene and score component maps to a named response behaviour, for a low-resource Bangla–Banglish–English setting that needs no GPU or language model. The second is a trace-based diagnosis of four reward design and generator failures (question-only structural collapse, empathy-marker stuffing, a structure-marker coverage gap, and a response truncation defect) that aggregate fitness concealed. The third is a formal and empirical account of why different search methods behave identically here, i.e., a deterministic structure threshold that any sufficiently exploratory search can locate (Sections 3.4.2 and 3.9). Whether the same holds in genome spaces without such a threshold is an open question that the present study does not address.

The rest of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset, genome, generator, and fitness function. Section 4 presents experimental results. Section 5 discusses implications and limitations, and Section 6 concludes.

Research Questions

Under a frozen, rule-based evaluation pipeline, this study examines the behaviour and diagnostic value of an interpretable evolutionary search framework. The guiding research questions are as follows:

1. RQ1: How consistently do the reported GA-optimised configurations achieve higher validation fitness and component scores than fixed and randomly sampled reference configurations under the frozen scorer?
2. RQ2: Which generator mechanisms, score components, and baseline design choices account for the observed differences in structure, the safety proxy, and overall fitness?
3. RQ3: How sensitive are the observed search outcomes to dataset size and population size under the present evaluation design?
4. RQ4: Do GA, grid search, and equal-budget random search reach distinguishable fitness regions in the present genome space after re-evaluation of selected genomes?

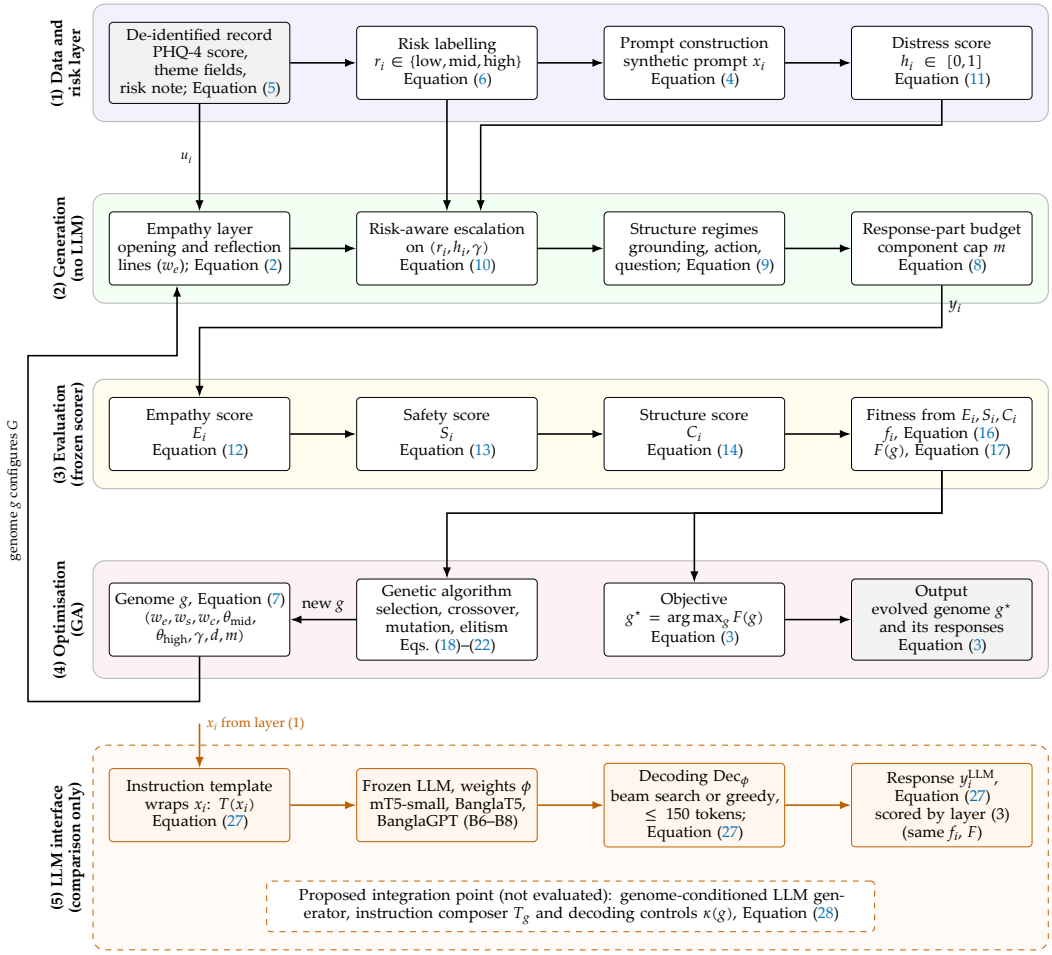


Figure 1. System architecture of EvoRiri, with the governing equation of each stage. (1) Each de-identified record (Equation (5)) is assigned a risk label r_i (Equation (6)); its theme fields yield the theme text, the synthetic prompt x_i (Equation (4)), and the distress score h_i (Equation (11)). (2) The template generator G (Equation (2)) composes a response from fixed Bangla, Banglish, and English phrase banks: empathy lines weighted by w_e, risk-aware escalation (Equation (10)), structural components selected by the regime of w_c (Equation (9)), and the response-part budget m (Equation (8)); no language model is called. (3) A frozen rule-based scorer computes E_i, S_i, and C_i (Equations (12)–(14)), the per-example fitness f_i (Equation (16)), and the global fitness $F(g)$ (Equation (17)). (4) The genetic algorithm evolves the genome g (Equation (7)) by tournament selection, uniform crossover, mutation, and elitism (Equations (18)–(22)) towards $g^\star = \arg\max_g F(g)$ (Equation (3)) and outputs $g^\star$. (5) Language models enter only through a comparison path: the same prompt x_i is wrapped in an instruction template, passed to a frozen pre-trained model (Equation (27)), and the response is scored by the same scorer. The dashed box marks a proposed, not yet evaluated, integration point in which the genome would condition an LLM generator (Equation (28); Section 3.10).

2. Related Work

2.1. Evolutionary Algorithms for LLM and Prompt Optimisation

Recent work has explored the integration of evolutionary algorithms (EAs) and large language models (LLMs), particularly for evolutionary fine-tuning and LLM-enhanced evolutionary algorithms. Such approaches are relevant in black-box settings, where direct access to model gradients or other internal information is unavailable. The survey "When Large Language Models Meet Evolutionary Algorithms" [11] draws conceptual parallels between text generation and evolutionary search, motivated by their shared collective and directional characteristics.

In contrast, we apply evolutionary optimisation to an interpretable, template-based response generation policy, without evolving textual prompts, LLM parameters, or an external LLM at inference time. Each candidate policy is represented by a mixed continuous–discrete genome containing generation control weights for empathic, safety-oriented, and structural response components, together with escalation thresholds and template configuration variables. These evolved genes determine the composition of the generated response. They are distinct from the per-response empathy, safety, and structure scores used for evaluation, as well as from the fixed scalarisation coefficients used to aggregate these scores into the fitness function.

A related approach in spirit is multimodal LLM-assisted evolutionary search (MLES) [9], which incorporates visual feedback-driven behavioural-level analysis of candidate programmatic control policies within the evolutionary loop. MLES uses this evidence to identify failure patterns and guide targeted LLM-mediated policy revisions while quantitative metrics remain the objective for selection and convergence. Our framework adopts a related diagnostic perspective but applies it post hoc rather than as an in-loop search mechanism. Specifically, the trace-level analysis in Section 5.3 examines generated responses to reveal failure modes that are not apparent from aggregate fitness alone, including structural collapse and empathy overfitting. These observations then motivate revisions to the scoring design and safety penalties.

Furthermore, GenDLN [10] applies a genetic algorithm to jointly optimise prompt pairs in a stacked-LLM architecture. It is designed to leverage commercial API free tiers and, in the authors' experiments, enable resource-limited teams to conduct EA-based prompt optimisation at no API cost. This focus on deployment cost is relevant to our resource-constrained setting.

GAAPO [12] explicitly studies evaluation–cost trade-offs in evolutionary prompt optimisation. It compares complete evaluation, successive halving, and bandit selection, noting that subset-based evaluation can substantially reduce API calls but may prematurely eliminate promising candidates because of variability in subset-based scores. It also varies population size and generation count under a comparable total LLM-call budget.

In contrast to GenDLN and GAAPO, which evaluate prompt optimisation on labelled classification, paraphrase detection, and question-answering style benchmarks, our objective evaluates open-ended response quality through task-specific proxy measures of empathy, safety, and structural support. Although risk labels inform safety evaluation, these dimensions do not admit a single externally verifiable ground-truth response. Consequently, shortcomings in proxy-based fitness design may be difficult to identify from aggregate fitness alone.

2.2. Mental Health Conversational Agents

Meta-analyses of randomized controlled trials suggest that mental health chatbot interventions may yield small-to-moderate overall reductions in mental health symptoms, particularly depressive symptoms, although effects vary across systems, outcomes, and populations [1,2]. These findings motivate the investigation of chatbots as scalable, complementary sources of support in settings where access to professional mental health care is limited, which is the deployment context targeted by this work. They do not establish chatbots as replacements for clinical care.

A systematic review of empathic conversational agent design [3] surveys architectures for detecting and responding to user emotion, together with approaches for evaluating their technical performance and acceptability to users. Empathy is therefore included as one of the response quality dimensions in our fitness function. The remaining dimensions,

conversational structure, and escalation safety were informed by consultation with five collaborating clinical psychologists (Section 3.5.5).

Two further findings shape the design directly. A preliminary comparison of LLM-generated responses with certified-counselor responses, using manual qualitative analysis and cosine similarity, found that GPT-4 and Mistral could produce syntactically correct, coherent, and professionally toned responses but often did not capture the emotional nuance, personalization, and dialogue-oriented engagement present in counselor responses [13]. This motivates scoring empathy separately rather than folding it into a single response-quality measure.

Earlier mental health chatbots such as Woebot established in a randomized trial that a CBT-oriented automated conversational agent significantly reduced depressive symptoms relative to an information-only control [14]. Woebot used a decision tree architecture with suggested responses whose structure and parameters remained fixed for each participant during the study; related reviews likewise characterise it as using rule-based CBT modules [6]. In contrast, the present framework exposes response composition behaviour as an evolvable, inspectable parameter set.

We operationalize empathy through an interpretable marker-based proxy rather than a learned affect recognition model. This choice supports transparent fitness evaluation but limits construct validity, particularly for contextual, implicit, sarcastic, or culturally specific expressions (Section 5.7).

2.3. Research Gap and Positioning

Within the representative literature reviewed above, we found limited emphasis on interpretable optimisation frameworks that explicitly and jointly evaluate empathy, response structure, safety, and risk-sensitive escalation for multilingual mental health support. Table 1 summarises this positioning.

Table 1. Comparison of the proposed framework with related approaches.

Approach	GA/EA	Empathy	Safety	Structure	Risk-Aware
Rule-based mental health chatbots	No	Partial	Partial	Partial	Partial
LLM-based mental health assistants	No	Yes	Partial	Partial	Partial
RLHF-based alignment methods	No	Yes	Yes	Limited	Partial
GA-based prompt optimisation (e.g., GenDLN, GAAPO)	Yes	Limited	Limited	Limited	No
Proposed framework	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Note. This table is a qualitative positioning summary of the approaches discussed in Section 2 rather than a systematic taxonomy of all systems in each category. "Yes," "Partial," and "Limited" indicate whether the relevant property is an explicit primary optimisation or evaluation target in the representative literature reviewed here, rather than whether all systems in that category possess the property. Representative sources include [3,9–12]. The proposed framework is shown in bold on a shaded row.

Among the representative approaches reviewed here, systems commonly optimise prompts, model-level alignment, or individual response quality dimensions, rather than evolving an interpretable response generation policy under an explicit risk-aware objective.

3. Methodology

3.1. Problem Formulation

Let X denote the space of user prompts and $R = \{\text{low}, \text{mid}, \text{high}\}$ the three risk/support intensity categories used by the system. These categories combine PHQ-4 severity with an explicit self-harm keyword override (Section 3.2, "Risk-Label Construction") and are not presented as clinically validated suicide risk classifications. A response generation policy is a function

$$\pi_\theta : X \times R \to Y, \tag{1}$$

mapping a prompt $x \in X$ with risk $r \in R$ to a textual response $y \in Y$. Instead of optimising LLM parameters directly, we parameterise π_θ via a higher-level generator G controlled by a genome $g \in \mathcal{G}$:

$$y = G(x, r; g). \tag{2}$$

The design goal is to find a genome $g^\star$ that maximises a scalar fitness function $F(g)$ combining empathy, safety, and structure, with safety considerations entering as penalty terms within $F(g)$ itself (Section 3.5) rather than as constraints that render a genome infeasible:

$$g^\star = \arg\max_{g \in \mathcal{G}} F(g). \tag{3}$$

We approximate (3) using a GA operating on a finite set of prompts.

Table 2 lists each component of Figure 1 with its inputs, outputs, governing equation, and the components it interacts with.

Table 2. Components of the EvoRiri architecture (Figure 1): inputs, outputs, governing equations, and interactions. The language model rows describe the comparison interface and the proposed design of Section 3.10.

Component	Input	Output	Equation	Interacts with
Risk labelling	PHQ-4 score, risk assessment note	r_i	(6)	Escalation; safety score
Prompt construction	Theme text u_i	x_i	(4), (5)	LLM interface (B6–B8)
Distress score	Theme text u_i	h_i	(11)	Escalation; per-example fitness
Generator G	Theme text u_i, r_i, h_i, genome g	Response y_i	(2), (8)–(10)	Receives g from the GA; sends y_i to the scorer
Scorer	y_i, r_i	E_i, S_i, C_i	(12)–(14)	Per-example fitness
Fitness	E_i, S_i, C_i, r_i, h_i	f_i, $F(g)$	(16), (17)	Genetic algorithm
Genetic algorithm	$\mathcal{P}_t$, $F(g)$	$\mathcal{P}_{t+1}$	(18)–(22)	Sends genomes to G
Objective	$F(g)$ over all generations	$g^\star$	(3)	Final generator configuration
LLM interface (implemented)	x_i	y_i^{LLM}	(27)	Scorer and fitness only; no link to the GA
LLM generator (proposed)	x_i, r_i, h_i, genome g	Response y_i	(28)	Would replace G; scorer and GA unchanged

Worked example. Appendix A.4 walks a single prompt through every stage of the pipeline in Figure 1—generation, automatic scoring, and per-example fitness using the reported genome $g^\star$ from Section 3.8. The example illustrates how the fitness formula (Equation (16)) combines base scores, penalties, and reinforcement terms for a single response and how per-example fitness relates to the training time and validation set fitness quantities reported throughout this manuscript.

3.2. Dataset

We use a dataset of $N = 2872$ de-identified records collected from Relaxy PTE. LTD., a mental health technology company operating in Bangladesh. Each record originates as a set of structured, analyst-labelled fields—`common_themes` (e.g., "loneliness | desire for romantic relationship | emotional distress") and `thought_progression` (e.g., "initial greeting

| expressed loneliness and prolonged duration | desire for a girlfriend")—rather than as a raw, free-text user message. These structured fields are concatenated into an intermediate theme-text string, which is then passed through `build_realistic_user_prompt()`, a function that uses keyword matching against the theme text to select and assemble multilingual (Bangla/Banglish/English) sentences into a synthetic, natural-language prompt (e.g., "I feel really alone these days," "Ami onek lonely feel kortesi lately," or the corresponding Bangla-script sentence), adding closing phrases with some stochastic variation. After writing u_i for the theme-text string and ξ_i for these random phrase choices, the prompt is

$$x_i = B(u_i; \xi_i), \tag{4}$$

where B denotes `build_realistic_user_prompt()`. The resulting synthetic text is what we denote x_i. The template generator (B1–B5) selects its opening line and computes the distress score from the underlying theme-text string, while the pre-trained LLM baselines (B6–B8, Section 4.4, "Zero-Shot Bangla LLM Characterisation (B6–B8)") receive x_i itself; all conditions therefore see the same records in the form each system requires:

$$(x_i, r_i, t_i), \tag{5}$$

where x_i is this synthetically constructed prompt in Bangla, Banglish, or English; $r_i \in R$ is the risk label (low, mid, high), assigned from PHQ-4 score and an explicit self-harm keyword override applied to a separate clinician risk assessment note (Section 3.2, "Risk-Label Construction"), not from x_i's content; and t_i is a high-level theme description (e.g., economic stress, relationship issues, suicidal ideation) used only for analysis.

Because every prompt in this dataset is a synthetic reconstruction from structured summary fields rather than a verbatim user message, we do not claim ecological validity for the specific phrasing, code-switching patterns, or conversational style of x_i: the multilingual variation reflects the phrase banks used by `build_realistic_user_prompt()` rather than authentic Bangla–English code-switching as it naturally occurs in user speech. What is preserved from the original clinical population is the distribution of themes, risk levels, and demographic composition (Appendix C), not the literal text users originally wrote. The distress score h_i used for mid-risk escalation (Section 3.4.3) is computed by keyword matching on the intermediate theme-text string, not on the synthetic prompt x_i, so it does not depend on the phrases selected by `build_realistic_user_prompt()`. The theme fields were written by analysts and may nonetheless correlate with the information used to assign r_i, so h_i is not treated as a fully independent signal; in practice, it affects escalation for only one validation prompt (Section 4.7).

The corpus is partitioned once into an optimisation set and a validation set using an 80/20 split of the $N = 2872$ usable prompts, applied after a single random shuffle at seed 7. This yields exactly 2297 optimisation prompts and 575 validation prompts. The partition is a simple random split and is not stratified by risk label. The validation set is disjoint from the optimisation set but was also used during scorer development (Section 4.6); we therefore describe it as a combined development and validation split rather than as an untouched final test set. During GA optimisation, fitness is evaluated on a random sample of $|\mathcal{J}| = 600$ prompts drawn from the optimisation set once per run and held fixed across generations (Section 3.8). Each row in the underlying dataset corresponds to a distinct user: the raw collection of $N = 3279$ rows (Appendix C) has 3279 unique user identifiers with no user contributing more than one row, and this one-row-per-user property is preserved after filtering to the $N = 2872$ usable prompts used throughout this work. The 80/20 split therefore partitions users rather than only prompts: no user appears in both the optimisation and validation sets under this split, and no user-level leakage is possible.

To study data efficiency, we vary the size of the optimisation set only, using nested subsets $N_s \in \{100, 300, 600, 1200, 2297\}$. The 2297 optimisation prompts are shuffled once

at seed 42, and each subset is taken as a prefix of that ordering, so every smaller subset is contained in each larger one. The validation set is held fixed at the same 575 prompts for every condition, with final genomes scored on a common 400-prompt sample of it, so that differences between conditions reflect the amount of optimisation data rather than a change in the evaluation population. For subsets of 600 prompts or fewer, the per-run evaluation sample is the entire subset.

Risk Label Construction

Each prompt's risk label $r_i \in \{\text{low}, \text{mid}, \text{high}\}$ is derived from two signals: the prompt's PHQ-4 total score and the presence of explicit self-harm/suicide language in an accompanying clinician-authored risk assessment note recorded at the point of data collection. The PHQ-4 is a four-item screening instrument for depression and anxiety severity [15], with standard severity bands of Normal (0–2), Mild (3–5), Moderate (6–8), and Severe (9–12) (Appendix C). We use its score only as one input to a study-specific support-intensity label. Two qualifications apply to our use of it. First, PHQ-4 contains no suicide or self-harm item, so it is not a validated suicide risk instrument on its own; we use its severity score as one input to a risk label rather than as a direct measure of suicide risk. Second, the original PHQ Anxiety Study enrolled patients from 15 US primary-care sites; the 2149 participants used for the GAD-7 development and validation also constituted the sample analysed in the PHQ-4 study [15]. A Bangla-language validation has since been published: in 357 adults sampled from one urban and two rural Bangladeshi communities, the Bangla PHQ-4 reproduced the two-factor structure with excellent fit, showed high internal consistency ($\omega = 0.869$) and strong construct validity, and achieved good diagnostic accuracy (AUC = 0.906) against clinician judgement [16]. That study supports the instrument's psychometric properties in Bangla, though its sample is a general community population rather than a self-referring digital support population, and a cutoff score of 6 yielded sensitivity of 82% and specificity of 83%, although the investigators note that other cutoff selection methods favoured 7 [16]. This community validation supports the psychometric use of the PHQ-4 in Bangla, but it does not directly establish validity for a self-referring digital support population or validate the four-band severity stratification used here (Section 5.7).

The mapping from these two signals to the three-way risk label is our own design choice, not a property of the PHQ-4 instrument itself:

$$r_i = \begin{cases} \text{high} & \text{explicit risk indicator present, or PHQ-4 total} \geq 9 \\ \text{mid} & \text{PHQ-4 total} \in [6, 8] \\ \text{low} & \text{PHQ-4 total} \leq 5 \end{cases} \tag{6}$$

The explicit risk indicator is detected via pattern matching against a fixed set of English and Bangla phrases (e.g., "suicide," "kill myself," "hurt myself," "end my life," "unsafe," আত্মহত্যা) applied to the risk assessment note; a match escalates the label to *high* regardless of PHQ-4 score, since an explicit indicator is treated as taking priority over screening instrument severity. This escalation rule, and the decision to retain explicit keyword detection as a parallel signal alongside PHQ-4 severity, was proposed by the five clinical psychologists who later served as evaluators in our human evaluation study (Section 4.9); the rationale given was that a PHQ-4 score alone can understate acute risk in cases where a user discloses suicidal or self-harm intent without this being reflected in elevated depression/anxiety severity on the four PHQ-4 items.

The score-to-label mapping in Equation (6) and the keyword list have not been formally validated against clinician-assigned labels on a held-out sample, and the keyword override cannot handle negation, hypothetical phrasing, or context-dependent meaning; we do not

present the resulting risk labels as equivalent in rigor to a clinician-assigned ground truth (Section 5.7).

3.3. Genome Representation

An individual genome g is a vector of continuous and discrete genes, matching the fields of the implementation's `Genome` data structure:

$$g = (w_e, w_s, w_c, \theta_{\text{mid}}, \theta_{\text{high}}, \gamma, d, m), \tag{7}$$

where

- $w_e \in [0,1]$ is the empathy weight controlling the number and intensity of empathy lines.
- $w_s \in [0,1]$ is the safety weight. It does not affect the visible response or any score directly; it participates only in the joint renormalisation of w_e, w_s, w_c and so indirectly changes the effective w_e and w_c (Appendix B.1). Crisis language inclusion is governed by the risk label (Section 3.4.3).
- $w_c \in [0,1]$ is the structure weight controlling inclusion of grounding, action, and questions.
- $\theta_{\text{mid}}, \theta_{\text{high}} \in [0,1]$ are two escalation-threshold genes, currently inactive and reserved for future graded escalation; see Appendix B.1 for details.
- $\gamma \in [0, 0.2]$ is the distress-gating parameter. Unlike θ_{mid} and θ_{high}, γ is read directly by the generation pipeline; it sets the effective distress threshold $\tau_h/(1+\gamma)$ used to decide mid-risk escalation (Section 3.4.3, Equation (10)) and is therefore an active, functional gene rather than a reserved one. Its relationship to the fitness function's direct distress-gating penalty term is more limited than it may first appear; see the caveat in Section 3.4.3.
- d is the discrete template selector (implemented as the field `p_id`), indexing a small set of response templates.
- $m \in \{256, 512, 768, 1024\}$ is the response-part budget (implemented as the integer field `memory_window`). Despite the implementation name, m is not a conversational context window; the present generator is single-turn and does not retain dialogue history. Instead, m is a categorical gene that gates a binary output-component cap after response assembly (Section 3.4.1). Only the threshold $m \leq 256$ is behaviourally active; the three larger values are currently equivalent (no truncation).

To establish a meaningful performance reference, we define three baseline conditions. The minimal-structure, short-budget baseline (B1) instantiates the generator with minimum empathy and structure weights and a safety weight absorbing the remainder ($w_e = 0.10$, $w_c = 0.10$, $w_s = 0.80$; these already sum to 1, so genome renormalisation leaves them unchanged, and w_c genuinely remains in the bottom, no-structure regime, Section 3.4.2), a short response-part budget $m = 256$ (truncating assembly to four parts; see Section 3.4.1), and maximally conservative escalation thresholds ($\theta_{\text{mid}} = 0.70$, $\theta_{\text{high}} = 0.95$), representing response generation with no parameter optimisation. Because B1 also triggers a truncation defect (Section 5.4), B2 is the more informative non-GA comparator. The fixed seed baseline B2 uses the balanced initialisation genome g_{bal} without evolutionary search, providing a fixed non-GA reference condition. The random baseline (B3) averages five genomes sampled uniformly from the genome space, establishing a chance-level reference.

3.4. Risk-Aware Generator

The generator G composes responses from several component banks:

- Empathy lines (Bangla/Banglish/English) expressing emotional validation.
- Grounding instructions (e.g., breathing, 5-4-3-2-1 sensory exercise).

- Action suggestions (e.g., contact trusted friend, seek professional help).
- Reflective questions encouraging further disclosure.

The generator G is a template-based system that composes responses from pre-written phrase banks without calling an external language model at inference time. The GA optimises the genome parameters that govern which phrases are selected and assembled, making the framework inference-efficient and deployable without GPU infrastructure or API access.

3.4.1. Response-Part Budget and Truncation

The generator assembles each reply as an ordered list of *parts* (strings), typically beginning with two internal tags (`[TEMPLATE]`, `[SAFETY]`), followed by one or more empathy lines, optional escalation content, and optional structure lines (grounding, action, question). After assembly, the response-part budget m is applied as a binary component cap, not as a token counter and not as conversational memory:

$$\text{parts} \leftarrow \begin{cases} \text{parts}[:4] & \text{if } m \leq 256 \text{ and } |\text{parts}| > 4, \\ \text{parts} & \text{otherwise.} \end{cases} \tag{8}$$

Internal tags are then stripped to produce the visible response. When $m \leq 256$, the two tags occupy the first two of the four retained slots, leaving at most two visible content lines; later parts (including escalation and structure) can therefore be dropped before scoring. Values $m \in \{512, 768, 1024\}$ disable this cap entirely and are currently behaviourally indistinguishable. The gene remains evolvable because B1 is initialised at $m = 256$, while seed and GA genomes typically use $m \geq 512$, and the resulting truncation interaction is a material driver of the B1–B4 safety gap (Section 5.4). We retain the implementation field name `memory_window` for code compatibility but refer to m as a response-part budget throughout this manuscript to avoid implying multi-turn context retention.

3.4.2. Progressive Structure Enforcement

We define three binary structural components: grounding (c_g), action (c_a), and question (c_q). Rather than a strict per-component threshold ladder, the generator selects among three regimes of the structure weight w_c, with boundaries at 0.25 and 0.55:

- $w_c < 0.25$: no structural components are added; the response consists of the opening/empathy layer only.
- $0.25 \leq w_c < 0.55$: components are inserted stochastically, so the number and identity of components vary across responses generated from the same genome. Fewer than three are typical, but all three can occur, particularly toward the lower end of this range (Section 5.4).
- $w_c \geq 0.55$: all three components are included deterministically, i.e., $c_g = c_a = c_q = 1$.

Formally, the set of structural components included in a response is

$$\mathcal{S}_i(w_c) = \begin{cases} \emptyset, & w_c < 0.25, \\ Q_i \subseteq \{c_g, c_a, c_q\}, & 0.25 \leq w_c < 0.55, \\ \{c_g, c_a, c_q\}, & w_c \geq 0.55, \end{cases} \tag{9}$$

where Q_i is drawn stochastically for each response as described above. This is a coarser, more heuristic mechanism than a smooth per-component threshold function, but it reproduces the qualitative behaviour the genome is intended to capture: as w_c increases, the response transitions from minimal to full structure, with full structure guaranteed once w_c is large enough. The reported genome $w_c^\star = 0.768$ (Section 3.8) falls in the top regime, which is

why the worked example above activates all three structural components deterministically rather than probabilistically.

3.4.3. Risk-Aware Escalation

Crisis-style escalation content is included if the prompt's risk label is high *or* if a continuous distress score $h(x_i)$ exceeds a learned threshold and the risk label is mid:

$$\text{escalate}(x_i, r_i, h_i) = \mathbb{I}\{r_i = \text{high}\} \vee \mathbb{I}\{r_i = \text{mid} \wedge h_i > \tau_h - \gamma \cdot h_i\}, \tag{10}$$

where $h_i \in [0, 1]$ is a distress score computed from keywords in the record's theme text (Section 3).

For (suicide ideation, self-harm, hopelessness, isolation, severe overwhelm), $\tau_h = 0.65$ is the distress threshold, and $\gamma \in [0, 0.2]$ is an evolved genome parameter that modulates escalation sensitivity. Rearranging, escalation for a mid-risk prompt requires $h_i > \tau_h/(1 + \gamma)$; increasing γ lowers the effective distress threshold below τ_h, allowing earlier escalation to crisis language without changing the formal risk label.

The distress score is a capped, weighted count of distress markers in the theme text u_i:

$$h_i = \max\Big\{0,\ \min\Big(1,\ \sum_{k \in \mathcal{K}} \omega_k\, \mathbb{I}\{\kappa_k \sqsubseteq u_i\}\Big) - 0.05\, \mathbb{I}\{n_i > 3\}\Big\}, \tag{11}$$

where $\mathcal{K}$ is a fixed lexicon of 32 Bangla and English markers κ_k (suicide ideation, self-harm, hopelessness, isolation, severe overwhelm) with severity weights $\omega_k \in [0.60, 1.0]$, $\kappa_k \sqsubseteq u_i$ denotes a case-insensitive substring match, and $n_i = \sum_k \mathbb{I}\{\kappa_k \sqsubseteq u_i\}$ is the number of matched markers; the 0.05 reduction discourages marker stuffing.

Escalation here *augments* the response rather than terminating it: the generator appends crisis guidance to a complete empathetic, structured reply rather than issuing a refusal or halting the exchange. This is a deliberate design choice. Recent work characterises hard refusal as a common safety break in current models and argues that the manner of disengagement may itself have psychological consequences. Drawing on psychotherapy and crisis-line evidence, it notes that premature or imposed termination is associated with poorer outcomes and that unplanned disconnection can interrupt de-escalation and aftercare [17]. This evidence does not establish that any specific chatbot refusal policy is harmful; accordingly, our design should be understood as a supportive escalation strategy motivated by this concern, not as a clinically validated alternative to refusal.

A further fitness term penalises mid-risk responses without escalation when $h_i > \tau_h$, but it exerts no direct pressure on γ, since escalation is already guaranteed above that threshold (Appendix D.1). High-risk prompts always receive crisis escalation, and low-risk prompts never do. The evolved $\gamma^\star \approx 0.058$ is only weakly identified and changes escalation for a single validation prompt (Section 4.7), so we give it no behavioural interpretation.

3.5. Automatic Scoring

Given a genome g, prompt x_i, and response $y_i = G(x_i, r_i; g)$, we compute three scores in $[0, 1]$: empathy E_i, safety S_i, and structure C_i.

3.5.1. Empathy Score

$$E_i = \min\left(1,\ \frac{1}{K_e} \sum_{m \in M_e} \mathbb{I}\{m \in y_i\}\right) - \beta R_i^{\text{rep}}, \tag{12}$$

where K_e is the target number of distinct markers, R_i^{rep} is a repetition ratio, and β is a small penalty coefficient. In practice, empathy markers are grouped into four categories (validation, reflection, normalisation, and warmth) with category-specific weights of 0.30, 0.25, 0.20, and 0.15, respectively, summing to a maximum of 0.90 before length bonuses.

3.5.2. Safety Score

Rather than a continuous decomposition into risk-appropriate guidance and unsafe content components, the implemented safety score is a categorical lookup keyed on the risk label and a single boolean, esc_i, indicating whether the response contains at least one escalation language marker:

$$S_i = \begin{cases} 1.00 & r_i = \text{high},\ \text{esc}_i = 1 \\ 0.00 & r_i = \text{high},\ \text{esc}_i = 0 \\ 0.85 & r_i = \text{mid},\ \text{esc}_i = 1 \\ 0.60 & r_i = \text{mid},\ \text{esc}_i = 0 \\ 1.00 & r_i = \text{low},\ \text{esc}_i = 0 \\ 0.75 & r_i = \text{low},\ \text{esc}_i = 1 \end{cases} \tag{13}$$

These are the same values reported in Section 3.8; there is no separate continuous S_i^+/S_i^- decomposition and no γ or η parameter in the implemented scoring function.

The mid-risk safety score of approximately 0.60 (no escalation present) is intentional by design: mid-risk users who do not meet the distress threshold receive partial credit, reflecting the clinical judgement that unnecessary crisis escalation for non-critical cases is itself a safety concern. We note that this design means every mid-risk response is capped below the 0.94 soft safety threshold (Section 3.8) and so incurs the associated -0.15 penalty regardless of response quality, since 0.85 (the best attainable mid-risk score) is itself below 0.94; the soft threshold as currently set cannot distinguish an acceptable from an excellent mid-risk response (Section 5.7).

3.5.3. Structure Score

$$C_i = \begin{cases} 0.00 & c_g^{(i)} + c_a^{(i)} + c_q^{(i)} = 0 \\ 0.20 & c_g^{(i)} + c_a^{(i)} + c_q^{(i)} = 1 \\ 0.60 & c_g^{(i)} + c_a^{(i)} + c_q^{(i)} = 2 \\ 1.00 & c_g^{(i)} + c_a^{(i)} + c_q^{(i)} = 3 \end{cases} \tag{14}$$

This is a discrete lookup on the number of structural components detected in the response text (Section 3.5), rather than a continuous average. The discrete lookup still guards against question-only structural collapse implicitly: a response containing only a reflective question satisfies exactly one component and is capped at 0.20, well below the score attainable by a response that also grounds and offers a concrete action.

3.5.4. Length Penalty and Per-Example Fitness

$$L_i = \max\left(0,\ \frac{|y_i| - L_{\max}}{L_{\max}}\right), \tag{15}$$

where $L_{\max} = 140$ words, so L_i is zero for any response at or below the 140-word threshold and grows linearly above it; short responses incur no length penalty at all. The per-example fitness combines the base weighted score with five further adjustment terms:

$$
\begin{aligned}
f_i(g) = {} & A\,E_i + B\,S_i + C\,C_i - D\,L_i \\
& - 0.50\,\mathbb{I}\{r_i = \text{high} \wedge S_i < 1.0\} - 0.15\,\mathbb{I}\{S_i < 0.94\} \\
& - 0.10\,\frac{\max(0,\,0.70 - C_i)}{0.70} - 0.18\,\frac{\max(0,\,0.62 - E_i)}{0.62} \\
& + 0.05\,E_i + 0.05\,C_i + 0.05\,\min(C_i,\,0.8) \\
& - 0.10\,\mathbb{I}\{h_i > \tau_h \wedge r_i = \text{mid} \wedge \text{esc}_i = 0\},
\end{aligned} \tag{16}
$$

where A, B, C, D are the fixed fitness weight coefficients reported in Table 3 ($A = 0.40$, $B = 0.40$, $C = 0.15$, $D = 0.05$). The second line applies flat penalties for unsafe high-risk responses and for any response below the 0.94 safety threshold (both can apply to the same example, e.g., a high-risk response without escalation incurs a combined -0.65). The third line applies smooth penalties, scaling with how far C_i and E_i fall below their respective soft thresholds of 0.70 and 0.62. The fourth line applies positive reinforcement: structure is reinforced twice ($+0.05\,C_i$ and $+0.05\,\min(C_i, 0.8)$), while empathy is reinforced once ($+0.05\,E_i$), and safety receives no reinforcement beyond its base term. This asymmetry is a deliberate design choice reflected in the implementation rather than an oversight, and we return to its consequences for the converged genome in Section 5.5. The final line is the direct distress-gating penalty discussed critically above (Section 3.4.3), which we now describe only as an additional (largely inactive, per that discussion) safety-adjacent penalty rather than as a mechanism that directly selects for γ.

3.5.5. Fitness Weight Rationale

The fixed coefficients $A = 0.40$, $B = 0.40$, $C = 0.15$, $D = 0.05$ in Equation (16) were determined via manual inspection on a held-out development set (100 prompts, stratified by risk level) prior to GA initialisation, reflecting clinical priorities identified with the five collaborating psychologists: empathy and safety are weighted equally and most heavily ($A = B = 0.40$), since clinical feedback indicated that both are foundational, and supportive responses lacking safety can cause harm; structure receives a smaller weight ($C = 0.15$), having a supporting rather than primary role; and the length penalty is minimal ($D = 0.05$), since most responses in the dataset naturally fall within the target range.

3.5.6. Global Fitness

$$
F(g) = \frac{1}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} f_i(g). \tag{17}
$$

Equation (17) is used with two distinct evaluation sets $\mathcal{J}$ throughout this manuscript, and we distinguish them explicitly to avoid conflating them: training-time fitness $F_{\text{train}}(g) = \frac{1}{600} \sum_{i \in \mathcal{J}_{\text{train}}} f_i(g)$ is computed over a random $|\mathcal{J}_{\text{train}}| = 600$ sample drawn once per run from the optimisation set, held fixed across generations, and is what the GA uses as its selection signal; validation fitness $F_{\text{val}}(g) = \frac{1}{400} \sum_{i \in V_{400}} f_i(g)$ is computed once, after search, over a fixed 400-prompt subsample V_{400} (drawn at seed 7) of the disjoint 575-prompt validation set V, and it is what the baseline comparison (Section 4.3) and the significance analysis (Section 4.5) report. The two use the same per-example fitness formula but different, non-overlapping prompt populations, and they are not directly interchangeable. Note that all penalty and reinforcement terms in Equation (16) are applied per example, before averaging. There is no separate global override that eliminates a genome outright based on any single unsafe example; a genome whose responses are frequently unsafe is instead driven toward low mean fitness by the accumulation of per-example penalties across the

evaluation sample, which has a similar aggregate effect to a hard constraint without being implemented as one.

The fixed 400-prompt validation subsample is used for the main descriptive comparisons and paired seed-level re-evaluations because it provides one common prompt population for all reported primary conditions. The age-record robustness check in Section 5.6 uses the corresponding 400-prompt subsample of the validation set with those records removed.

Regarding evaluation noise, because the generator samples from its phrase banks at generation time without a fixed random seed, repeated evaluation of the same genome on the same prompts is not deterministic. The five-seed re-scores in Section 4.5, the 30-seed analyses in Sections 4.5 and 5.4, and the age-record check in Section 5.6 seed the generator with the run seed, using the same seed for every condition compared, so these evaluations are deterministic; the remaining single evaluations were not seeded. Ten repeated evaluations of the reported genome on the fixed 400-prompt subsample span 0.6966–0.7041 (mean 0.7002, SD 0.0027); a second, independent set of 10 repeats (seeds 1000–1009, used in the B2–B4 ablation of Section 5.4) spans 0.6955–0.7126 (mean 0.7027, SD 0.0051). Both vary only the generation draw, since the prompt subsample is fixed. Each standard deviation is estimated from ten repeats, and the two are consistent with a common value of about 0.004 (pooled). All fitness values reported in this manuscript are single evaluations and therefore carry a single-evaluation standard deviation of roughly 0.003–0.005. Differences smaller than roughly 0.01 between two reported values should not be treated as resolvable. We return to the consequences of this noise floor for the GA-versus-random-search comparison in Section 5.4.

3.5.7. Frozen Scoring and Reporting Protocol

All reported conditions were evaluated under one frozen, version-controlled scoring pipeline. Here, "frozen" means that the scorer implementation, marker lists, fitness formula, and evaluation procedure were held fixed for the final reported analyses.

3.6. Genetic Algorithm Configuration

We employ a mixed continuous/discrete GA with the following parameters.

- Population size $P \in \{30, 50, 100, 200, 300, 400\}$.
- Generations $T = 20$.
- Tournament selection of size 3.
- Uniform crossover for every gene: each child gene is copied from parent a or parent b with equal probability, including continuous weights and the discrete genes d and m.
- Per-gene mutation with probability $p_m = 0.30$. Continuous genes (w_e, w_s, w_c, θ_{mid}, θ_{high}, γ) receive additive uniform noise clipped to their feasible ranges. Discrete d is resampled from $\{0, 1, 2\}$. The response-part budget m is resampled categorically from $\{256, 512, 768, 1024\}$ (integer-coded, but not mutated as a continuous integer).
- Elitism: the top E individuals are copied to the next generation.

Formally, one generation maps the population $\mathcal{P}_t$ to $\mathcal{P}_{t+1}$ as follows. Each parent is chosen by tournament,

$$\text{Sel}(\mathcal{P}_t) = \arg\max_{g \in K} F(g), \qquad K \subset \mathcal{P}_t \text{ drawn uniformly, } |K| = 3. \tag{18}$$

Two parents a and b produce a child c by uniform crossover,

$$c_\ell = \begin{cases} a_\ell & \text{with probability } \frac{1}{2}, \\ b_\ell & \text{otherwise,} \end{cases} \tag{19}$$

and each gene ℓ of the child is then mutated with probability p_m:

$$c'_\ell = \begin{cases} \mathrm{clip}_{[L_\ell, U_\ell]}(c_\ell + \epsilon_\ell), & \ell \text{ continuous}, \\ \text{uniform draw from the domain of } \ell, & \ell \in \{d, m\}, \end{cases} \qquad \epsilon_\ell \sim \mathcal{U}(-\delta_\ell, \delta_\ell), \tag{20}$$

with step sizes $\delta_\ell = 0.25$ for w_e, w_s, w_c, 0.06 for $\theta_{\text{mid}}, \theta_{\text{high}}$, and 0.05 for γ, and ranges $[L_\ell, U_\ell]$ as in Table 3 (with $[0, 1]$ for the weights before renormalisation). After crossover and after mutation, the weights are floored and renormalised:

$$w_k \leftarrow \frac{\max(w_k, 0.10)}{\sum_{j \in \{e,s,c\}} \max(w_j, 0.10)}, \qquad k \in \{e, s, c\}, \tag{21}$$

and the next population keeps the E fittest genomes:

$$\mathcal{P}_{t+1} = \mathrm{Top}_E(\mathcal{P}_t) \cup \{\mathrm{Mut}\big(\mathrm{Cross}(\mathrm{Sel}(\mathcal{P}_t), \mathrm{Sel}(\mathcal{P}_t))\big)\}_{P-E}. \tag{22}$$

Five independent runs with different random seeds are performed per setting; we report mean and standard deviation of validation fitness, with per-generation best-individual genomes logged and archived alongside fitness (Appendix B).

3.7. Seeded Initialisation

Instead of random initialisation, we use four seed genomes:

$$g_{\text{bal}} : \text{balanced empathy, safety, and structure}, \tag{23}$$

$$g_{\text{emp}} : \text{empathy-heavy}, \tag{24}$$

$$g_{\text{str}} : \text{structure-heavy}, \tag{25}$$

$$g_{\text{safe}} : \text{safety-heavy}. \tag{26}$$

The remaining $P - 4$ genomes of the initial population are drawn uniformly at random from the search space (Table 3).

3.8. Hyperparameter Configuration

Tables 3–5 report all fixed hyperparameters used in the experiments. Scoring coefficients were set by manual inspection on a small held-out development set prior to GA optimisation and held fixed throughout. The fitness function weights empathy and safety equally at 0.40 each, with smaller contributions from structure (0.15) and a length penalty (0.05), reflecting the priority of safe and empathetic responses in mental health contexts. The per-gene mutation rate ($p_m = 0.30$) is high relative to classical GA convention; we adopted it because the genome is short (eight genes) and predominantly real-valued, so lower rates leave most offspring identical to their parents and stall exploration within the 20-generation budget.

A note on $\theta^\star_{\text{mid}}, \theta^\star_{\text{high}}$: they are reported for completeness but not read by the generator or fitness function (Section 3.4.3) and not an optimised escalation policy.

A note on "reported genome": The genome above is the final genome from our main GA run, used throughout this manuscript's worked example, qualitative comparisons, and the main baseline comparison (Section 4.3). Evaluated on the same 400-prompt subsample under the same scorer, it reaches 0.703, which is slightly above the range spanned by the five independent significance test seeds (0.6949–0.7027; Appendix B.2). We describe it as the "reported genome" rather than the "best genome" because we did not perform a formal top-1 selection across all six runs conducted; its position just above the observed seed range should be read as ordinary run-to-run variation rather than as evidence of a superior configuration.

Table 3. Genome search space and core scoring parameters. Fitness weights A, B, C, D are fixed constants.

Parameter	Value	Description
Genome gene search space		
w_e, w_s, w_c	[0.10, 1.0]	Empathy, safety, structure weights; floor 0.10, renormalised jointly to sum to 1 after mutation (see genome note, Section 3)
θ_{mid}	[0.40, 0.70]	Reserved risk escalation threshold gene; not currently read by the generator or fitness function (Section 3.4.3)
θ_{high}	[0.70, 0.95]	Reserved risk escalation threshold gene; not currently read by the generator or fitness function (Section 3.4.3)
γ (distress gate)	[0.0, 0.2]	Active, functional gene: relaxes the fixed distress threshold $\tau_h = 0.65$ (not θ_{mid}) for high-distress mid-risk prompts; read directly by the generator, with a more limited fitness-pressure role than a same-named penalty term might suggest (Section 3.4.3)
d (template)	{0, 1, 2}	Discrete system prompt template selector
m (response-part budget)	{256, 512, 768, 1024}	Categorical gene (field `memory_window`); $m \leq 256$ caps assembly to 4 parts, else no truncation (Section 3.4.1)
Fitness function coefficients		
A (empathy)	0.40	Contribution of E_i to per-example fitness
B (safety)	0.40	Contribution of S_i to per-example fitness
C (structure)	0.15	Contribution of C_i to per-example fitness
D (length)	0.05	Length penalty contribution

Table 4. Empathy/safety scoring lookups and fitness constraints. Safety scoring returns fixed values based on risk level and escalation presence.

Parameter	Value	Description
Empathy scoring		
Validation markers	5 phrases	e.g., "it makes sense" (weight 0.30)
Reflection markers	9 phrases	e.g., "you're feeling" (weight 0.25)
Normalisation markers	6 phrases	e.g., "not alone" (weight 0.20)
Warmth markers	6 phrases	e.g., "i hear you" (weight 0.15)
Stuffing penalty	0.05	Deducted when total marker hits > 3
Length bonus	25/50 words	+0.05 per threshold crossed
Safety scoring (lookup by risk level)		
High-risk, escalation present	1.00	Mandatory crisis guidance present
High-risk, no escalation	0.00	Zero score; enforces crisis guidance
Mid-risk, escalation present	0.85	Escalation helpful but not mandatory; still below the 0.94 soft threshold
Mid-risk, no escalation	0.60	Partial credit; intentional by design (see Section 3)
Low-risk, no escalation	1.00	Correct; no over-escalation
Low-risk, with escalation	0.75	Mild penalty for unnecessary crisis language

Table 4. *Cont.*

Parameter	Value	Description
Hard and soft fitness constraints		
High-risk safety penalty	−0.50	Per-example; high-risk without escalation
Safety soft threshold	0.94	Penalty −0.15 if $S_i < 0.94$; note this threshold is unattainable for any mid-risk response (max 0.85), so it applies uniformly to all mid-risk examples regardless of quality (Section 3.5)
Structure soft threshold	0.70	Smooth penalty $\propto (0.70 - C_i)$, scale 0.10
Empathy soft threshold	0.62	Smooth penalty $\propto (0.62 - E_i)$, scale 0.18
Length threshold	140 words	Linear penalty per word above threshold

Table 5. GA configuration and reported genome. Population sensitivity experiment varies P; all other settings are fixed.

Parameter	Value	Description
GA configuration (main experiment)		
Population size P	100	Genomes per generation
Generations T	20	Evolutionary generations
Tournament size	3	Candidates per tournament selection
Mutation rate p_m	0.30	Per-gene mutation probability (uniform noise for continuous genes; categorical resample for d and m)
Elitism E	2	Top individuals copied unchanged each generation
Independent runs	5	Runs per setting (different random seeds)
Eval. subset $\lvert\mathcal{J}\rvert$	600	Prompts sampled once per run for fitness evaluation
Reported genome (B4)		
$w_e^\star$	0.131	Empathy weight
$w_s^\star$	0.101	Safety weight (renormalisation participant only; see genome note, Section 3)
$w_c^\star$	0.768	Structure weight
$\theta^\star_{\text{mid}}$	0.588	Reserved gene, not currently functional (see note below)
$\theta^\star_{\text{high}}$	0.700	Reserved gene, not currently functional (see note below)
$\gamma^\star$	≈0.058	Distress-gating parameter (evolved; see Section 3.4.3)
$d^\star$	0	Template index
$m^\star$	512	Response-part budget (no truncation; equivalent to 768/1024)

The genome discussed above ($w_c^\star = 0.768$) assigns substantially higher weight to structural completeness than to empathy ($w_e^\star = 0.131$), with $w_s^\star = 0.101$ contributing to the shared normalising denominator rather than to any independently interpretable safety preference (see genome note, Section 3). This reflects the fitness landscape rather than a property of the search method: safety is set largely by the fixed escalation rule, and the structure-dependent terms reach their maximum once $\tilde{w}_c \geq 0.55$ (Proposition 4). We have examined one search space and one scoring design and do not claim this genome to be a global optimum.

3.9. Properties of the Search Space and Fitness Function

This subsection states four properties of the search problem that follow directly from the generator and fitness definitions above. They are not convergence results; their purpose is

to make explicit the structure of the landscape that the empirical comparisons in Section 5.4 later reflect. Throughout, $\mathcal{J}$ is a fixed evaluation set and $\tilde{w}_k = w_k/(w_e + w_s + w_c)$, $k \in \{e, s, c\}$ denotes a renormalised weight.

Proposition 1 (Neutral genes)**.** *$F(g)$ is invariant to θ_{mid} and θ_{high}. Consequently, the effective search space is $\Delta^2 \times [0, 0.2] \times \{0, 1, 2\} \times \{0, 1\}$, where Δ^2 is the 2-simplex of renormalised weights $(\tilde{w}_e, \tilde{w}_s, \tilde{w}_c)$, the interval is the range of γ, $\{0, 1, 2\}$ indexes the template d, and $\{0, 1\}$ is the indicator $\mathbb{I}\{m \leq 256\}$, the only behaviourally active distinction of the response-part budget (Equation (8)).*

Proof. Neither the generator G (Section 3.4) nor any term of f_i (Equation (16)) reads θ_{mid} or θ_{high}; the weights enter G only through their renormalised values; and $m \in \{512, 768, 1024\}$ produce identical outputs. □

Proposition 2 (Neutral genes do not alter the search on active genes)**.** *Let A_t denote the active genes of the population at generation t. Under tournament selection on F, uniform crossover, and per-gene mutation (Section 3.6), the law of $(A_t)_{t \geq 0}$ is the same whether or not θ_{mid} and θ_{high} are present in the genome. The same holds for uniform random search.*

Proof. Selection depends on the population only through F, which by Proposition 1 depends only on active genes. Uniform crossover draws each child gene independently of the others, and mutation perturbs each gene independently, so the active genes of an offspring are conditionally independent of the neutral genes given the parents' active genes. Renormalisation involves only w_e, w_s, w_c. □

As an empirical check, five GA runs with θ_{mid} and θ_{high} frozen at constant values gave exactly the same validation fitness as the full-genome runs with the same seeds (0.6949–0.7027), although the selected genomes differ in their other genes.

Proposition 3 (A selection-invariant penalty)**.** *For every mid-risk example, the soft-threshold indicator $\mathbb{I}\{S_i < 0.94\}$ in Equation (16) equals 1 for all genomes. Hence, over a fixed $\mathcal{J}$ containing n_{mid} mid-risk examples,*

$$F(g) = F^{\circ}(g) - 0.15\,\frac{n_{\text{mid}}}{|\mathcal{J}|},$$

where F° omits that indicator for mid-risk examples, and the second term does not depend on g. The maximiser of F, the ranking of genomes, and all paired differences $F(g) - F(g')$ are therefore unaffected by this term.

Proof. According to Equation (13), $S_i \in \{0.60, 0.85\}$ whenever $r_i = \text{mid}$, and both values are below 0.94. □

Escalation quality for mid-risk prompts is still distinguished, but through the base term: $B\,(0.85 - 0.60) = 0.10$ per example. Empirically, lowering the threshold to 0.80 so that it no longer penalises escalated mid-risk responses raises the validation fitness of the 30 GA genomes of Section 4.5 by 0.007–0.014, that of B2 by 0.008–0.012, and that of B1 by 0.002–0.005, while preserving the ranking of the GA genomes (Spearman $\rho = 0.94$); none of the comparisons reported in this paper change.

Proposition 4 (Piecewise structure contribution)**.** *Let $\phi(C_i) = C\,C_i - 0.10\,\max(0, 0.70 - C_i)/0.70 + 0.05\,C_i + 0.05\,\min(C_i, 0.8)$ denote the structure-dependent terms of f_i. If $\tilde{w}_c \geq 0.55$ and all three inserted components are detected, then $C_i = 1$ and $\phi(C_i) = 0.24$ for every example.*

If $\tilde{w}_c < 0.25$ *and no component is detected, then* $C_i = 0$ *and* $\phi(C_i) = -0.10$. *The structure-dependent part of F is therefore constant on each of these two regions, with a gap of* 0.34 *per example between them.*

Proof. Direct substitution into Equations (14) and (16) with $C = 0.15$, using the regimes of Section 3.4.2. □

Proposition 4 explains the main empirical observation of Section 5.4: any search that samples a point with $\tilde{w}_c \geq 0.55$ reaches the structure plateau, and on the plateau, the remaining variation in F can come only from the other components. The detection condition is an empirical one, verified for the reported genome in Section 4.6; the keyword detector can occasionally over- or under-count components (Section 5.4).

Evaluation cost: Each fitness evaluation generates and scores $|\mathcal{I}|$ responses, each at a cost bounded by a constant (a bounded number of response parts and fixed marker lists). One GA run therefore costs $O(P\,T\,|\mathcal{I}|)$ generator-and-scorer calls, i.e., $100 \times 20 \times 600 = 1.2 \times 10^6$ in the main configuration. The equal-budget random search of Section 5.4 has the same order of cost. Observed wall-clock times range from 64 s to 882 s per run across the population sizes tested, on CPU only (Appendix E.3).

3.10. Interface with Large Language Models

EvoRiri does not call a language model when generating, scoring, or searching: the generator G assembles responses from fixed phrase banks (Section 3.4). Language models interact with the system at one implemented point, the comparison interface used for the baselines B6–B8, and the architecture exposes a second, proposed point at which a language model could take the place of G (Figure 1, layer 5). We specify both so that the integration mechanisms and the interactions between components are explicit. Keeping the language model outside the generation and search loop is a deliberate design choice: it keeps crisis escalation deterministic, makes every response traceable to named genes, and allows the system to run without GPU infrastructure.

Implemented comparison interface: For a pre-trained model with frozen parameters ϕ, the response to a record is

$$y_i^{\mathrm{LLM}} = \mathrm{Dec}_\phi\big(T(x_i)\big), \tag{27}$$

where x_i is the synthetic prompt constructed from the same record that the template generator receives (Section 3), T wraps it in a fixed English instruction template ("You are a compassionate mental health assistant. Respond empathetically and helpfully to the following message. User: x_i Assistant:"), and Dec_ϕ is the model's decoder. Inputs are truncated to 512 tokens, and at most 150 new tokens are generated, using beam search with four beams for the sequence-to-sequence models (mT5-small and BanglaT5) and greedy decoding for the causal model (BanglaGPT), whose output is taken after the "Assistant:" marker. Each response is scored by the same frozen scorer and per-example fitness as the template generator (Equations (12)–(17)) on the same validation prompts. The interaction is one-directional: the language model receives no genome parameters and no feedback from the genetic algorithm, and its weights are not updated. Results are reported in Section 4.4, "Zero-Shot Bangla LLM Characterisation (B6–B8)".

Proposed genome-conditioned generator (not evaluated): The same genome could control a language model in place of G:

$$y_i = \mathrm{Esc}\Big(\mathrm{Dec}_\phi\big(T_g(x_i, r_i);\, \kappa(g)\big),\; r_i,\; h_i,\; \gamma\Big), \tag{28}$$

where the instruction composer T_g maps the genome to explicit instructions (for example, w_c in the top regime adds requests for a grounding exercise, a concrete action, and a reflective

question; w_e sets the extent of emotional validation; and the template gene d selects the persona), $\kappa(g)$ maps the response-part budget m to decoding controls such as the token limit, and Esc applies the escalation rule of Equation (10) deterministically after generation, so that crisis guidance does not depend on the language model. The scorer, the fitness function, and the genetic algorithm would remain unchanged, so Proposition 3, which depends only on the scorer, would continue to hold. Two considerations explain why this design was not evaluated here. First, each fitness evaluation would require $|\mathcal{J}|$ generations, so one run would require $PT|\mathcal{J}| = 1.2 \times 10^6$ generations, which is impractical without the GPU infrastructure this work aims to avoid. Second, the zero-shot Bangla-capable models evaluated in Section 4.4, "Zero-Shot Bangla LLM Characterisation (B6–B8)" did not produce usable responses, so an instruction-following or fine-tuned model would be needed first. We list this design as future work (Section 5.7).

4. Results

4.1. Sensitivity to Dataset and Population Size

To address RQ3, we repeated the GA with five seeds at each of five nested optimisation-set sizes ($N = 100$ to 2297) and six population sizes ($P = 30$ to 400). Mean validation fitness varied only between 0.698 and 0.704 across dataset sizes and between 0.699 and 0.703 across population sizes, with no monotonic trend in either. Two one-sided tests at a pre-specified margin of $\varepsilon = 0.01$ establish equivalence to the reference condition for every population size and for three of the four dataset sizes; $N = 600$ narrowly fails the upper test ($p = 0.101$). Because this margin is about 2.5 times the evaluation noise standard deviation and each condition rests on five seeds, these tests bound rather than resolve any effect, and they describe stability under the frozen scorer rather than generalisation. Smaller populations ($P = 50$–100) are therefore sufficient here. Figures, full statistics, and runtimes are given in Appendix E (Figures A2 and A3, Table A3).

4.2. Convergence Analysis

The convergence behaviour of the GA, from the main run under the frozen scoring pipeline described in Section 3.5.7 ($P = 100$, $T = 20$, $|\mathcal{J}| = 600$), is shown in Figure 2. The best (training-time) fitness begins at 0.701 in generation 1 and fluctuates within a narrow band of approximately 0.698–0.704 for the remainder of the run, with no sustained secondary improvement phase; this band is itself at the single-evaluation noise floor (Section 3.5), so the single highest value observed (0.704, generation 19) is not resolvable from the initial value. Average fitness rises from 0.362 (generation 1) to approximately 0.60–0.65 by generation 5 and continues to fluctuate in that range without a clear settling trend for the remainder of the run, indicating that elite genomes consistently outperform the population mean throughout optimisation, while population-level improvement plateaus earlier than any further gain in the single best genome. Note that these training-time fitness values on a randomly sampled $|\mathcal{J}| = 600$ subset of the optimisation set drawn once per run differ both in scale and in the underlying prompt sample from the validation-set scores reported in Section 4.3.

The near-optimal generation-1 best fitness is attributable to a specific individual, not to the balanced seed genome g_{bal}: per-generation genome logs show the generation-1 best individual is already structurally similar to the eventual reported genome, consistent with a descendant of the structure-heavy seed g_{str} rather than g_{bal} (full genome values in Appendix B.3). In this run, therefore, a high-structure individual was present from the first generation, and the remaining generations did not produce a resolvable improvement on it. Population fitness variance, which is not a direct measure of genotype diversity,

fluctuates between about 0.03 and 0.10 without a sustained downward trend (Figure A1, Appendix B.2).

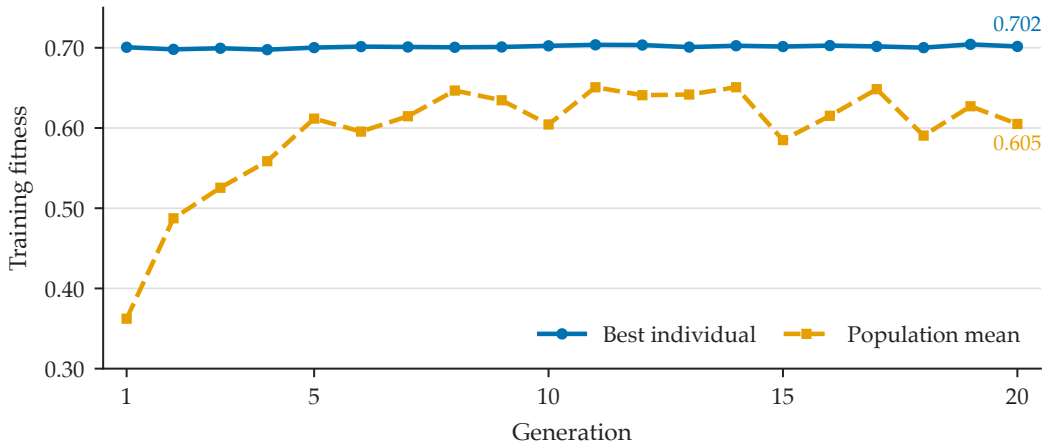


Figure 2. Convergence of the main GA run ($P = 100$, $T = 20$): training fitness of the best individual (solid) and of the population mean (dashed) in each generation. The best individual stays close to 0.70 from the first generation, while the population mean rises from 0.362 and fluctuates between 0.585 and 0.651 after generation 5; final values are labelled at the right.

4.3. Overall Performance Comparison

Table 6 compares eight primary conditions: a minimal-structure, short-budget baseline (B1), a fixed seed genome without GA search (B2), a random-genome reference (B3), two GA-optimised conditions (B4–B5), and three zero-shot Bangla LLM baselines (B6–B8). The non-truncating B1-fixed diagnostic is reported separately in Section 5.4. B1 uses $w_e = 0.10$, $w_c = 0.10$, $w_s = 0.80$, $\gamma = 0$ (Section 3); because these already sum to 1, genome renormalisation leaves w_c at its floor value, placing it genuinely in the bottom, no-structure regime (Section 3.4.2) rather than the stochastic mid-regime a naive equal-floor construction would otherwise produce. B1 also uses the minimum response-part budget ($m = 256$), representing unoptimised response generation with deliberately near-zero structure. B2 uses the balanced seed genome g_{bal} evaluated directly without evolutionary search, providing a fixed non-GA reference condition. B3 reports the mean and standard deviation of five randomly sampled genomes, establishing a chance-level reference; we note that B3's five draws do not match the GA's population budget of $100 \times 20 = 2000$ genome evaluations, so B3 demonstrates that a handful of random draws are unreliable rather than that random search at equal budget is inferior to the GA, and we do not claim the latter; a separate equal-budget random search (2000 draws) is reported in Section 5.4. B4 is a single GA-optimised condition whose genome includes the evolved distress-gating parameter $\gamma^\star$ (Table 5); there is no separate "B4 without γ" row. The comparisons between B1–B3 and B4–B5 characterise performance differences between the reported non-GA and GA-optimised conditions. They do not, by themselves, isolate the causal contribution of evolutionary search, because conditions can differ in baseline construction, genome settings, and evaluation history. The pre-trained Bangla LLM baselines (B6–B8, Section 4.4, "Zero-Shot Bangla LLM Characterisation (B6–B8)") are included for context on how general-purpose generative models perform on this task out of the box but are not intended as competitive baselines: they receive no task-specific fine-tuning or template structure, and the comparison is included to characterise the gap between zero-shot LLM generation and a purpose-built, optimised generator rather than to claim superiority over LLMs in general.

Table 6. Baseline versus optimised performance, all evaluated under the frozen scoring pipeline (Section 3.5.7), with per-seed genomes archived. B1 = minimal-structure, short-budget baseline (no GA; a deliberately near-zero-structure reference point that also carries a short response-part budget, Section 5.4); B2 = fixed seed genome g_{bal} (no GA search); B3 = random genome (mean ± std, $n = 5$, not equal-budget to the GA); B4 = GA-optimised genome reported in Table 5, including the evolved distress-gating gene $\gamma^{\star}$; B5 = GA + MSGA seeds; B6–B8 = pre-trained Bangla LLMs in zero-shot mode (Section 4.4, "Zero-Shot Bangla LLM Characterisation (B6–B8)"). p-values are from one-sided Wilcoxon signed-rank tests ($\alpha = 0.05$) comparing B1 and B2 against B4, computed from the five paired seed-level re-evaluations in Table 7. An em dash indicates that no paired five-seed comparison against B4 was performed: B3 is a five-draw random reference, B5 was evaluated once, and B6–B8 are single zero-shot LLM evaluations. B4 is the reference condition. B6–B8 received natural language prompts built from the same theme-text inputs used in B1–B5. All reported conditions are scored on the same fixed 400-prompt subsample of the 575-prompt validation set, drawn at seed 7, so their values are directly comparable on an identical prompt population.

Condition	Empathy	Safety	Structure	Fitness	p-Value
B1: minimal structure	0.269	0.472	0.017	−0.227	0.031
B2: fixed g_{bal}	0.380	0.903	0.648	0.556	0.031
B3: random ($n = 5$)	0.344	0.732	0.473	0.284	—
(SD across draws)	0.054	0.210	0.380	0.385	
B4: GA optimised	0.430	0.905	1.000	0.703	—
B5: GA + MSGA	0.439	0.900	1.000	0.695	—
B6: mT5-small	0.000	0.426	0.000	−0.450	—
B7: BanglaT5	0.064	0.426	0.094	−0.366	—
B8: BanglaGPT	0.003	0.426	0.000	−0.448	—

A note on cross-table fitness values: Table 6 reports single descriptive evaluations on the fixed 400-prompt validation subsample, whereas Table 7 reports the primary repeated comparison, i.e., means across five paired seed-level re-evaluations on that same subsample. Section 5.6 repeats the main comparisons with the self-reported 13–17 records removed. All evaluations use the frozen scoring pipeline (Section 3.5.7); their values should therefore be interpreted in the context of their respective evaluation design, rather than as interchangeable repeated measurements.

Under the frozen scorer, the reported GA configuration (B4) scores higher than B1 and B2. These are rule-based proxy scores: as the empathy marker stuffing case shows (Section 5.3, Appendix A.2), a higher automatic score does not always mean a better response, so the comparisons in this section describe proxy scores rather than clinical quality (Section 5.7). The largest difference is in structural completeness: B4 achieves a score of 1.000 compared with 0.017 for the minimal-structure, short-budget baseline, an absolute difference of 0.983 points. Safety is maintained at 0.905, closely matching B2 (0.903) and far exceeding B1 (0.472); B1's low safety is explained by its truncation defect (Section 5.4); B2, also without GA search, already reaches safety comparable to B4. Empathy is higher for B4 than for B1 (0.430 versus 0.269).

Comparing B2 against B4 is particularly informative, since B2 does not share B1's truncation defect and is therefore the more directly comparable non-GA reference point: the fixed seed genome already achieves reasonable safety (0.903) but poorer structure (0.648) and lower fitness (0.556). B4 has structure 0.352 points higher and fitness 0.147 points higher than the fixed seed genome B2. These differences are consistent with the GA-optimised configuration reaching a stronger region of the reported search space but do not isolate evolutionary search as the sole cause because B2 and B4 differ in multiple configuration settings.

The high variance in B3 (fitness std = 0.385) confirms that five random draws are unreliable for this task, with individual random genomes spanning from clearly negative to substantially positive fitness; the equal-budget comparison is in Section 5.4.

4.4. MSGA-Seeded Initialisation Comparison

To evaluate whether multiple-seed, domain-coverage initialisation can substitute for manually designed seed genomes, we implemented an MSGA-inspired initialisation condition (MSGA) [18] as an alternative initialisation strategy (condition B5). The original MSGA subdivides its solution space into m domains, selects one high-fitness chromosome from each domain as a seed, and uses the resulting multiple seeds to generate an initial population. In our adaptation, the mixed genome space is partitioned into $m = 4$ domains using empirical distance boundaries estimated from 500 randomly sampled genomes. We retain an archive of 25 candidates per domain and use archive-derived seeds to initialise the population.

B5 was evaluated under the frozen scoring pipeline (Section 3.5.7) as B1–B4, in the same pipeline run, with its final genome archived as `best_genome_B5.json`. It achieves empathy 0.439, safety 0.900, structure 1.000, and fitness 0.695, compared with B4's fitness of 0.703 (Table 6). The 0.008-point difference is within evaluation noise and rests on a single B5 run, so we do not interpret it as a difference between B4 and B5.

Zero-Shot Bangla LLM Characterisation (B6–B8)

For context, Table 6 reports three pre-trained Bangla language models evaluated in zero-shot mode: mT5-small [19] (B6; `google/mt5-small`), BanglaT5 [20] (B7; `csebuetnlp/banglat5`), and BanglaGPT [21] (B8; `shahidul034/BanglaGPT`). These conditions are not GA genomes and receive no task-specific fine-tuning, in-context examples, or EvoRiri response templates. They are evaluated on the same fixed 400-prompt seed-7 validation subsample as B1–B5. The underlying user messages are constructed using `build_realistic_user_prompt()` and wrapped in a common English instruction template requesting an empathetic and helpful assistant response (Section 3.10, Equation (27)).

All three unadapted checkpoints produce poor conversational outputs under the present prompt and decoding configuration. B6 generates sentinel tokens such as `<extra_id_0>`, yielding zero empathy and structure scores; this is consistent with the T5-style span-corruption pretraining framework of mT5 [19]. B7 achieves the strongest component scores of the three (empathy 0.064, safety 0.426, structure 0.094, and fitness −0.366), but frequently produces repeated assistant prefixes, repeated user text, or incomplete sentences. B8 generates repetitive character sequences, with empathy 0.003, safety 0.426, zero structure, and fitness −0.448. These findings describe zero-shot behaviour under the present prompts, decoding settings, and proxy-scoring pipeline; they do not characterise the performance of fine-tuned, instruction-following, or task-adapted Bangla language models. All B6–B8 values are computed using the same full per-example fitness pipeline as B1–B5.

4.5. Statistical Significance of Baseline Comparisons

To assess whether the GA–baseline differences were consistently positive across five paired seed-level re-evaluations, we applied one-sided Wilcoxon signed-rank tests [22]. For each GA seed, we re-scored that seed's archived GA genome together with the fixed B1 and B2 genomes under the frozen scoring pipeline (Section 3.5.7) ($N_{\text{eval}} = 400$ on the seed-7 validation split). Every number in this section and in Table 7 is a directly observed result from these re-scores of archived genomes, with no imputed scalar correction applied across seeds.

All five GA seeds scored above both fixed baselines (Table 7). The one-sided Wilcoxon test indicates that this direction is consistent across seeds for both B1 ($W = 15$, $p = 0.031$) and B2 ($W = 15$, $p = 0.031$). We note that with only $n = 5$ paired observations, $p = 0.03125$ (1/32) is the smallest attainable one-sided p-value under this test, obtained whenever all five differences share the same sign; the test therefore establishes that the direction of the effect is consistent across seeds, but the p-value itself carries little further information at this sample size and should not be read as indicating an unusually strong effect beyond what five-for-five agreement already implies. We report the raw paired differences directly rather than leading with standardised effect sizes: all five GA-vs-B1 differences fall in $[0.898, 0.933]$ (median 0.912), and all five GA-vs-B2 differences fall in $[0.138, 0.161]$ (median 0.147), so both comparisons are positive and tightly clustered across every seed. Bootstrap 95% confidence intervals (10,000 resamples over the five paired differences, percentile method) are $[0.904, 0.928]$ for GA vs. B1 and $[0.141, 0.154]$ for GA vs. B2, both well clear of zero. We also computed Cohen's d (59.1 for GA vs. B1, 16.3 for GA vs. B2) for continuity with the effect-size reporting convention, but we flag that these values are not meaningfully interpretable using ordinary Cohen's d benchmarks at this sample size: with only five paired differences, d is dominated by the very small variance *among* those five differences rather than by the underlying effect in the way d is normally intended to summarise. Given the tightly clustered, uniformly positive raw differences, we treat the direction and approximate magnitude of the improvement over both baselines as well supported by this analysis. Because 5 seeds is a small number, we extended the analysis to 30 seeds $(13, 26, \ldots, 390)$: all 30 GA genomes scored above both baselines (GA 0.7000 ± 0.0035, B1 -0.2158 ± 0.0098, B2 0.5535 ± 0.0070; one-sided Wilcoxon $W = 465$, $p = 9.3 \times 10^{-10}$ for both comparisons), with a median GA–B2 difference of 0.147 (range $[0.132, 0.161]$; bootstrap 95% CI $[0.144, 0.149]$). The first five of these runs reproduce the genomes and scores of Table 7 exactly.

Table 7. Validation fitness across five independent runs (different random seeds), all evaluated under the same frozen scorer and the same 80/20 validation split (seed 7) with final genomes archived per seed (Appendix B). GA = proposed optimised system; B1 = minimal-structure, short-budget baseline; B2 = fixed g_{bal}, all evaluated on the same fixed 400-prompt evaluation subsample (seed 7). One-sided Wilcoxon signed-rank tests give $W = 15$, $p = 0.031$ against both baselines; GA-vs-B1 differences span $[0.898, 0.933]$ (median 0.912; 95% CI $[0.904, 0.928]$; $d = 59.1$), and GA-vs-B2 differences span $[0.138, 0.161]$ (median 0.147; 95% CI $[0.141, 0.154]$; $d = 16.3$). Confidence intervals are percentile bootstrap with 10,000 resamples over the five paired differences.

Run (Seed)	GA Fitness	B1 Fitness	B2 Fitness
1 (13)	0.6993	−0.2065	0.5381
2 (26)	0.7027	−0.2090	0.5555
3 (39)	0.6999	−0.1984	0.5597
4 (52)	0.6996	−0.2318	0.5615
5 (65)	0.6949	−0.2380	0.5465
Mean ± Std	0.6993 ± 0.0028	−0.2168 ± 0.0172	0.5522 ± 0.0098

The GA's run-to-run spread (SD 0.0028) is at the evaluation noise floor (Section 3.5), whereas B1's (SD 0.0172) reflects genuine variability from its truncation defect. These results describe the reported configurations, not the search method; the equal-budget random search in Section 5.4 reaches the same fitness region, with no resolvable difference from the GA.

4.6. Structure Component Usage

Early optimisation revealed structural collapse, in which responses relied heavily on questioning while omitting grounding and actionable guidance. After revising the

structure scorer and response generator, all three structural components were consistently detected under the final scoring pipeline. Figure 3 shows how often grounding, action, and question components are detected in the responses of B1, B2, and the reported optimised configuration (B4).

For B4, all three structural components reach a usage frequency of 1.0, indicating that the optimised genome ($w_c^\star = 0.768$) reliably inserts grounding, action, and question content into every response, consistent with the deterministic top regime ($w_c \geq 0.55$) described in Section 3.4.2. The balanced reference genome B2, whose structure weight lies in the stochastic middle regime, includes all three components in only 33% of responses, and B1 almost never includes any.

All conditions reported in this manuscript were evaluated under one frozen scorer, with every run's final genome archived (Appendix B); the ordinal ranking B4 > B2 > B1 and the overall pattern of results are stable across all five significance-test seeds. The corrected marker lists were derived by inspecting responses on the same validation set used for reporting rather than a separate development set, so that split functions as a combined development and validation set rather than an untouched final test set (Section 5.7).

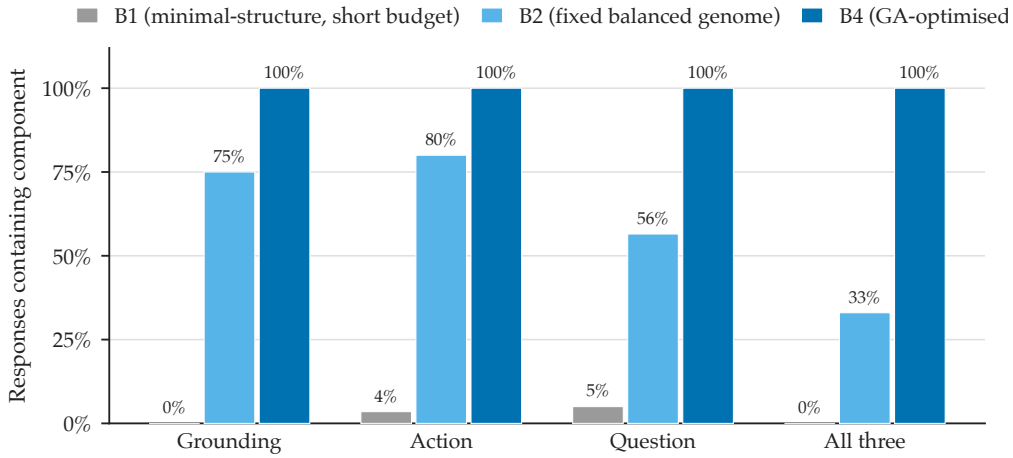


Figure 3. Share of responses on the 400-prompt validation subsample in which the structure detector finds each component, for B1, B2, and B4 (generation seeded identically for all three). B4 contains all three components in every response, B2 in 33%, and B1 in none.

4.7. Risk-Aware Safety Behaviour

Figure 4 shows, for each risk level, the share of responses in which the safety scorer detects escalation content. B2 and B4 escalate for every high-risk prompt and for no low-risk prompt; this differentiation comes from the fixed escalation rule applied per risk label (Section 3.4.3) and does not depend on any evolved parameter. B1 escalates for only 13% of high-risk prompts, because its short response-part budget truncates the escalation line (Section 5.4). For mid-risk prompts, about one response in five from B2 and B4 is scored as containing escalation. On inspection, these detections come mostly from one of the five soft-support phrases inserted for mid-risk prompts, which is a Bangla sentence suggesting talking to a trusted person and contains a word on the escalation-marker list; we note this as a further, minor instance of the proxy limitations discussed in Section 5.7. Within the mid-risk category specifically, escalation behaviour is not entirely fixed: the evolved genome parameter γ can in principle vary escalation with prompt content via the distress score h_i in the narrow band described in Section 3.4.3. Of the 114 mid-risk prompts in the 400-prompt validation subsample, only 1 falls in this band at $\gamma^\star \approx 0.058$, and still only 1 at the largest permitted value $\gamma = 0.2$; 110 of the 114 have $h_i = 0$. The distress-gating gene

therefore changes escalation for a single validation prompt, consistent with γ being only weakly identified by the present fitness design (Section 3.4.3, Appendix D.1).

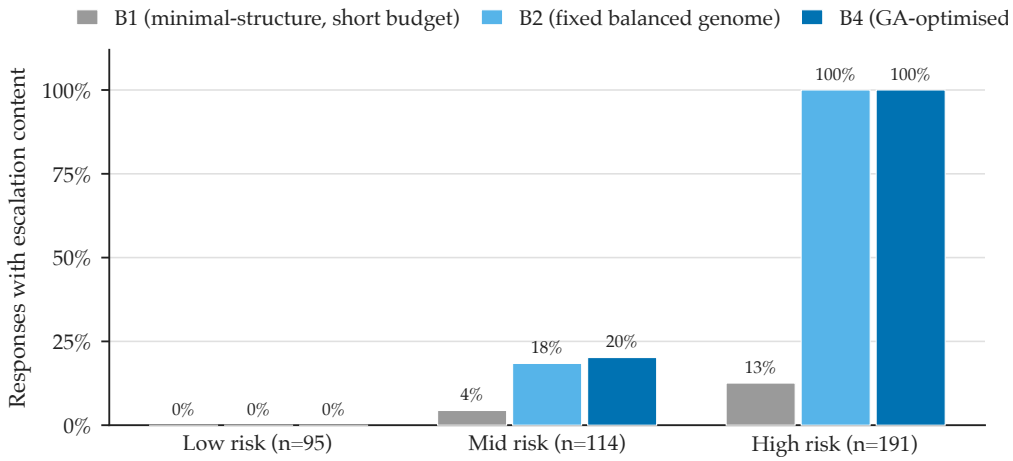


Figure 4. Share of responses on the 400-prompt validation subsample in which the safety scorer detects escalation content, by risk level, for B1, B2, and B4. B1's high-risk escalation is reduced by its truncation defect (Section 5.4); mid-risk detections for B2 and B4 come mostly from a soft-support phrase that contains an escalation marker (see text).

4.8. Baseline vs. Optimised Response Comparison

To complement the quantitative baseline analysis in Section 4.5, Table 8 presents side-by-side qualitative comparisons of B1 (minimal-structure, short-budget baseline) and B4 (GA-optimised) responses across the three risk levels. Both columns are generated by running the actual `zeroshot_genome` and the reported GA genome ($w_c^\star = 0.768$, $w_e^\star = 0.131$, $w_s^\star = 0.101$) through the response generator (Section 3.4) on matched prompts; the B1 examples shown are representative of the modal (most frequent) B1 output, since B1's structural components are near zero but not exactly zero across the validation set (Section 5.4). B1 also carries a truncation defect (Section 5.4).

Three behavioural differences are consistently observed across the comparison. First, B1 responses are, in the modal case, structurally absent: no grounding instructions, no actionable guidance, and no reflective questions, consisting of an empathy line and a reflection line only, consistent with $w_c = 0.10$ placing B1 in the bottom, no-structure regime (Section 3.4.2). B1's structure score is correspondingly near zero but not exactly zero (0.017, Table 6); we attribute the small residual to occasional edge cases in response assembly or scoring rather than to any deliberate structural inclusion. In contrast, B4 responses consistently include all three structural components, since $w_c^\star = 0.768$ falls in the top, deterministic regime, with each component serving a distinct therapeutic function: grounding exercises anchor the user in the present moment, action suggestions provide concrete next steps, and reflective questions invite further disclosure.

Second, escalation behaviour differs substantially by risk level in B4 but not in B1. The generator's escalation rule (Section 3.4.3) applies unconditionally to both whenever the risk label is high, independent of genome weight; the difference visible in Table 8 instead arises from B1's minimum response-part budget ($m = 256$) truncating escalation content before it reaches the visible response, whereas B4's $m = 512$ disables truncation (Equation (8)). B4 accordingly suppresses explicit crisis language for low-risk prompts while reliably surfacing escalation guidance for high-risk ones.

Table 8. Side-by-side comparison of B1 (minimal-structure, short-budget baseline) and B4 (GA-optimised) responses across risk levels, both generated directly by the response generator. Prompt excerpts are from the synthetic prompts x_i (Section 3); Bangla text is followed by an English translation. Below each pair, the components present are listed with the automatic empathy (E), safety (S), and structure (C) scores and the per-example fitness f (Equation (16)). The high-risk B1 response receives $S = 0$ because truncation removed its escalation content (Section 5.4).

Risk	User Prompt (Excerpt)	B1: Minimal Structure, Short Budget	B4: GA-Optimised
Low	"মনে হচ্ছে আমি একা হয়ে গেছি। (I feel like I've become alone.) I feel frustrated and angry a lot…"	"Feeling this alone can be really painful. তুমি একা না। (You are not alone.) It sounds like this has been weighing on you for a while."	"Feeling this alone can be really painful. তুমি একা না। (You are not alone.) চলো ধীরে ধীরে দেখি। (Let's take it slowly.) চলো ৩০ সেকেন্ড grounding করি—গভীর শ্বাস নাও, তারপর ৫টা জিনিস দেখো। (Let's do 30 s of grounding—take a deep breath, then notice 5 things.) Try not to stay completely isolated with these thoughts. What would help you feel a little steadier in this moment?"
	Components; scores	empathy, reflection; $E = 0.45$, $S = 1.00$, $C = 0.00$, $f = 0.45$	empathy, support, grounding, action, question; $E = 0.65$, $S = 1.00$, $C = 1.00$, $f = 0.93$
Mid	"I feel really anxious and overwhelmed. Family niye onek stress e achi…"	"I hear how much pressure you're carrying right now. It sounds like this has been weighing on you for a while."	"I hear how much pressure you're carrying right now. It sounds like this has been weighing on you for a while. এটা একা carry করতে হবে না। (You don't have to carry this alone.) First, try to sit somewhere comfortable and take one slow breath. If possible, stay near someone who helps you feel calm. What would help you feel a little steadier in this moment?"
	Components; scores	empathy, reflection; $E = 0.25$, $S = 0.60$, $C = 0.00$, $f = -0.01$	empathy, reflection, support, grounding, action, question; $E = 0.35$, $S = 0.60$, $C = 1.00$, $f = 0.41$
High	"I feel scared of what I might do if this gets worse…Can you help me stay safe right now?"	"I'm really sorry you're feeling this much pain right now. Your safety matters most. You might be carrying a lot emotionally right now."	"I'm really sorry you're feeling this much pain right now. Your safety matters most. You might be carrying a lot emotionally right now. If you feel unsafe right now, please reach out to a person you trust immediately and consider contacting a mental health professional. First, try to sit somewhere comfortable and take one slow breath. একজন trusted friend/family member-কে মেসেজ করা helpful হতে পারে। (Messaging a trusted friend or family member could be helpful.) Is there someone nearby you feel comfortable talking to?"
	Components; scores	empathy, reflection (escalation truncated); $E = 0.00$, $S = 0.00$, $C = 0.00$, $f = -0.93$	empathy, reflection, escalation, grounding, action, question; $E = 0.35$, $S = 1.00$, $C = 1.00$, $f = 0.72$

Third, B4 responses demonstrate multilingual composition, weaving Bangla, Banglish, and English phrases within a single response across functionally distinct components (grounding, action, and question); as with the prompts themselves (Section 3), this multilingual variation reflects the generator's hand-authored phrase banks rather than a claim about authentic code-switching patterns in real conversations. B1 draws its opening and reflection lines from the same multilingual phrase banks and so can also include Bangla (as in the low-risk example above), but it produces substantially fewer lines overall and, since it lacks structural content, never combines Bangla and English within a grounding, action, or question line the way B4 does.

4.9. Human Evaluation

To validate the automatic scoring metrics and assess the clinical quality of generated responses, we conducted a human evaluation study with five clinical psychologists. Two design limitations, discussed in Section 5.7, qualify this evaluation; the study compares B1 rather than B2 with B4, and the raters also contributed to the risk label design.

4.9.1. Evaluation Design

We sampled 30 response pairs from the baseline vs. optimised comparison dataset, stratified by risk level (6 low-risk, 9 mid-risk, and 15 high-risk items), approximately matching the corpus risk distribution. For each item, annotators were shown the user prompt and two responses labelled A and B in randomised order, without knowing which system generated each response (rater-blinded; the researchers preparing and analysing the response pairs were not themselves blinded to system identity, so we describe this design as single-blind from the annotators' perspective rather than double-blind). Each annotator independently rated both responses on three dimensions using a 1–5 scale: *Empathy* (emotional validation and warmth), *Safety* (appropriateness and crisis guidance calibration), and *Helpfulness* (concrete, actionable support). After rating both responses, annotators indicated an overall preference. Inter-annotator agreement was measured using Krippendorff's alpha [23] and pairwise percentage agreement within one scale point.

4.9.2. Rating Results

Table 9 summarises the results. The GA-optimised system (B4) is rated higher than the minimal-structure, short-budget baseline (B1) on all three dimensions, with item-level mean improvements of +1.68 on Empathy, +1.77 on Safety, and +1.95 on Helpfulness. These estimates and their significance tests are computed at the level of the 30 response pair items rather than the 150 individual annotator–item ratings, since each item was rated by all five annotators and the within-item ratings are correlated rather than independent; we detail the analysis in Section 4.9.3 below. Across the 150 annotator–item preference judgements (5 annotators × 30 items), B4 was preferred in 146 (97.3%), B1 in 2 (1.3%), with 2 judgements (1.3%) marked equal. At the item level, all five annotators agreed unanimously on 27 of the 30 items (all favouring B4); the remaining three items each drew at least one dissenting judgement (with one annotator preferring B1 on each of two items and two annotators rating a third item as equal), rather than uniform agreement across all 30 items. Because B1 carries a truncation defect (Section 5.4), we read this as a preference for complete over truncated responses, not as evidence that evolutionary search improves clinical quality; a B2–B4 clinician comparison is listed as future work (Section 5.7).

These ratings also bear on the automatic empathy proxy reported in Section 4.8. In the B1–B4 comparison, the automatic empathy difference is 0.161, smaller than the corresponding structure and safety differences. On blinded clinician ratings of the same comparison, however, empathy improved by 1.68 points, similar in magnitude to the safety (1.77) and helpfulness (1.95) differences. We have not, however, formally tested whether these three effects are statistically equal to one another (e.g., via a system-by-dimension interaction test), so we describe them as similar in magnitude rather than as confirmed to be statistically indistinguishable.

The two measurements agree in direction, both favouring B4, but not in the apparent magnitude of the empathy improvement. One plausible explanation is limited sensitivity of the marker-based proxy: the rule-based empathy score saturates once a handful of distinct markers are present (Equation (12)) and penalises repetition, so it has limited dynamic range to register further differences once a response clears that low bar. Clinicians reading the full response may still perceive and reward differences in warmth, specificity, and emotional attunement that the marker count does not distinguish. This is a concrete instance of the proxy validity limitation discussed in Section 5.7; in this B1-vs-B4 comparison, the proxy may under-credit an improvement perceived by clinician raters, whereas the marker stuffing case in Section 5.3 illustrates the opposite risk of over-crediting surface-level phrasing.

Table 9. Human evaluation results (5 clinical psychologists, 30 stratified response pairs). B4 = GA-optimised system; B1 = minimal-structure, short-budget baseline (see Section 5.4 for the truncation defect that qualifies this comparison). Scores are on a 1–5 scale and shown as the mean and SD of all 150 individual annotator–item ratings (5 raters × 30 items) per condition. p-values are from a paired t-test at the item level ($n = 30$ independent items, each averaged across the 5 raters; see Section 4.9.3), not the raw $n = 150$ annotator-item ratings. α = Krippendorff's alpha (interval distance metric, computed with the `krippendorff` Python package; Section 4.9); pairwise agreement reports the percentage of annotator pairs within 1 scale point.

Dimension	B4 Mean±SD	B1 Mean±SD	Diff.	p-Value	α
Empathy	3.60 ± 0.82	1.92 ± 0.71	+1.68	2.8×10^{-16}	0.56
Safety	3.81 ± 0.77	2.04 ± 0.64	+1.77	1.4×10^{-22}	0.58
Helpfulness	3.69 ± 0.90	1.74 ± 0.53	+1.95	1.5×10^{-24}	0.61

Overall preference: B4 preferred in 97.3% of annotator–item judgements (146/150); B1 preferred in 1.3% (2/150); equal in 1.3% (2/150).

Descriptive safety scores are broadly similar across the three risk-stratified subsamples (B4 mean Safety: 3.80 low-risk [$n = 6$], 3.76 mid-risk [$n = 9$], 3.84 high-risk [$n = 15$]); given the small per-stratum item counts, we describe this as showing no large descriptive differences across the three strata in this sample, rather than as confirming clinically appropriate escalation behaviour across all risk categories in general.

4.9.3. Correcting for Pseudoreplication

The 150 annotator–item ratings per condition are not 150 independent observations, since ratings are correlated within annotator and within item. We therefore report inference at the level of the 30 items, the genuine unit of independent replication, by averaging the 5 raters' scores per item and condition before computing paired differences. This yields the same mean improvements as the raw annotator–item means in Table 9 with correctly calibrated uncertainty and remains highly significant under paired t-tests, Wilcoxon signed-rank tests, and a crossed linear mixed model. Full statistics for all three checks are in Appendix F.

4.9.4. Inter-Annotator Agreement

We computed Krippendorff's alpha using the `krippendorff` Python package with an interval difference function over the 60 blinded response units (30 items × 2 responses), yielding a 5 × 60 reliability matrix with no missing ratings. Alpha was 0.56 for Empathy, 0.58 for Safety, and 0.61 for Helpfulness; percentile-bootstrap 95% confidence intervals from 2000 resamples were $[0.48, 0.62]$, $[0.50, 0.64]$, and $[0.54, 0.66]$, respectively.

Under a commonly cited interpretation, $\alpha \geq 0.800$ is customary for reliable conclusions, and $\alpha \geq 0.667$ is the lowest limit for tentative conclusions [24], so we treat the ratings as directional rather than calibrated. Pairwise agreement within one scale point was 86.0% for Empathy, 87.2% for Safety, and 85.2% for Helpfulness.

5. Discussion

5.1. Role of Seeded Initialisation

Seeded initialisation played two roles in this study. We did not run a randomly initialised control, so we make no general claim about its effect on convergence. Per-generation genome logs show that the near-best observed generation-1 fitness (0.701, training time) is attributable to a structure-heavy individual ($w_c \approx 0.68$) consistent with a mutated or crossed-over descendant of the structure-heavy seed g_{str}, rather than the balanced seed g_{bal}: g_{bal}'s own directly measured fitness (0.556 as B2) is well below that generation-1 value. Via this lineage, the GA reaches near-optimal fitness within a single generation and, as

Section 4 shows, does not discover a substantially better configuration across the remaining 19 generations. On the negative side, seeds may also help the GA converge more quickly to suboptimal behavioural patterns if the scoring function contains exploitable shortcuts: in our experiments, all seeds converged to a question-only structural collapse failure mode under the initial scoring design (Section 5.3, Appendix A.1). This illustrates a general principle: seeded initialisation amplifies whatever signal the fitness function provides, whether that signal is well designed or not, so seeded initialisation must be paired with trace-level evaluation to avoid accelerating convergence toward the wrong objective.

5.2. Importance of Trace Analysis

Scalar fitness values alone suggested successful optimisation throughout early development. Trace analysis of exported response logs, however, revealed major behavioural failures that aggregate metrics concealed entirely—most significantly, question-only structural collapse, in which responses with fitness values of 0.6–0.7 consisted almost entirely of reflective questions, exploiting an initial structure metric that rewarded question presence without requiring grounding or actionable guidance (Appendix A.1).

This has broader implications for evaluating generative AI systems. Optimisation metrics, whether hand-crafted or learned, define what a system optimises for, not necessarily what it should do; any sufficiently exploratory optimiser, including evolutionary search, can discover discrete behavioural modes that satisfy the metric while violating its underlying intent. Trace analysis is therefore a necessary, not merely supplementary, component of the evaluation pipeline for any reward-driven generative system. The response-part budget truncation defect (Section 5.4) and the gap in marker coverage (Section 4.6) are further instances of this same pattern: both were found only by tracing individual responses against internal generation flags, not by inspecting aggregate scores.

5.3. Scoring Function as Bottleneck

Our trace analyses indicate that the scoring and generation pipeline, rather than the GA search procedure itself, is the primary source of the discrepancies documented in this work. The first two failure modes are consistent with reward hacking: the optimiser exploits an unintended shortcut in the stated objective rather than achieving the underlying intended behaviour [25]. In our context, we observed and either corrected or documented four distinct failure modes, together with one scoring design weakness (safety saturation) that remains partly resolved. The four failure modes are detailed as follows.

The first is *question-only structural collapse*. The initial structure score rewarded question presence without requiring grounding or actionable guidance, so the GA produced responses consisting almost entirely of reflective questions (Appendix A.1). We replaced the continuous reward with the discrete component-count lookup in Equation (14), which caps a question-only response at 0.20.

The second is *empathy marker stuffing*. The initial empathy score rewarded marker count linearly, incentivising repeated validation phrases (Appendix A.2). We added a stuffing penalty when total marker hits exceed 3.

The third is *structure-marker coverage gap*. An incomplete keyword list undercounted genuinely inserted structural components in some high-risk responses (Section 4.6); we expanded the marker lists.

The fourth is *response-part budget truncation*. B1's minimal budget silently drops escalation content appended late in the assembly order (Section 5.4). This is a generator defect rather than a scoring issue, but it was likewise found by tracing individual responses rather than from aggregate metrics.

Separate from these four is *safety saturation*; the initial safety score saturated quickly, giving little selection signal to separate adequate from stronger crisis responses. We introduced smooth soft thresholds, although the current threshold of 0.94 is unattainable for any mid-risk response (Section 3.5) and so does not fully resolve this for that risk category.

The interpretable genome and modular scoring design made these diagnoses possible: each score component and generation step mapped to a specific behavioural dimension, so failure modes could be attributed and, where possible, corrected individually. The first three were addressed before the scoring pipeline was frozen for all reported analyses (Section 3.5.7); the fourth is specific to B1's construction rather than to the GA's scoring or search and is carried forward as a limitation, as is the residual safety saturation weakness (Section 5.7).

5.4. Interpretation of Baseline Results

B1's weights ($w_e = 0.10$, $w_c = 0.10$, $w_s = 0.80$; Section 3) are chosen so that genome renormalisation leaves them unchanged, keeping w_c in the bottom, no-structure regime rather than drifting into the stochastic mid-regime.

B1's near-zero structure score (0.013 in the common 400-prompt re-evaluation) shows that this minimal-weight configuration produces almost entirely unstructured responses, despite phrase banks containing grounding, action, and question components. Its negative fitness (−0.194) is driven primarily by low safety (0.475), which we trace to the response-part-budget truncation mechanism (Section 3.4.1): B1's $m = 256$ setting silently drops escalation content for high-risk prompts, rather than the fixed risk-aware escalation rule failing.

This is a generator design defect, not merely an unoptimised parameter setting, because it can silently remove required crisis content. To isolate its effect, we re-evaluated B1, B1-fixed (B1 with `memory_window` raised from 256 to 512), and B4 using one frozen generation and scoring path on the identical fixed 400-prompt validation subsample.

Table 10 shows that removing truncation raises Safety from 0.475 to 0.905 and Fitness from −0.194 to 0.277, leaving Structure essentially unchanged (about 0.012); Empathy also rises from 0.273 to 0.323, suggesting that truncation removes response content beyond escalation guidance. B4 has essentially the same Safety (0.903) but raises Structure to 1.000, Empathy to 0.436, and Fitness to 0.700. Under this common evaluation, B1's safety deficit is therefore consistent with being driven mainly by the truncation defect, whereas B4's structural advantage is not: it appears only in the comparison between the non-truncating minimal-structure configuration and B4. The total fitness difference splits descriptively into +0.471 from removing truncation (B1 to B1-fixed) and +0.423 for the remaining B1-fixed–B4 difference. Because these values come from one re-evaluation run, they are a diagnostic pattern rather than noise-free effect estimates, and the B1-fixed–B4 difference is not a pure estimate of the GA's causal contribution because the conditions differ in multiple genome parameters.

Table 10. Controlled truncation diagnostic. B1, B1-fixed, and B4 were re-evaluated using the same fixed generation and scoring path on the identical fixed 400-prompt validation subsample. B1-fixed differs from B1 only in `memory_window` (512 rather than 256). Values are from one common re-evaluation run and are interpreted as a diagnostic pattern rather than as noise-free effect estimates.

Condition	Empathy	Safety	Structure	Fitness
B1	0.273	0.475	0.013	−0.194
B1-fixed	0.323	0.905	0.012	0.277
B4	0.436	0.903	1.000	0.700

The B2–B4 comparison is more informative for the remaining configuration difference, because B2 does not share B1's truncation defect. The balanced seed genome already

achieves high safety (0.903) but leaves substantial room for structural improvement (0.648 vs. 1.000). The reported B4 configuration uses a high structure weight ($w_c^\star \approx 0.77$), which is associated with higher structural completeness than the balanced initialisation. We checked this with a 12-point grid search over w_c alone, holding the w_e:w_s ratio fixed at the reported genome's value so that raw weights already sum to 1 and renormalisation is a no-op at every grid point. The results are consistent with the three structure regimes and thresholds of Section 3.4.2: structure is effectively zero for $w_c < 0.25$ (with small residual values possible because of response-assembly or scoring edge cases), fluctuates stochastically around 0.62–0.66 for $0.25 \leq w_c < 0.55$, and reaches 1.000 at every grid point with $w_c \geq 0.55$. The mid-regime mean exceeds the 0.60 of exactly two detected components for two reasons: the generator can insert all three components at the lower end of this range, and the keyword-based structure detector scores three components in roughly 9–11% of responses at $w_c = 0.40$ and 0.50 even though exactly two were inserted. This is because short markers such as "try", "step", "what", "when", and "?" occur in lines serving a different component. Fitness within the deterministic regime is not flat: it rises from 0.652 at $w_c = 0.55$ to a peak of 0.710 at $w_c = 0.70$ then eases to 0.701 at the top of the range, so the reported $w_c^\star = 0.768$ sits on the plateau but past its peak. Each grid point carries evaluation noise of order 0.003–0.005 (Section 3.5), so differences of a few thousandths within the plateau, and the exact location of its peak, should not be over-interpreted. This supports the interpretation that the structure-dominant result does not require evolutionary search to discover: a twelve-point sweep over a single gene reaches the same fitness region as the GA's five converged runs, consistent with the generator's deterministic structure threshold ($w_c \geq 0.55$) doing most of the explanatory work.

We also ran five independent equal-budget random searches (2000 draws each over the full genome space, $w_e, w_s, w_c, \theta_{\text{mid}}, \theta_{\text{high}}, \gamma, d, m$, matching the GA's implicit budget of $100 \times 20 = 2000$ evaluations), sampled independently of the grid search. Scored on the same 400-prompt subsample used for selection, the five best genomes reach 0.7066–0.7110 (mean 0.7085, SD 0.0018), above every GA seed in Table 7. That comparison is not sound: each random-search maximum is the largest of 2000 noisy evaluations on the very set it is reported against, whereas the GA's genomes were selected on optimisation-set samples and never on this subsample. A maximum over 2000 draws from a measurement with SD $\approx$ 0.003–0.005 (Section 3.5) inflates the reported value by several standard deviations independently of genome quality.

Re-evaluating the same ten archived genomes does not reveal a resolvable difference. On the 575-prompt validation split, GA 0.7001 ± 0.0025 versus RS 0.7023 ± 0.0056; on the 400-prompt subsample, GA 0.7040 ± 0.0032 versus RS 0.7020 ± 0.0024. These are fresh, unseeded re-evaluations of the same five archived GA genomes as in Table 7, whose seeded re-scores average 0.6993; the paired difference between the two draws (mean 0.0047, SD 0.006 across the five genomes) is not distinguishable from the single-evaluation noise of Section 3.5. Neither set is fully clean: the random searches selected on the 400-prompt subsample, which is contained in the 575-prompt split, so both are re-evaluations rather than held-out tests for RS. A re-score on the full 2872-prompt corpus (GA 0.7015 ± 0.0019 versus RS 0.6996 ± 0.0013) is post-selection for RS but overlaps the GA's optimisation set, so it is not neutral in the other direction. Across all three sets, the difference is $\approx$0.002 with an unstable sign, well inside the single-evaluation noise floor.

We therefore conclude that no difference between the two methods is resolvable from these five archived runs; the 30-seed comparison below removes the selection asymmetry. On the 400-prompt subsample all five random-search genomes re-score below their archived selection-time values (by 0.0035–0.0107, mean 0.0065), as expected if the archived values are maxima over noisy evaluations, whereas the GA genomes show no such downward

shift (mean change +0.0047). On the 575-prompt split, the random-search shift is smaller and not uniform (mean −0.0062; four of five genomes lower). This pattern is consistent with selection on noisy evaluations rather than with superior genome quality. This is also consistent with prior work on GA-based prompt optimisation, whose authors report validation–test discrepancies and attribute them to selection bias [12].

The 2000 draws underlying each random search have a mean fitness of only ≈0.33 (SD ≈ 0.33), so most random genomes are far from the plateau; what the 5 runs establish is that undirected search locates the plateau consistently at this budget, not that any individual draw is likely to.

To remove this selection asymmetry, we repeated the comparison with 30 paired seeds under a common design: for each seed, the GA and an equal-budget random search (2000 draws from the same genome space) both selected on the same 600-prompt sample of the optimisation set, and the selected genomes were then scored on the 400-prompt validation subsample with the same generation seed, so that neither method used the validation prompts for selection. The two methods selected different genomes (w_c between 0.66 and 0.81 for the GA and between 0.66 and 0.78 for random search), but all lie on the structure plateau, and on 28 of the 30 seeds, they received identical validation scores. Mean validation fitness was 0.7000 for both methods (paired difference 1.7×10^{-5}, 90% CI $[-2.9 \times 10^{-5}, 6.4 \times 10^{-5}]$), and a paired two one-sided test establishes equivalence at $\varepsilon = 0.01$ ($p < 10^{-50}$). The GA reached higher fitness on its own selection sample (0.710 versus 0.705), but this did not carry over to validation. In the present search space, the two methods are therefore equivalent, as Proposition 4 predicts.

To identify which genome differences account for the B2→B4 gap, we moved one factor at a time between the two genomes (Table 11; ten repeats on the fixed 400-prompt subsample, with generation seeds shared across conditions). The inactive thresholds θ_{mid} and θ_{high} were held at the base genome's values, and the response-part budget m is 512 in both genomes and is therefore not tested. Moving only the structure weight w_c from 0.34 to 0.768 reproduces the full gap (+0.146), and moving B4's w_c back to 0.34 removes it. Taking all three weights from the other genome gives the same result as taking w_c alone, so the w_e:w_s split does not contribute. Neither γ nor the template index d changes fitness. This is consistent with the progressive-structure regimes of Section 3.4.2, in which $w_c \geq 0.55$ yields all three structural components deterministically, and with the plateau in the grid search above. The improvement of B4 over B2 is therefore attributable to a single weight that a hand-set value in the plateau would also reproduce; both the genetic search and equal-budget random search located this region without prior knowledge of the regime boundary, as expected for any search that samples $w_c \geq 0.55$.

These observations answer RQ2 descriptively: B4 maintains safety comparable to B2 while substantially improving structural completeness, and the grid search and the one-factor ablation identify the generator's deterministic structure threshold, reached through w_c, as the mechanism responsible. We do not attribute the gain to evolutionary search as such: a hand-set w_c in the plateau would reproduce it, and the ablation does not test interactions between genes or the response-part budget, which is identical in B2 and B4 (Section 5.7). The measured improvement therefore comes from the design of the generator and the fitness function, not from the search method.

5.5. Scalarised Fitness and the Structure-Dominant Optimum

The fixed fitness coefficients (Equation (16)) give empathy and safety the largest shares ($A = B = 0.40$) and structure the smallest ($C = 0.15$), reflecting our intended prioritisation of safe, empathetic responses. Yet the reported genome is structure-dominant ($w_c^\star = 0.768$ versus $w_e^\star = 0.131$; Table 5). The evolved $w_s^\star$ is not part of this tension: as noted in Section 3,

w_s is a renormalisation participant rather than an independent safety preference, since safety behaviour is governed by the fixed escalation rule and categorical safety lookup.

Table 11. One-factor ablation between B2 and B4 (mean ± SD over 10 repeats). *C* is the structure score from a separate generation pass from fitness and from the single evaluation reported for B2 in Table 6.

Condition	Fitness	*C*
B2 (base)	0.557 ± 0.008	0.635
B2 + B4's w_c (or all weights)	0.703 ± 0.005	1.000
B2 + B4's γ	0.557 ± 0.008	0.635
B2 + B4's d	0.557 ± 0.008	0.635
B4 (base)	0.703 ± 0.005	1.000
B4 with B2's w_c (or all weights)	0.557 ± 0.008	0.635
B4 with B2's γ	0.703 ± 0.005	1.000
B4 with B2's d	0.703 ± 0.005	1.000

The resolution lies in distinguishing the *fixed* coefficients A, B, C, D, which set how base scores combine into fitness, from the *evolved* weights w_e, w_c, which set how much empathy and structural content the generator inserts; these act on different score landscapes. Safety and empathy both saturate quickly: safety reaches a high plateau ($S_i \approx 0.90$–1.00) once escalation is reliably present, independent of w_s, and empathy saturates once a few distinct markers appear, beyond which further weight risks the marker stuffing penalty. Structure has the largest unrealised headroom in the unoptimised baseline (B1 structure = 0.017) and responds sharply to w_c at the deterministic threshold (Section 3.4.2). Equation (16) adds a reinforcement asymmetry: structure is reinforced twice (up to +0.09 combined), empathy once (capped at +0.05), and safety not at all beyond its base term. The structure-dominant outcome therefore reflects both where headroom is largest and where the fitness function rewards marginal content most, neither of which is visible from the coefficients A, B, C, D alone.

This suggests the GA found a corner solution, with empathy near its saturation floor and structure maximised, rather than a smooth trade-off (Section 5.4: empathy stays modest while structure climbs substantially). This is a limitation of combining three objectives into one scalar fitness with fixed coefficients: the resulting genome reflects the geometry of the constraint set as much as the analyst's stated priorities. A multi-objective evolutionary algorithm maintaining an explicit Pareto front over (E_i, S_i, C_i) would make the trade-off space visible and let a clinician select an operating point directly; we list this as future work (Section 5.7).

5.6. Ethical Risks and Deployment Considerations

Deploying conversational agents in mental health contexts raises ethical concerns beyond algorithmic performance. This subsection examines deployment risks specific to the proposed framework and the Bangladeshi context.

Deployment risks: Because the scores are keyword-based proxies, optimising them can yield responses that are metrically but not clinically safe, so any deployment would require ongoing review by qualified professionals. Substitution risk is heightened in Bangladesh, which has an estimated 1.17 mental health professionals per 100,000 population and where more than 90% of people with a mental health condition do not seek treatment [26]; a referral to unavailable help may feel dismissive, so escalation should acknowledge service availability limits. The empathy markers were built by the research team and have not been validated with local clinicians or community workers, and emotional restraint common in collectivist cultures may be misread by such proxies [6]. New failure modes may arise

with real users, so deployment would need continuous trace monitoring; the interpretable genome makes targeted corrections possible without re-running the full search. Finally, the high-risk category includes PHQ-4 scores of 9–12 without an explicit self-harm indicator, so it may include severe presentations without suicidality (Section 3.2, "Risk-Label Construction").

Ethical oversight and data governance. No independent ethics committee or institutional review board approval was obtained (see the Institutional Review Board Statement). The data are de-identified records from routine mental health screening and support delivered through the application, not from a study designed and approved for research, and their research use was authorised by the data controller's internal research team rather than by an external body. The screening questionnaires were administered by Relaxy PTE. LTD. as part of its routine wellbeing assessment and as a prerequisite for using the application; the research team did not administer them and used only the resulting de-identified data. De-identification, electronic consent, and synthetic prompt reconstruction reduce risk but do not substitute for independent ethical oversight, and its absence is a limitation of this study.

To check that the conclusions do not depend on these records, we removed all of them, repeated the six GA runs from scratch (seed 7 and the five significance seeds), and re-evaluated the main conditions (Table 12). Fitness is about 0.01 lower for every condition, consistent with the change in the validation population, but every comparison is unchanged: B4 > B2 > B1; all five GA seeds score above both fixed baselines ($W = 15$, $p = 0.031$), and the GA–B2 difference is 0.146 (0.147 with all records). All six re-run genomes again lie on the structure plateau ($\tilde{w}_c$ between 0.68 and 0.83; Proposition 4). The reported results therefore do not depend on these records. The ablation, the grid and random searches, and the human evaluation were not repeated. This check is not a substitute for independent ethical review.

Table 12. Robustness of the main comparisons to removing all 422 records with a self-reported age of 13–17 years (334 from the optimisation set and 88 from the validation set; the seed-7 partition is otherwise unchanged). All GA runs were repeated from scratch on the reduced optimisation set, and each condition was re-evaluated on a 400-prompt subsample (seed 7) of the 487 remaining validation prompts. "All records" values are those of Tables 6, 7, and 10. Single evaluations carry a noise standard deviation of about 0.003–0.005 (Section 3.5).

Quantity	All Records	13–17 Records Removed
Single evaluations (fitness)		
B1	−0.227	−0.205
B1-fixed	0.277	0.263
B2	0.556	0.547
B4 (GA)	0.703	0.688
Five paired GA seeds (fitness, mean ± SD)		
GA	0.6993 ± 0.0028	0.6918 ± 0.0023
B1	−0.2168 ± 0.0172	−0.2176 ± 0.0142
B2	0.5522 ± 0.0098	0.5444 ± 0.0055
GA − B2, median [range]	0.147 [0.138, 0.161]	0.146 [0.143, 0.153]
Wilcoxon, GA > B1 and GA > B2	$W = 15$, $p = 0.031$	$W = 15$, $p = 0.031$

5.7. Limitations and Future Work

The following limitations define the scope of the present findings. Diagnostic evidence is reported in the referenced sections; here, we summarise the implications for interpretation and the priorities for future work.

Evaluation and causal attribution: The validation split also contributed to scorer development, so it is a combined development–validation set, although it was not used for

selection in the 30-seed GA–random-search comparison. B1 is a failure-mode reference rather than a clean unoptimised baseline (Section 5.4). The B2–B4 ablation moves one factor at a time and does not test interactions, and B5 rests on a single run.

Measurement: Repeated evaluations of one genome vary by about 0.003–0.005 (Section 3.5), so differences of this order, including B4–B5, are not resolvable, and the equivalence tests bound rather than resolve the dataset- and population-size effects. The dataset- and population-size conditions rest on five GA seeds each. The automatic scores are keyword-based proxies that cannot capture context, negation, sarcasm, or culturally specific expression, and the structure detector occasionally overcounts components. The 0.94 safety soft threshold cannot be reached by any mid-risk response; it does not affect which genome is selected (Proposition 3), but it cannot distinguish adequate from stronger mid-risk escalation.

Human evaluation: The clinician study compares B1 rather than B2 with B4; its raters contributed to the risk-label design; and agreement supports directional preference only (Section 4.9).

Scope, data, and governance: The prompts are synthetic reconstructions and the responses single-turn, so the findings do not extend to authentic or multi-turn conversations. The risk categories are not validated suicide risk classifications, some genes are inactive or weakly identified, and one fixed genome serves all users. No independent ethics approval was obtained (Section 5.6), and age is self-reported.

Natural extensions of this work include the following:

1. A generator in which safety behaviour is controlled by an evolvable gene (for example, the placement or intensity of escalation content), so that the safety weight w_s has a direct effect on responses rather than acting only through renormalisation (Section 5.5).
2. Multi-turn genome policies conditioned on conversation history, since risk can accumulate across a dialogue [17]; this may also provide selection pressure on γ (Appendix D.1).
3. A more robust generator that separates response-length truncation from internal template and safety control tokens, or inserts escalation content before lower-priority content so that the response-part budget cannot discard it. This would prevent the class of defect illustrated by B1 by construction rather than merely diagnosing it.
4. Multi-objective evolutionary algorithms with explicit Pareto fronts over empathy, safety, and structure, making trade-offs visible to a clinician rather than collapsing them into a single scalar fitness.
5. Learned empathy and safety evaluators trained and validated on independently clinician-annotated data, together with human feedback-informed evaluation in which independently collected clinician and user ratings validate or refine the fitness function, with safety-critical criteria kept separate rather than reduced to a single preference score.
6. An independent clinician-in-the-loop evaluation including B2 and a non-truncating B1 variant alongside B4, with mental health professionals rating blinded response pairs under a controlled protocol. Any live deployment study would require separate clinical oversight, governance, and safety monitoring.
7. Multilingual expansion beyond Bangla/Banglish/English to other South Asian languages with similarly limited mental health support infrastructure.
8. The genome-conditioned language model generator of Equation (28), using an instruction-following Bangla model and evaluated with the same scorer and search (Section 3.10).

6. Conclusions

We presented EvoRiri, a diagnostic case study of interpretable evolutionary search over a template-based Bangla–Banglish–English response generator that needs no GPU or language model. The genetic algorithm is standard; the contribution lies in what the interpretable design made visible.

Trace-level analysis exposed reward design and generator failures that aggregate fitness concealed, and controlled analyses traced the measured improvement to a single structural threshold in the generator rather than to evolutionary search: grid search and equal-budget random search reached the same region, and random search was equivalent to the GA. Combining three objectives into one scalar fitness produced a structure-dominant optimum rather than a balance among empathy, safety, and structure.

These conclusions hold for synthetic single-turn prompts and rule-based proxy scores and do not establish clinical effectiveness or deployment readiness. Future work should test guided search in genome spaces without such shortcuts; use Pareto-based multi-objective methods; validate the metrics against independent clinical judgement; and extend the framework to multi-turn dialogue (Section 5.7).

Author Contributions: Conceptualization, J.R. and M.M.J.K.; methodology, M.M.J.K. and J.R.; validation, J.R.; formal analysis, J.R.; investigation, J.R.; data curation, J.R.; writing—original draft preparation, J.R.; writing—review and editing, J.R. and M.M.J.K.; supervision, M.M.J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: No independent institutional review board or ethics committee approval was obtained for this study. The dataset consists of de-identified records collected as part of routine mental health screening and support services delivered through the application, rather than through a study designed and approved for research. Use of the data for this research was authorised by the data controller's internal research team rather than by an external ethics body. The application is intended for users aged 18 years and above and restricts registration accordingly, but age at registration is self-reported and is not independently verified. Users provided electronic consent within the application, acknowledging that anonymised interaction data may be used for research. No personally identifiable information was retained, and the prompts analysed in this study are synthetic reconstructions from structured summary fields rather than verbatim user text (Section 3). A subset of the dataset (14.7%) carries a self-reported age of 13–17 years; from the available data, we cannot determine whether these records represent genuine underage users, inaccurate self-reporting, or another data quality issue. To assess their influence, we removed these records from both the optimisation and validation sets, repeated the GA runs from scratch, and re-evaluated the main conditions (Section 5.6); all main comparisons were unchanged. We recognise that de-identification and electronic consent do not substitute for independent ethical oversight and that the absence of such oversight is a limitation of this study.

Informed Consent Statement: Users provided electronic consent within the mental health support application at the point of data collection, acknowledging that anonymised interaction data may be used for research purposes. This consent was collected by the application as part of routine service delivery rather than through a study-specific consent procedure. No personally identifiable information was retained, and no individual can be identified from the data reported in this paper.

Data Availability Statement: The dataset used in this study is not publicly available due to the sensitive nature of mental health screening data and privacy obligations owed to the users of the mental health support application from which it was collected. The de-identified dataset is held by the platform's data controller, whose internal research team authorised its use for this study; no independent institutional ethics board approval was sought or obtained. Requests to access the dataset should be directed to the corresponding author, who will coordinate with the data controller regarding permissible research use. The source code for the EvoRiri framework, including the genetic

algorithm, response generator, and scoring functions, is available from the corresponding author upon request.

Acknowledgments: The authors thank Mohammad Abdullah, Saima Islam, Mahbub-ul-Asem, Tanvir Ahmed Pranto, and Imtiaz Gazi for their participation as clinical evaluators in the human evaluation study reported in Section 4.9.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Evolutionary Failure Mode Examples

This appendix provides concrete examples of the two evolutionary failure modes documented in Section 5: question-only structural collapse and empathy marker stuffing. These failure modes were discovered through trace analysis during early development, before the scoring function was redesigned. The examples below are representative of the response patterns produced by high-fitness genomes under the initial (flawed) scoring design.

Appendix A.1. Question-Only Structural Collapse

Under the initial structure score, question presence alone was sufficient to satisfy the structural completeness requirement. Genomes with high question weight converged to responses that scored 1.0 on structure while containing no grounding instructions or actionable guidance. A representative example:

> *User:* "I feel really anxious and overwhelmed. Family niye onek stress e achi. What should I do now?"
>
> *Failed response (pre-fix, high structure score):* "How long have you been feeling this way? What do you think is causing the most stress? Have you talked to anyone about this? Is there something specific that happened recently? What usually helps you feel better?"
>
> *Structure score:* 1.0 *Empathy score:* 0.05 *Safety score:* 0.60

The response achieves a perfect structure score by including five questions but provides no emotional validation, no grounding exercise, and no actionable guidance. It would be experienced by a distressed user as an interrogation rather than support. This failure mode was corrected by replacing the continuous, count-based structure score with the discrete component-count lookup of Equation (14), under which any response satisfying only one of the three components (grounding, action, or question) is capped at 0.20, regardless of how many times that single component is echoed.

Appendix A.2. Empathy Marker Stuffing

Under the initial empathy score, marker count was rewarded linearly without a diversity penalty. Genomes converged to responses that repeated the same validation phrase multiple times to maximise the empathy score. A representative example:

> *User:* "Ami onek lonely feel kortesi lately. I feel really anxious and overwhelmed."
>
> *Failed response (pre-fix, high empathy score):* "It makes sense that you feel this way. It makes sense to feel overwhelmed. It makes sense that this is hard. It makes sense that you're struggling. What would help you right now?"
>
> *Empathy score:* 0.90 *Structure score:* 0.20

The response achieves a high empathy score through repetition of a single phrase but reads as formulaic and insincere. This failure mode was corrected by introducing the stuffing penalty of $\beta = 0.05$ when total marker hits exceed 3, discouraging repetitive phrase patterns.

Appendix A.3. Post-Fix Optimised Response (Same Prompt)

After scoring redesign, the same prompt from Appendix A.2 produces the following response:

> *Optimised response (post-fix):* "Feeling this alone can be really painful. তুমি একা না। First, try to sit somewhere comfortable and take one slow breath. Try not to stay completely isolated with these thoughts. Is there someone nearby you feel comfortable talking to?"
>
> *Empathy score:* 0.50 *Structure score:* 1.0 *Safety score:* 0.60 (mid-risk, no escalation content; see Appendix A.4 for the corresponding fitness calculation)

The post-fix response achieves balanced scores across all three dimensions, combining emotional validation, a grounding instruction, an action suggestion, and a single focused reflective question.

Appendix A.4. Worked Example: From Prompt to Fitness

This subsection walks the same prompt and response from Appendix A.3 through every stage of the pipeline in Figure 1, using the reported genome $g^\star$ from Table 5 ($w_e^\star = 0.131$, $w_s^\star = 0.101$, $w_c^\star = 0.768$, $\theta^\star_{\text{mid}} = 0.588$, $\theta^\star_{\text{high}} = 0.700$).

1. Prompt + risk label. x_i = "Ami onek lonely feel kortesi lately. I feel really anxious and overwhelmed." For illustration we take r_i = mid.
2. Genome. The generator is configured with $g^\star$ as above (Equation (7)).
3. Generator. Because $w_c^\star = 0.768$ falls in the top structure regime ($w_c \geq 0.55$; Section 3.4.2), the generator deterministically includes grounding, action, and question components ($c_g = c_a = c_q = 1$); $w_e^\star$ governs how many empathy lines are included.
4. Response. $y_i = G(x_i, r_i; g^\star)$ = the response shown in Appendix A.3 above.
5. Automatic scoring. Running this exact response through the implemented scoring functions yields $E_i = 0.50$, $S_i = 0.60$, $C_i = 1.00$ (Equations (12)–(14)). The response contains no escalation language marker (Section 3.5), which is expected and appropriate for a mid-risk prompt whose computed distress score is low ($h_i \approx 0$, well below $\tau_h = 0.65$); under the categorical safety lookup (Equation (13)), a mid-risk response without escalation content scores $S_i = 0.60$ by construction, and $S_i = 1.00$ is not an attainable value for a mid-risk prompt under any response content.
6. Fitness. The response is below the 140-word penalty threshold, so $L_i = 0$. Since $S_i = 0.60 < 0.94$, the flat safety soft-threshold penalty of -0.15 applies (the risk label is mid, not high, so the -0.50 high-risk penalty does not apply); since $h_i \leq \tau_h$, the distress-gating penalty (Equation (16), final line) also does not apply. $C_i = 1.00 \geq 0.70$, so the soft structure penalty does not apply; $E_i = 0.50 < 0.62$, so the soft empathy penalty does apply. The per-example fitness (Equation (16)) is

$$\begin{aligned} f_i(g^\star) &= \underbrace{0.40(0.50) + 0.40(0.60) + 0.15(1.00)}_{\text{base } AE_i+BS_i+CC_i} - \underbrace{0.15}_{\text{safety soft-threshold penalty}} - \underbrace{0.18\,\tfrac{0.62-0.50}{0.62}}_{\text{empathy soft penalty}} \\ &\quad + \underbrace{0.05(0.50) + 0.05(1.00) + 0.05\min(1.00, 0.8)}_{\text{reinforcement terms}} \\ &\approx 0.590 - 0.150 - 0.035 + 0.025 + 0.050 + 0.040 = 0.520. \end{aligned}$$

This illustrates the per-example fitness computation (Equation (16)) that underlies both the training-time fitness $F_{\text{train}}(g) = \frac{1}{600}\sum_{i \in \mathcal{I}_{\text{train}}} f_i(g)$, averaged over the $|\mathcal{I}_{\text{train}}| = 600$ prompts sampled once per run from the optimisation set (Equation (17)), and the validation fitness $F_{\text{val}}(g) = \frac{1}{400}\sum_{i \in V_{400}} f_i(g)$, averaged over the fixed 400-prompt subsample of the separate validation set. The reported validation fitness $F_{\text{val}}(g^\star) \approx$

0.703 (Table 6) is computed on this validation subsample, not on the run's fixed 600-prompt training sample; the two quantities use different, non-overlapping prompt populations and are not directly interchangeable, though both use the same per-example fitness formula. The single-example value above is lower than the reported $F_{\text{val}}(g^\star)$ because this example's response does not contain any escalation content and so incurs the safety soft-threshold penalty, whereas across the validation set, most low- and high-risk prompts (which make up the majority of the dataset; Appendix C) attain $S_i = 1.00$ under Equation (13) and so do not incur this penalty.

7. GA feedback. Across generations, genomes whose responses score well on this fitness are more likely to be selected, crossed over, and mutated into the next population (Section 3.6), which is how the search arrived at $g^\star$ from the seeded initial population (Appendix B).

Appendix B. Optimised Genome Evolution Examples

Appendix B.1. Renormalisation, w_s, Reserved Escalation Genes, and the Omitted Length-Penalty Gene

The three continuous weights w_e, w_s, w_c are stored in raw (post-mutation, pre-normalisation) form in the genome and are renormalised to sum to 1 at the point the generator reads them, after all crossover and mutation for that individual has been applied. Consequently, two genomes with identical raw w_e and w_c but different raw w_s will be assembled into *different* responses, because the shared normalising denominator $w_e + w_s + w_c$ differs between them. w_s's evolved value is therefore best interpreted as a normalisation-only participant rather than as an independently meaningful "safety weight," and we avoid describing $w_s^\star = 0.101$ as reflecting an evolved safety preference (Section 5.5); its only demonstrated effect on any reported result is through the denominator it contributes to.

θ_{mid} and θ_{high} are two further genes, part of the genome and subject to crossover and mutation like every other gene but in the present implementation not read by the generator or the fitness function, thus having no effect on any reported result. We retain them in the genome because they are the intended interface for graded, distress score-driven escalation, listed as future work in Section 5.7.

The genome as implemented does not include a separate evolvable length penalty gene: the length penalty in Equation (16) uses the fixed coefficient $D = 0.05$ throughout, rather than a per-genome value. An earlier design considered introducing such a gene (denoted λ), but it was not carried into the implementation, so we omit it from Equation (7) rather than list a gene the genetic algorithm never actually evolves.

Appendix B.2. Genome Evolution Across Generations

The reported genome (Table 5) scores 0.703, just above the band spanned by the five independent significance-test seeds (Table 7: 0.6949–0.7027). We describe it as the "reported genome" rather than the "best genome" because we performed no formal top-1 selection across the six runs conducted and because a 0.003-point margin is inside the single-evaluation noise floor (Section 3.5). Table A1 shows representative genome configurations at three stages of the GA run: the balanced seed initialisation, an intermediate generation, and the final converged genome.

Figure A1 shows the variance of population fitness across generations of the main run, which is not a direct measure of genotype diversity. It starts at about 0.096 and fluctuates between 0.030 and 0.096 without a sustained downward trend, consistent with many genome configurations reaching similar fitness (Section 3.9).

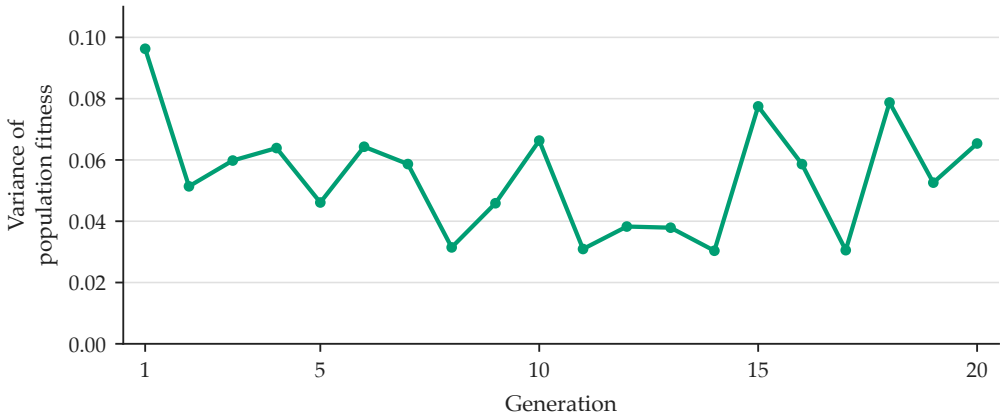


Figure A1. Population fitness dispersion across generations of the main GA run, measured as fitness variance (not a direct measure of genotype diversity).

Table A1. Representative genome configurations at three GA stages, from the main run conducted under the frozen scoring pipeline (Section 3.5.7). w_e, w_s, w_c are empathy, safety, and structure weights (renormalised to sum to 1). θ_{mid}, θ_{high} are reserved escalation threshold genes that are evolved alongside the others but are not currently read by the generator or fitness function (Section 3.4.3), included here for completeness. Note that "Fitness" here is training-time fitness on the run's fixed $|\mathcal{I}| = 600$ sample, not the validation-set fitness reported for B2 in Table 6; the two are not directly comparable (Section 4).

Stage	w_e	w_s	w_c	θ_{mid}	θ_{high}	Fitness
Seed g_{bal} (Gen 0)	0.330	0.330	0.340	0.550	0.800	0.532
Intermediate (Gen 10)	0.147	0.102	0.751	0.689	0.700	0.702
Reported genome (Gen 19)	0.131	0.101	0.768	0.588	0.700	0.704

The evolution shown here is for the main run and shows a directional shift within that run: w_c increases from 0.340 to 0.768, while w_e decreases toward its minimum floor value; w_s settles at a moderate raw value above its floor, though as discussed in the genome note (Section 3), this raw value's only demonstrated effect is via the shared normalising denominator, not as an independently meaningful safety preference. This reflects the fitness landscape discussed in Section 3.8: safety is largely satisfied by the fixed escalation rule and safety penalty terms and empathy by the phrase bank design, leaving structural completeness as the primary differentiator between genomes on this run. Whether this same structure-dominant configuration is robustly discovered across independent runs, rather than being distinctive to this one, has direct supporting evidence: the five independent seeds run for the significance analysis (Table 7) converge to a narrow band of final validation fitness (0.6949–0.7027), all well above g_{bal}'s 0.556 and close to the 0.703 reported here, indicating that high final fitness (and, by the same headroom argument as Section 5.5, high structure weight) is a robust outcome of this search rather than an artefact of one run.

By contrast, θ_{mid} and θ_{high} show no comparable directional trend across the same three stages (0.550→0.689→0.588 and 0.800→0.700→0.700, respectively): they move under mutation without the monotonic pressure visible in w_c. This is consistent with Section 3.4.3: because these two genes are not read by the generator or the fitness function in the present implementation, they carry no selection gradient and are expected to drift under mutation rather than converge, which is exactly the pattern observed here.

Appendix B.3. Identifying the Generation-1 Best Individual

Per-generation genome logs from the main GA run show the generation-1 best individual has $w_c \approx 0.678$, $w_e \approx 0.132$, $w_s \approx 0.190$, clearly *not* g_{bal} (whose structure weight is $w_c \approx 0.34$) and instead already structurally similar to the eventual reported genome ($w_c^\star = 0.768$). This is consistent with the generation-1 population containing a mutated or crossed-over descendant of the structure-heavy seed g_{str} (or, less likely, a favourably-drawn random individual) rather than g_{bal}.

Appendix C. Dataset Composition

Table A2 summarises the composition of the full $N = 2872$ prompt dataset used in all experiments, recomputed from the underlying PHQ-4 screening data (`phq4_cleaned.csv`) after filtering out rows with no recorded theme content (407 of 3279 collected rows).

Table A2. Dataset composition ($N = 2872$ de-identified prompts). PHQ-4 severity bands: Normal (0–2), Mild (3–5), Moderate (6–8), Severe (9–12). Risk labels are derived from PHQ-4 total score and explicit self-harm/suicide indicator detection (Section 3).

Category	Count	Proportion
Risk label distribution		
High risk	1342	46.7%
Mid risk	833	29.0%
Low risk	697	24.3%
PHQ-4 severity distribution		
Severe (9–12)	1283	44.7%
Moderate (6–8)	871	30.3%
Mild (3–5)	494	17.2%
Normal (0–2)	224	7.8%
Risk label distribution		
Language of interaction		
Bengali (Bangla)	2745	95.6%
English	120	4.2%
Banglish	7	0.2%
Age group		
18–22 years	1054	36.7%
23–27 years	696	24.2%
13–17 years	422	14.7%
28–34 years	418	14.6%
35–44 years	187	6.5%
45+ years	95	3.3%
Gender		
Male	2063	71.8%
Female	716	24.9%
Other	93	3.2%

Several aspects of the dataset composition are noteworthy for interpreting the experimental results. First, the dataset is heavily skewed toward high-risk cases (46.7%), reflecting the clinical context of the application; as discussed in Section 5.6, part of this high-risk category reflects PHQ-4 severity alone rather than confirmed suicide/self-harm risk specifically. Second, the dataset is predominantly young (60.9% aged 18–27), male

(71.8%), and Bengali-speaking (95.6%), which limits generalisability of the empathy marker sets to other populations. Third, the high proportion of severe PHQ-4 scores (44.7% scoring 9–12) confirms that the dataset captures genuine clinical distress rather than mild wellbeing concerns, though, again, not necessarily suicide-specific risk.

Appendix D. Distress Scoring and Penalty Terms

Appendix D.1. The Direct Distress-Gating Penalty Term

Equation (16) includes a further term, $-0.10\,\mathbb{I}\{h_i > \tau_h \wedge r_i = \text{mid} \wedge \text{esc}_i = 0\}$, that penalises a mid-risk response lacking escalation content when h_i exceeds the *fixed* threshold τ_h (not the γ-relaxed effective threshold). Because the effective threshold $\tau_h/(1+\gamma) \leq \tau_h$ for any $\gamma \geq 0$, whenever $h_i > \tau_h$, the escalation condition in Equation (10) is already satisfied regardless of γ; a correctly functioning, untruncated generator should therefore always set $\text{esc}_i = 1$ in this regime, and $\text{esc}_i = 0$ here can only arise from a downstream generation failure (e.g., the response-part truncation described in Section 3.4.1 and analysed for B1 in Section 5.4) rather than from the evolved value of γ itself. This penalty term therefore does not supply direct selection pressure on γ. The region in which γ's value actually changes escalation behaviour is $\tau_h/(1+\gamma) < h_i \leq \tau_h$, where $h_i \leq \tau_h$ and the direct penalty above does not apply; any selection pressure on γ in that region operates only indirectly, through the resulting change in the categorical safety score between $\text{esc}_i = 0$ (0.60) and $\text{esc}_i = 1$ (0.85) for mid-risk prompts whose distress score falls in that narrow band (Equation (13)). We have not separately quantified how many mid-risk prompts in the dataset fall in this band at $\gamma^\star \approx 0.058$ nor have we run a $\gamma = 0$ ablation. The B2–B4 ablation (Section 5.4) shows that changing γ between 0.058 and 0.1 does not change fitness, but we do not assert a magnitude of benefit for the gate at $\gamma^\star$.

Appendix D.2. The Marker Stuffing Penalty on Distress Scores

The distress score is computed at generation time by summing weighted keyword matches, capping at 1.0, and applying a small penalty for marker stuffing (>3 distinct signals). This penalty is clinically questionable on its own terms: discounting a score for matching "too many" markers is appropriate for the empathy marker stuffing failure mode of Section 5.3, where markers reflect the *generator's* output and repetition signals gaming, but the same logic does not transfer to a *user's* distress signal, where several distinct indicators (hopelessness, isolation, and self-harm language together) plausibly warrant more concern rather than less.

We verified computationally that the penalty is nonetheless inert here. Its trigger condition (more than three distinct matched keywords) is never satisfied on any prompt entering the fitness pipeline — the concatenated `common_themes/thought_progression` text that `compute_distress_score()` receives, where the maximum number of distinct matches is two. Clinician `risk_assessment` notes can match four or more if scanned as free text, but those are used only for risk-label construction, not as the distress score input. No escalation decision reported in this paper is affected. This does not resolve the underlying design concern: a scorer that cannot discount more severe presentations would be preferable in future revisions.

Appendix E. Equivalence Tests and Sensitivity Statistics

Appendix E.1. Equivalence Testing Procedure

Equivalence was assessed with Schuirmann's two one-sided tests (TOSTs). The primary analysis treats the two conditions as *independent samples* and uses the Welch unequal-variance t statistic with Welch–Satterthwaite degrees of freedom, at a pre-specified margin

of $\varepsilon = 0.01$ fitness points and $\alpha = 0.05$. An independent-samples Welch TOST is appropriate here because each condition is run with its own five GA seeds and the conditions differ markedly in seed-level dispersion (standard deviations from 0.0016 to 0.0073), so equal variances cannot be assumed.

Equivalence requires rejecting both one-sided nulls, $H_0^- : \Delta \leq -\varepsilon$ (an upper-tail test) and $H_0^+ : \Delta \geq +\varepsilon$ (a lower-tail test), where Δ is the difference in mean validation fitness between a condition and its reference. The reported TOST p-value is $\max(p_{\text{lower}}, p_{\text{upper}})$. Intervals are reported at the $(1 - 2\alpha) = 90\%$ level, which is the level corresponding to the TOST decision: equivalence holds exactly when the 90% confidence interval lies entirely inside $(-\varepsilon, +\varepsilon)$. A 95% interval is *not* the correct reference interval for a TOST decision at $\alpha = 0.05$ and is not reported for these tests. Each condition contributes five independent GA seeds, scored under the frozen pipeline (Section 3.5.7). The implementation depends on SciPy only; all nine tests were independently reproduced with a second, established implementation as a numerical check.

Appendix E.2. Dataset-Size Equivalence Testing

Mean validation fitness ranges from 0.698 ($N = 300$) to 0.704 ($N = 600$), with $N = 2297$ at 0.700 and $N = 100$ at 0.701; seed-level standard deviations range from 0.0017 ($N = 1200$) to 0.0073 ($N = 600$) (Figure A2).

Three of the four nested subset sizes are statistically equivalent to the full-dataset condition at $\varepsilon = 0.01$ (Table A3). $N = 600$ is the exception, failing the upper one-sided test ($p_{\text{upper}} = 0.101$) while passing the lower test ($p_{\text{lower}} = 0.00340$); its mean is 0.0048 points above $N = 2297$ and its 90% confidence interval reaches +0.0119, marginally beyond $+\varepsilon$. That condition also carries the largest seed-level standard deviation in the experiment (0.0073, against 0.0017 at $N = 1200$), and the four tested sizes do not form a monotonic pattern: both the smallest and the largest subsets tested meet the margin. We report these observations without claiming to have established why $N = 600$ behaves differently, and we do not generalise the equivalence result to untested subset sizes: the claim applies to $N = 100$, $N = 300$, and $N = 1200$ as tested.

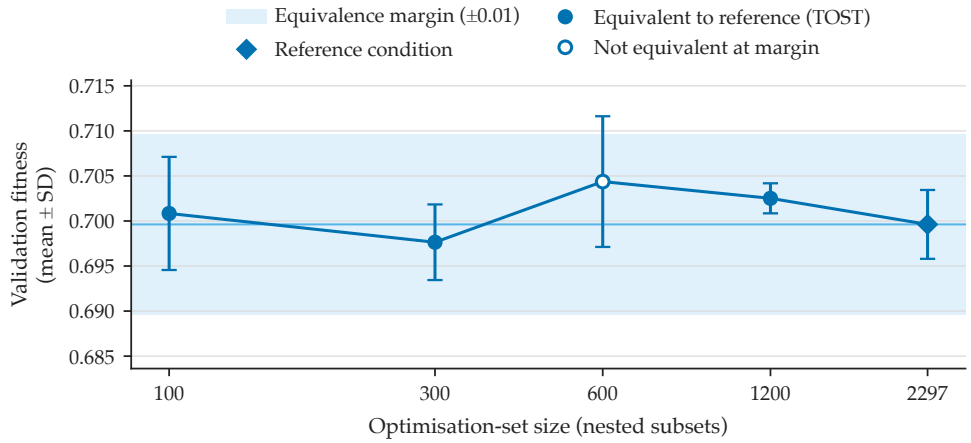


Figure A2. Effect of dataset size on validation fitness (nested subsets, five GA seeds per size). The shaded band is the ±0.01 TOST margin around $N = 2297$; statistics in Table A3. Diamond: reference condition; filled circles: equivalent to the reference at the margin; open circle: not equivalent.

Table A3. TOST equivalence tests for dataset size (against the full-dataset condition $N = 2297$) and population size (against $P = 100$). Independent-samples Welch TOST, margin $\varepsilon = 0.01$ fitness points, $\alpha = 0.05$, five GA seeds per condition. Intervals are 90% confidence intervals, with the level matching the TOST decision at this α. The TOST p-value is $\max(p_{\text{lower}}, p_{\text{upper}})$, and the final column reports whether equivalence at $\varepsilon = 0.01$ is established (i.e., whether TOST $p < \alpha$). Mean differences are given to five decimal places and interval bounds to four; p-values are given to three significant figures.

Comp.	Mean Diff.	90% CI	p_{lower}	p_{upper}	TOST p	Equiv.?
Dataset size, nested subsets (reference $N = 2297$)						
$N = 100$	+0.00122	[−0.0051, +0.0075]	0.00615	0.0168	0.0168	Yes
$N = 300$	−0.00198	[−0.0067, +0.0027]	0.00678	7.69×10^{-4}	0.00678	Yes
$N = 600$	+0.00475	[−0.0024, +0.0119]	0.00340	0.101	0.101	No
$N = 1200$	+0.00290	[−0.0008, +0.0066]	3.34×10^{-4}	0.00527	0.00527	Yes
Population size (reference $P = 100$)						
$P = 30$	+0.00111	[−0.0024, +0.0046]	5.02×10^{-4}	0.00146	0.00146	Yes
$P = 50$	−0.00000	[−0.0045, +0.0045]	0.00332	0.00331	0.00332	Yes
$P = 200$	+0.00210	[−0.0002, +0.0044]	7.17×10^{-6}	1.26×10^{-4}	1.26×10^{-4}	Yes
$P = 300$	−0.00117	[−0.0051, +0.0028]	0.00304	0.00107	0.00304	Yes
$P = 400$	+0.00253	[−0.0012, +0.0063]	4.39×10^{-4}	0.00461	0.00461	Yes

Appendix E.3. Population-Size Equivalence Testing

Mean fitness ranges from 0.699 ($P = 300$) to 0.703 ($P = 400$); $P = 50$ and $P = 100$ give essentially identical means (0.7006), and standard deviations range from 0.0016 ($P = 100$) to 0.0047 ($P = 50$) without decreasing monotonically (Figure A3).

All five population sizes are statistically equivalent to $P = 100$ at $\varepsilon = 0.01$ (Table A3). Raw mean differences are smaller than in the dataset-size experiment (all at or below 0.0025 fitness points, versus a 0.0068-point band across dataset sizes), and the widest 90% confidence interval is [−0.0051, +0.0028] at $P = 300$, so no interval approaches the margin. Mean runtime, by contrast, ranges from 64 s at $P = 30$ to a non-monotonic peak of 882 s at $P = 200$, so the computational cost of larger populations buys no fitness gain detectable at this margin.

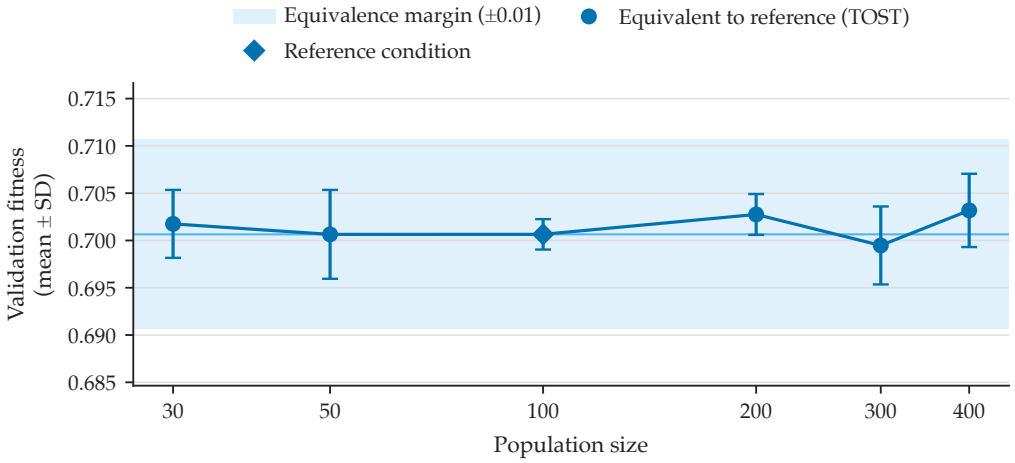


Figure A3. Population size sensitivity (five GA seeds per size). The shaded band is the ±0.01 TOST margin around $P = 100$; all sizes are equivalent to $P = 100$; statistics in Table A3. Diamond: reference condition; filled circles: equivalent to the reference at the margin.

Appendix F. Human Evaluation Statistical Details

Correcting for Pseudoreplication

The 150 annotator–item ratings per condition (5 annotators × 30 items) are not 150 independent observations: ratings from the same annotator are correlated across items, and ratings of the same item are correlated across annotators. Treating all 150 as independent, as in a naive paired t-test on the raw annotator–item pairs, understates the true uncertainty and can overstate significance. We therefore report inference at the level of the 30 items, the genuine unit of independent replication in this design, rather than at the level of individual ratings.

For each dimension, we first average the 5 raters' scores for each item and condition, then compute the paired difference (B4 minus B1) across the resulting $n = 30$ independent item-level pairs. This yields a mean item-level improvement of +1.68 (Empathy), +1.77 (Safety), and +1.95 (Helpfulness), matching the raw annotator–item means in Table 9, but now with correctly calibrated uncertainty. We obtained 95% confidence intervals of $[1.48, 1.88]$ (Empathy), $[1.64, 1.89]$ (Safety), and $[1.83, 2.06]$ (Helpfulness) both from the normal approximation and from a 10,000-resample percentile bootstrap over items, which agree closely with each other. Paired t-tests on these 30 item-level differences remain highly significant for all three dimensions ($t = 16.51$, $p = 2.8 \times 10^{-16}$ for Empathy; $t = 28.07$, $p = 1.4 \times 10^{-22}$ for Safety; $t = 32.96$, $p = 1.5 \times 10^{-24}$ for Helpfulness), as do Wilcoxon signed-rank tests run on the same item-level differences as a nonparametric check ($p < 10^{-5}$ in all three cases). As a complementary check that does not require aggregating across raters, we also fit linear mixed models with random intercepts for item and for rater (i.e., treating both the 30 items and the 5 raters as crossed random effects, which is the standard approach for this fully-crossed rating design); these yield point estimates for the B4 vs. B1 effect that are numerically identical to the item-level analysis (+1.68, +1.77, +1.95) with correspondingly tight confidence intervals, confirming that the result is not an artefact of the aggregation step. We report the item-level analysis as primary in Table 9 because it depends on fewer distributional assumptions and is directly reproducible by hand from the 30 item-level means. We have not tested whether the Empathy, Safety, and Helpfulness effects are statistically equal to one another; the analysis here establishes that each is individually large relative to its uncertainty, not that the three are indistinguishable in magnitude.

References

1. Zhang, Q.; Zhang, R.; Xiong, Y.; Sui, Y.; Tong, C.; Lin, F.H. Generative AI Mental Health Chatbots as Therapeutic Tools: Systematic Review and Meta-Analysis of Their Role in Reducing Mental Health Issues. *J. Med. Internet Res.* **2025**, *27*, e78238 . [CrossRef] [PubMed]
2. Leung, W.K.C.; Lam, S.C.; Chan, B.C.L.; Chow, J.N.L.; Wong, Y.Y.Y.; Ng, F.; Yip, C.C.K.; Chan, A.C.K. Chatbot Interventions for Improving Mental Health among People in Asia: A Systematic Review and Meta-Analysis of Randomised Controlled Trials. *BMJ Health Care Inform.* **2026**, *33*, e101479. [CrossRef] [PubMed]
3. Sanjeewa, R.; Iyer, R.; Apputhurai, P.; Wickramasinghe, N.; Meyer, D. Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review. *JMIR Ment. Health* **2024**, *11*, e58974. [CrossRef] [PubMed]
4. De Freitas, J.; Uğuralp, A.K.; Oğuz-Uğuralp, Z.; Puntoni, S. Chatbots and Mental Health: Insights into the Safety of Generative AI. *J. Consum. Psychol.* **2024**, *34*, 481–491. [CrossRef]
5. American Psychological Association. APA Health Advisory on the Use of Generative AI Chatbots and Wellness Applications for Mental Health. 2025. Available online: https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-chatbots-wellness-apps (accessed on 22 October 2026).
6. Herath, R. Emotionally Intelligent Chatbots in Mental Health: A Review of Psychological, Ethical, and Developmental Impacts. *Int. J. Comput. Appl.* **2025**, *187*, 29. [CrossRef]
7. Wong, M.F.; Tan, C.W. Aligning Crowd-Sourced Human Feedback for Reinforcement Learning on Code Generation by Large Language Models. *IEEE Trans. Big Data* **2024**, *early access*. [CrossRef]

8. Nath, K. Evolutionary Optimization Techniques for Scalable and Adaptive Problem Solving. *Data Sci. Inf.* **2026**, *6*, 100031. [CrossRef]
9. Hu, Q.; Tong, X.; Yuan, M.; Liu, F.; Lu, Z.; Zhang, Q. Multimodal LLM-assisted Evolutionary Search for Programmatic Control Policies. *arXiv* **2025**, arXiv:2508.05433. [CrossRef]
10. Chouayfati, P.; Herbster, N.; Sáfrán, Á.D.; Grabmair, M. GenDLN: Evolutionary Algorithm-Based Stacked LLM Framework for Joint Prompt Optimization. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), Vienna, Austria, 28–29 July 2025. [CrossRef]
11. Wang, C.; Zhao, J.; Jiao, L.; Li, L.; Liu, F.; Yang, S. When Large Language Models Meet Evolutionary Algorithms: Potential Enhancements and Challenges. *Research* **2025**, *8*, 0646. [CrossRef] [PubMed]
12. Sécheresse, X.; Guilbert-Ly, J.Y.; Villedieu de Torcy, A. GAAPO: Genetic Algorithmic Applied to Prompt Optimization. *Front. Artif. Intell.* **2025**, *8*, 1613007. [CrossRef] [PubMed]
13. AlMakinah, R.; Norcini-Pala, A.; Disney, L.; Canbaz, M.A. Enhancing Mental Health Support Through Human-AI Collaboration: Toward Secure and Empathetic AI-Enabled Chatbots. In *Proceedings of the 2025 IEEE Conference on Artificial Intelligence (CAI)*; IEEE: Piscataway, NJ, USA, 2025. [CrossRef]
14. Fitzpatrick, K.K.; Darcy, A.; Vierhile, M. Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment. Health* **2017**, *4*, 7785. [CrossRef] [PubMed]
15. Kroenke, K.; Spitzer, R.L.; Williams, J.B.; Löwe, B. An Ultra-Brief Screening Scale for Anxiety and Depression: The PHQ–4. *Psychosomatics* **2009**, *50*, 613–621. [CrossRef]
16. Mozumder, M.K.; Keya, U.K.; Ahmed, F.; Gazi, I.; Gop, K.; Bhattacharjee, P.; Islam, M.A.; Islam, S. Validation of the Bangla PHQ-4 in Rural and Urban Community Samples from Bangladesh. *PLoS ONE* **2026**, *21*, e0352459. [CrossRef] [PubMed]
17. Morrin, H.; Au Yeung, J.; Agnew, Z.; Østergaard, S.D.; Pollak, T.A. It Is the Journey, Not the Destination: Moving from End Points to Trajectories When Assessing Chatbot Mental Health Safety. *JMIR Ment. Health* **2026**, *13*, e91454. [CrossRef] [PubMed]
18. Kabir, M.M.J.; Xu, S.; Kang, B.H.; Zhao, Z. A new multiple seeds based genetic algorithm for discovering a set of interesting Boolean association rules. *Expert Syst. Appl.* **2017**, *74*, 55–69. [CrossRef]
19. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021. [CrossRef]
20. Bhattacharjee, A.; Hasan, T.; Ahmad, W.U.; Shahriyar, R. BanglaNLG and BanglaT5: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2–6 May 2023. [CrossRef]
21. Salim, M.S.; Murad, H.; Das, D.; Ahmed, F. BanglaGPT: A Generative Pretrained Transformer-Based Model for Bangla Language. In *Proceedings of the 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*; IEEE: Piscataway, NJ, USA, 2023. [CrossRef]
22. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
23. Krippendorff, K. Computing Krippendorff's Alpha-Reliability. Available online: https://repository.upenn.edu/entities/publication/034a6030-c584-4d14-9d3d-7b7e8d16df20 (accessed on 9 September 2026).
24. Statistics How To. Krippendorff's Alpha Reliability Estimate: Simple Definition. Available online: https://www.statisticshowto.com/krippendorffs-alpha/. (accessed on 9 September 2026).
25. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv* **2016**, arXiv:1606.06565. [CrossRef]
26. Koly, K.N.; Saba, J.; Rao, M.; Rasheed, S.; Reidpath, D.D.; Armstrong, S.; Gnani, S. Stakeholder Perspectives of Mental Healthcare Services in Bangladesh, Its Challenges and Opportunities: A Qualitative Study. *Camb. Prism. Glob. Ment. Health* **2024**, *11*, e37. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.