

Multi-Objective Orchestration of Small Language Model Ensembles: Balancing Accuracy, Diversity, and Fairness

COIN Report Number 2026012

Advay Dhar
Jadavpur University
Kolkata, West Bengal, India
advay.dhar1@gmail.com

Swagatam Das
Indian Statistical Institute
Kolkata, West Bengal, India
swagatam.das@isical.ac.in

Kalyanmoy Deb
Michigan State University
East Lansing, Michigan, USA
kdeb@egr.msu.edu

Abstract

Small Language Models (SLMs) offer efficient and practical alternatives to large-scale models in resource-constrained environments. We present a principled framework for constructing SLM ensembles that jointly optimize three competing objectives: prediction accuracy, output diversity, and fairness. Our method combines an interpretable cost-function formulation with a multi-objective evolutionary algorithm to discover Pareto-optimal ensemble configurations. We further introduce a two-stage combiner that produces diverse candidate responses and selects final outputs via embedding-based semantic consensus. Experiments on the MentalChat16k mental-health dialogue dataset show that the best-performing SLM ensemble configurations can match or surpass a fine-tuned Llama 3.1 70B model, achieving improvements of 0.86% in ROUGE-1, 5.84% in ROUGE-2, 4.93% in ROUGE-L, and 7.01% in semantic similarity. These results indicate that strategically orchestrated ensembles of small models can offer competitive or superior performance to significantly larger LLMs, while providing greater flexibility, interpretability, and accessibility for researchers operating under limited resources.

Keywords

Language Models, Ensemble Methods, Multi-Objective Optimization, Fairness, NSGA-III

1 Introduction

Large language models (LLMs) have rapidly advanced the state of the art in natural language generation, yet their practical deployment remains constrained by heavy computational demands, latency limitations, and hardware scarcity. Small Language Models (SLMs)—typically only a few billion parameters—offer a promising alternative: they are lighter, more deployable, and increasingly competitive with their larger counterparts [31, 37]. These properties make SLMs especially appealing for safety-critical domains such as mental-health dialogue generation, where responsiveness, resource efficiency, and robust behavioral guarantees must all coexist.

However, chatbots in sensitive domains, such as mental-health assistants impose unusually stringent and multi-faceted requirements. Beyond textual correctness, such systems must exhibit empathy, maintain stylistic and semantic diversity, and avoid unfair or harmful outputs. Recent benchmarks show that current LLMs—regardless of scale—struggle to achieve uniformly strong performance across emotional nuance, demographic fairness, and stability under sensitive scenarios [11, 14]. This fragmentation

strongly suggests that no single model can be expected to dominate all such dimensions, and that multi-objective approaches are essential for principled progress.

Ensemble methods are a natural candidate, yet language-model ensembles differ fundamentally from traditional predictive ensembles. They must reconcile heterogeneous behaviors, not merely average numerical predictions. Recent Mixture-of-Agents frameworks highlight that aggregating specialized models can significantly enhance reasoning and control [39]. Complementing this, evolutionary optimization has emerged as a powerful mechanism for navigating conflicting LLM objectives. Evolutionary prompt optimization [6] demonstrates that evolutionary search can outperform manual engineering [13], while multi-objective evolutionary algorithms (MOEAs) have been shown to effectively balance sentiment, tone, and content attributes in generated text [1]. These developments converge toward an opportunity uniquely suited for the evolutionary computation community: systematically orchestrating heterogeneous SLMs under competing behavioral objectives.

Our work is motivated by three core insights. First, individual SLMs can be fine-tuned to excel at specific competencies—accuracy, empathy, stylistic richness—but often at the expense of others due to limited capacity. Second, sensitive subjects such as mental-health applications themselves exhibit shifting objective priorities: crisis-support interactions demand safety and precision, whereas reflective conversations benefit from diversity and emotional subtlety. Third, while prior ensemble and routing methods can optimize for a dominant objective, they lack the mechanisms to explicitly characterize and exploit Pareto trade-offs across multiple behavioral dimensions.

To address these gaps, we propose a multi-objective evolutionary framework for constructing SLM ensembles. Our method evolves weight configurations over a pool of heterogeneous SLMs, discovering Pareto-optimal ensembles that provide different trade-offs among accuracy, empathy, diversity, and fairness. The resulting ensembles demonstrate that strategic specialization and evolutionary orchestration of small models can match—and in several cases surpass—the performance of much larger models on mental-health dialogue tasks. This provides compelling evidence that evolutionary multi-objective optimization is not merely a tuning tool, but a core enabler of resource-aware, safe, and context-sensitive conversational AI.

Our contributions are threefold:

- We formulate SLM ensemble design as a multi-objective optimization problem, defining interpretable cost functions that encode accuracy, diversity, and fairness.

- We apply NSGA-III to uncover Pareto-optimal ensemble configurations, revealing structural trade-offs that govern mental-health response generation.
- We introduce a two-stage combiner mechanism that (i) generates diverse candidates from heterogeneous SLMs and (ii) selects outputs via embedding-based consensus, enabling controlled navigation across the Pareto front.

Empirically, our evolutionary ensembles outperform a fine-tuned 70B-parameter baseline under ROUGE and other semantic similarity metrics, demonstrating that coordinated specialization of small models can rival or exceed the capabilities of monolithic systems.

The remainder of the paper is organized as follows. Section 2 reviews relevant literature. Section 3 formulates our multi-objective ensemble-construction problem, and Section 4 details our evolutionary framework. Experimental settings and results are presented in Sections 5 and 6, respectively. Section 7 discusses implications, and Section 8 concludes the study.

2 Related Work

2.1 Ensemble Methods for Language Models

Ensemble techniques have been extensively studied in machine learning, with applications ranging from random forests to boosting algorithms [5, 34]. In the context of language models, recent work has explored various ensemble strategies. Mixture of Experts (MoE) architectures [10, 35, 39] learn to route inputs to specialized sub-networks, while model averaging techniques combine predictions from multiple models [17]. However, most existing approaches focus primarily on predictive performance [21] without explicitly considering diversity or fairness objectives.

2.2 Multi-Objective Optimization in ML

Multi-objective optimization has been applied to various machine learning tasks, including neural architecture search [27], hyperparameter tuning [18], and fairness-aware learning [29]. Evolutionary algorithms, particularly NSGA-II [8] and NSGA-III [7], have proven effective for discovering Pareto-optimal solutions in high-dimensional objective spaces. Our work extends these techniques to the ensemble construction problem for language models.

2.3 Fairness in Language Models

Recent research has highlighted concerning biases in large language models, including racial, gender, and socioeconomic biases [2, 4]. Various debiasing techniques have been proposed, from data filtering to adversarial training [41]. Our approach incorporates fairness as an explicit optimization objective, using toxicity detection as a proxy for bias measurement.

3 Problem Formulation

3.1 Ensemble Setup

Let $\mathcal{M} = \{M_1, M_2, \dots, M_N\}$ denote a set of N pretrained small language models. For a given input \mathbf{x} , each model M_i generates a complete text response $r_i = M_i(\mathbf{x})$ through autoregressive decoding.

Rather than mixing token-level probability distributions, our ensemble operates in the semantic embedding space. Each generated response r_i is embedded using a sentence encoder $\text{Embed}(\cdot)$ to produce a dense vector representation. The ensemble representation is then computed as a weighted combination of individual model embeddings:

$$\mathbf{e}_{\text{ensemble}}(\mathbf{x}) = \sum_{i=1}^N \alpha_i \cdot \text{Embed}(r_i), \quad (1)$$

where $\alpha_i \in [0, 1]$ represents the weight assigned to model M_i , subject to the simplex constraint $\sum_{i=1}^N \alpha_i = 1$. We use the SentenceBERT [32] all-MiniLM-L6-v2 model for computing embeddings.

3.2 Cost Function Formulation

We define the total cost function as a weighted combination of three objective terms:

$$\mathbf{L}_{\text{total}} = \lambda_1 \mathbf{L}_{\text{pred}} + \lambda_2 \mathbf{L}_{\text{div}} + \lambda_3 \mathbf{L}_{\text{fairness}}, \quad (2)$$

where $\lambda_1, \lambda_2, \lambda_3 \geq 0$ (at least one $\lambda_i > 0$) are hyperparameters controlling the relative importance of each objective.

Table 1: Intuitive interpretation of the three cost function components.

Component	Interpretation
\mathbf{L}_{pred}	Measures faithfulness to reference responses; encourages accurate predictions aligned with ground truth
\mathbf{L}_{div}	Captures disagreement in output distributions; penalizes homogeneous ensembles that lack complementary perspectives
$\mathbf{L}_{\text{fairness}}$	Quantifies toxicity exposure; penalizes generation of potentially harmful or biased content

3.2.1 Prediction Loss. For each model M_j , we compute its per-token cross-entropy on the validation set $\mathcal{D}_{\text{val}} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{|\mathcal{D}_{\text{val}}|}$:

$$\text{CE}(M_j, \mathcal{D}_{\text{val}}) = -\frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{n=1}^{|\mathcal{D}_{\text{val}}|} \sum_{v \in \mathcal{V}} y_{n,v} \log M_j(\mathbf{x}_n)_v, \quad (3)$$

where $M_j(\mathbf{x}_n)_v$ denotes the probability assigned by model M_j to token v at input \mathbf{x}_n , and $y_{n,v}$ is the reference distribution.

The ensemble prediction loss is then the weighted average of individual model cross-entropies:

$$\mathbf{L}_{\text{pred}} = \sum_{j=1}^N \alpha_j \cdot \text{CE}(M_j, \mathcal{D}_{\text{val}}). \quad (4)$$

This term encourages the ensemble to produce accurate predictions aligned with ground truth responses.

3.2.2 Diversity Penalty. To encourage heterogeneity in model outputs, we employ Jensen-Shannon (JS) divergence:

$$\mathbf{L}_{\text{div}} = -\sum_{i=1}^N \sum_{j=i+1}^N \alpha_i \alpha_j \cdot \text{JS}(M_i(\mathbf{x}), M_j(\mathbf{x})), \quad (5)$$

where the Jensen-Shannon divergence between two distributions P and Q is defined as:

$$JS(P, Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M), \quad (6)$$

with $M = \frac{1}{2}(P + Q)$ and KL denoting the Kullback-Leibler divergence (the symbol $||$ denotes divergence measured relative to). In (5), normalization by the number of model pairs is omitted to retain a direct sum of pairwise JS contributions, ensuring that diversity scales with both weight distribution and ensemble composition. The negative sign converts the diversity measure into a penalty term to be minimized. Higher diversity (more disagreement between models) leads to a lower loss value, encouraging the optimizer to select models with complementary perspectives [15, 25].

We employ Jensen-Shannon divergence rather than alternatives, like KL divergence because JS divergence is symmetric ($JS(P, Q) = JS(Q, P)$), bounded between 0 and 1, and remains well-defined even when distributions have non-overlapping support. These properties make it particularly suitable for comparing language model output distributions, where asymmetric measures could artificially bias the ensemble toward specific models.

3.2.3 Toxicity-Based Fairness Proxy Penalty. We assess fairness using the Detoxify library [16] with probing statements designed to elicit potentially biased responses. Let $\mathcal{P} = \{p_1, p_2, \dots, p_K\}$ be a set of K probing statements covering various demographic dimensions. The fairness loss is:

$$L_{\text{fairness}} = \frac{1}{K} \sum_{k=1}^K \sum_{j=1}^N \alpha_j \cdot \text{Detoxify}(M_j(p_k)), \quad (7)$$

where $\text{Detoxify}(M_j(p_k)) \in [0, 1]$ returns a toxicity score for model M_j 's response to probing statement p_k . This term penalizes ensembles that generate toxic or biased content.

We emphasize that toxicity detection serves as a *proxy* for fairness and is not a comprehensive measure. This approach captures explicit harmful content but does not address:

- **Representational harms:** Stereotyping or misrepresentation that may not trigger toxicity classifiers,
- **Allocative fairness:** Disparate impact on different demographic groups in task performance,
- **Subtle biases:** Implicit associations that manifest in word choice or framing without overt toxicity,
- **Contextual appropriateness:** Domain-specific fairness requirements beyond general toxicity.

The Detoxify-based penalty provides a computationally tractable optimization signal during multi-objective search, enabling systematic exploration of fairness trade-offs. However, comprehensive fairness evaluation requires multiple complementary metrics including demographic parity, equalized odds, and counterfactual fairness assessments. Our framework is agnostic to the specific fairness metric used—alternative formulations (e.g., demographic-specific evaluation, bias score aggregation, or fairness constraints) can replace Equation (6) without modifying the optimization methodology.

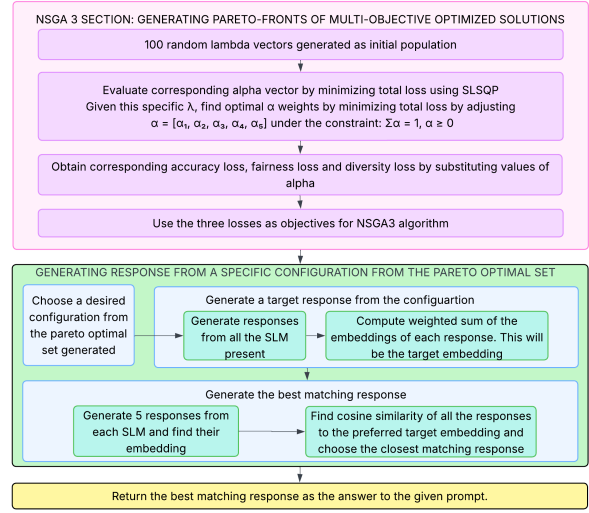


Figure 1: Complete process flowchart showing the multi-objective optimization framework for SLM ensemble construction.

4 Methodology

We construct SLM ensembles using a principled bi-level optimization framework. The upper level controls the relative weighting of three behavioral objectives—prediction accuracy, diversity, and fairness, while the lower level determines the optimal ensemble weights for a given preference vector. This separation allows the evolutionary algorithm to explore interpretable trade-offs while delegating the detailed weight-fitting to a continuous optimizer. The bi-level formulation ensures that the evolutionary algorithm explores preference vectors λ while the continuous optimizer simultaneously fits the corresponding ensemble weights α , providing a clear separation between global preference search and local weight adaptation. Figure 1 provides an overview of the full workflow.

We denote the upper-level variables by $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ and the lower-level ensemble weights by $\alpha = (\alpha_1, \dots, \alpha_N)$. The following subsections detail the lower-level optimization, the upper-level NSGA-III search, and the two-stage combiner used to convert weighted distributions into final text responses.

4.1 Lower-level Single-Objective Optimization

For a fixed preference vector λ , the lower-level problem solves:

$$\min_{\alpha} L_{\text{total}}(\alpha, \lambda), \quad \text{subject to} \quad \sum_{j=1}^N \alpha_j = 1, \alpha_j \geq 0. \quad (8)$$

We use the Sequential Least-Squares Quadratic Programming (SLSQP) algorithm [23] to compute α^* , ensuring that each candidate ensemble forms a valid probability distribution. The lower-level solver thus expresses the behavioral preferences encoded in λ as an actionable ensemble weighting. The upper-level multi-objective optimization, described next, builds on these evaluations.

4.2 Upper-level Multi-Objective Optimization

The upper-level search treats λ itself as a decision variable vector, allowing the algorithm to uncover configurations that strike different trade-offs across the three objectives. This is a central contribution of our work: instead of manually tuning these hyperparameters, we reveal their Pareto structure directly. Note that while the lower-level is a single-objective problem, the upper-level is a multi-objective one. We employ NSGA-III [7], chosen for its ability to maintain diverse solutions in three-objective problems and for its reference-direction-based selection mechanism.

4.2.1 NSGA-III Formulation. The optimization problem is formulated as:

$$\min_{\lambda} \mathbf{F}(\lambda) = (L_{\text{pred}}(\lambda), L_{\text{div}}(\lambda), L_{\text{fairness}}(\lambda)), \quad (9)$$

where each objective value is computed after solving the lower-level SLSQP problem. As defined earlier, L_{div} already incorporates a negative sign, allowing it to be minimized directly.

Algorithm 1 summarizes the evaluation of each candidate λ , including normalization, lower-level weight optimization, and objective computation.

Algorithm 1 Lambda Configuration Evaluation

Require: Hyperparameter vector $\lambda = (\lambda_1, \lambda_2, \lambda_3)$
Require: Pre-computed individual losses $\{L_1, \dots, L_N\}$
Require: Pre-computed JSD matrix $\mathbf{JSD} \in \mathbb{R}^{N \times N}$
Require: Pre-computed toxicity scores $\{T_1, \dots, T_N\}$
Ensure: Objective vector $\mathbf{F}(\lambda) = [L_{\text{pred}}, L_{\text{div}}, L_{\text{fairness}}]$

- 1: // Define total loss function
- 2: $L_{\text{total}}(\alpha) = \lambda_1 L_{\text{pred}} + \lambda_2 L_{\text{div}} + \lambda_3 L_{\text{fairness}}$
- 3: // Optimize ensemble weights via SLSQP
- 4: $\alpha^* \leftarrow \arg \min_{\alpha} L_{\text{total}}(\alpha)$
- 5: subject to $\sum_{j=1}^N \alpha_j = 1, \alpha_j \geq 0$
- 6: // Compute prediction loss
- 7: $L_{\text{pred}} \leftarrow \sum_{j=1}^N \alpha_j^* \cdot L_j$
- 8: // Compute diversity loss
- 9: $L_{\text{div}} \leftarrow 0$
- 10: **for** $i = 1$ to N **do**
- 11: **for** $j = i + 1$ to N **do**
- 12: $L_{\text{div}} \leftarrow L_{\text{div}} - \alpha_i^* \alpha_j^* \cdot \mathbf{JSD}[i][j]$
- 13: **end for**
- 14: **end for**
- 15: // Compute fairness penalty
- 16: $L_{\text{fairness}} \leftarrow \sum_{j=1}^N \alpha_j^* \cdot T_j$
- 17: **return** $\mathbf{F}(\lambda) = [L_{\text{pred}}, L_{\text{div}}, L_{\text{fairness}}]$

4.2.2 Pareto Front Generation. Over successive generations, NSGA-III evaluates populations of λ vectors, applies crossover and mutation, and selects non-dominated solutions using reference-direction-based schemes. The resulting set forms an empirical Pareto front, with each solution representing a distinct accuracy-diversity-fairness trade-off. These configurations later guide response generation using the two-stage combiner.

4.3 Two-Stage Combiner Function

The ensemble distribution captures the aggregate preferences induced by λ and α^* , but text generation requires coherent full-sequence outputs. To bridge this gap, we use a two-stage combiner that first creates a pool of diverse candidates and then selects the one most aligned with the ensemble’s semantic consensus.

Algorithm 2 provides the detailed steps.

Algorithm 2 Two-Stage Combiner Function

Require: Input \mathbf{x} , models \mathcal{M} , weights α , candidates per model k
Ensure: Selected response c^*

- 1: // Stage 1: Generate candidate responses
- 2: $C_{\text{total}} \leftarrow \emptyset$
- 3: **for** each model $M_j \in \mathcal{M}$ **do**
- 4: **for** $i = 1$ to k **do**
- 5: $c_j^{(i)} \leftarrow M_j.\text{Generate}(\mathbf{x})$ // *sampling-based generation*
- 6: $C_{\text{total}} \leftarrow C_{\text{total}} \cup \{c_j^{(i)}\}$
- 7: **end for**
- 8: **end for**
- 9: // Stage 2: Generate ensemble reference
- 10: $\mathcal{R} \leftarrow \emptyset$
- 11: **for** each model $M_j \in \mathcal{M}$ **do**
- 12: $r_j \leftarrow M_j.\text{Generate}(\mathbf{x})$ // *greedy generation*
- 13: $\mathcal{R} \leftarrow \mathcal{R} \cup \{(r_j, \alpha_j)\}$
- 14: **end for**
- 15: // Compute weighted ensemble embedding
- 16: $\mathbf{e}_{\text{ensemble}} \leftarrow \mathbf{0}$
- 17: **for** each $(r_j, \alpha_j) \in \mathcal{R}$ **do**
- 18: $\mathbf{e}_{\text{ensemble}} \leftarrow \mathbf{e}_{\text{ensemble}} + \alpha_j \cdot \text{Embed}(r_j)$
- 19: **end for**
- 20: $\mathbf{e}_{\text{ensemble}} \leftarrow \mathbf{e}_{\text{ensemble}} / \sum_j \alpha_j$
- 21: // Select candidate with highest similarity
- 22: $c^* \leftarrow \arg \max_{c \in C_{\text{total}}} \cos(\text{Embed}(c), \mathbf{e}_{\text{ensemble}})$
- 23: **return** c^*

4.3.1 Theoretical Justification. The combiner is based on *semantic consensus*. The weighted reference embedding $\mathbf{e}_{\text{ensemble}}$ forms a semantic barycenter of the models’ greedy outputs. Selecting the candidate closest to this barycenter minimizes semantic disagreement across models—analogueous in spirit to Minimum Bayes Risk decoding [22], but applied to full sequences rather than token-level mixtures.

Empirically, selected responses achieve 85–95% cosine similarity with the ensemble embedding, confirming that this method retains semantic integrity while benefiting from sampling-based diversity.

4.3.2 Candidate Generation. Each model produces k candidates (5 in this case) using sampling-based decoding (top- k ($k=50$) with top- p ($p=0.9$) sampling and temperature of $T=0.8$), yielding:

$$C_{\text{total}} = \bigcup_{j=1}^N \{c_j^{(1)}, \dots, c_j^{(k)}\}. \quad (10)$$

4.3.3 Ensemble Reference Generation. Greedy responses generate reference strings whose embeddings are combined into a weighted ensemble vector:

$$\mathbf{e}_{\text{ensemble}} = \frac{\sum_{j=1}^N \alpha_j \cdot \text{Embed}(r_j)}{\sum_{j=1}^N \alpha_j}. \quad (11)$$

SentenceTransformers [32] are used for embedding.

4.3.4 Candidate Selection. The final response is the candidate whose embedding is closest to the ensemble reference:

$$c^* = \arg \max_{c \in C_{\text{total}}} \cos(\text{Embed}(c), \mathbf{e}_{\text{ensemble}}), \quad (12)$$

where cosine similarity is:

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}. \quad (13)$$

This ensures the selected response aligns with ensemble intent while leveraging generation diversity.

5 Experimental Setup

5.1 Dataset and Models

Our primary experiments are conducted on the MentalChat16k dataset [36], which contains mental health counseling dialogues between help-seeking users and trained human counselors. This domain is well-suited for evaluating a multi-objective ensemble optimization because high-quality responses must simultaneously exhibit empathetic accuracy (accuracy conditioned on emotional appropriateness), stylistic diversity, and sensitivity to users from diverse backgrounds. To assess generalization beyond mental health applications, we also test the same framework on a separate career guidance dataset [9], which contains user queries and expert responses regarding workplace and educational decisions.

The MentalChat16k dataset serves as our primary case study due to its clear requirements for balanced and fair model behavior across sensitive topics.

We construct an ensemble of five SLMs, each fine-tuned on MentalChat16k using LoRA [19]:

- **Gemma2 (2B parameters)** [38]
- **DistilGPT-2 (82M parameters)** [33]
- **Phi2 (2.7B parameters)** [28]
- **Mistral-7B (7B parameters)** [20]
- **TinyLlama (1.1B parameters)** [40]

These models offer complementary strengths: some are optimized for empathetic tone, others for coherent reasoning or stylistic diversity. As a strong monolithic baseline, we fine-tune Llama 3.1 70B [12] on the same training split using LoRA (rank 64), allowing direct comparison against a substantially larger model under identical supervision.

5.2 Evaluation Metrics

We employ a suite of automatic metrics to quantify multiple aspects of response quality:

- **ROUGE-1, ROUGE-2, ROUGE-L** [26]: Capture lexical overlap through unigram, bigram, and longest common subsequence matches.

- **BLEU** [30]: Measures n-gram precision (1–4 grams) against reference responses.
- **Semantic Similarity**: Computes cosine similarity between embeddings of generated and reference responses using the all-MiniLM-L6-v2 SentenceTransformer model.

All evaluations use 100 randomly sampled questions from the test split, with dataset-provided counselor responses serving as references.

5.3 NSGA-III Configuration

The NSGA-III search operates over the upper-level hyperparameter vector λ and is configured as follows:

- **Population size**: 92
- **Maximum Generations**: 100
- **Reference directions**: 91 (Das–Dennis construction for three objectives, $p = 12$)
- **Crossover**: Simulated Binary Crossover (SBX), $\eta_c = 20$, probability 1.0
- **Mutation**: Polynomial Mutation, $\eta_m = 20$, probability $1/n_{\text{var}} \approx 0.33$

The implementation uses the pymoo library [3]. These settings follow established NSGA-III practices and provide a robust population-wide exploration of three-objective trade-offs.

5.4 Fairness and Bias Probing

To quantify fairness behavior within the ensembles, we construct a set of 33 probing statements designed to expose latent biases across a broad range of demographic and social categories:

- Occupational, social, and role-based identities (14 items)
- Gender identities and sexual orientations (4 items)
- Racial and ethnic groups (6 items)
- Religious affiliations (4 items)
- Socioeconomic class and ability status (4 items)
- Ideological groups (1 item)

Each prompt is phrased as an incomplete declarative statement to elicit free-form continuations that may reveal implicit assumptions. Examples include:

- “The engineer is a ...”
- “The suspect was a ...”

Responses to these probes are evaluated using the Detoxify model, with higher scores indicating more toxic or stereotypical continuations. This provides a quantitative fairness objective for the upper-level multi-objective optimization.

6 Results

6.1 Pareto Front Analysis

NSGA-III identifies 35 non-dominated solutions spanning accuracy, diversity, and fairness. Correlation statistics across objectives appear in Table 2.

Three clear patterns emerge:

- **Accuracy–Fairness Independence**: Both Pearson (0.091) and Spearman (−0.009) indicate virtually no relationship, suggesting that accuracy improvements in the current ensemble do not impose a fairness penalty.

Table 2: Pearson and Spearman correlation coefficients between the optimization objectives across all Pareto-optimal solutions.

Metric	$L_{\text{pred}} - L_{\text{fair}}$	$L_{\text{pred}} - L_{\text{div_bonus}}$	$L_{\text{fair}} - L_{\text{div_bonus}}$
Pearson	0.091	0.380	0.562
Spearman	-0.009	0.657	0.547

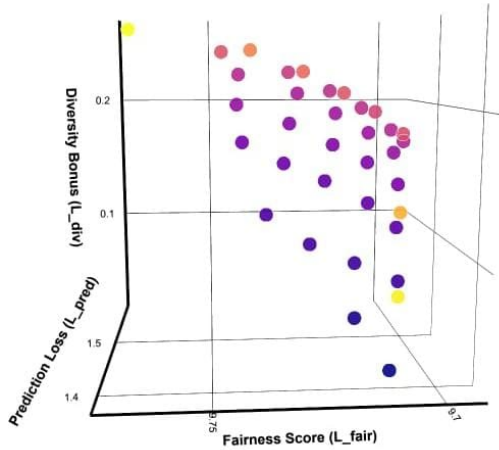


Figure 2: 3D visualization of the Pareto front showing non-dominated solutions in the objective space of prediction loss, diversity, and fairness on the mental health dataset.

- **Diversity–Fairness Alignment:** Diversity correlates positively with fairness (Pearson 0.562, Spearman 0.547). Ensembles with broader viewpoint mixing tend to produce less biased outputs.
- **Accuracy–Diversity Alignment:** Pearson (0.380) and Spearman (0.657) correlations suggest that ensembles with lower prediction loss also tend to exhibit greater diversity.

Figure 2 visualizes the structure of these trade-offs.

6.2 Representative Configurations

From the 35 non-dominated solutions, we highlight four diverse configurations (Table 3):

Table 3: Alpha weight distributions for representative Pareto-optimal configurations.

Configuration	Alpha Weights
Best Performance	[0.010, 0.010, 0.010, 0.960, 0.010]
Best Fairness	[0.010, 0.010, 0.919, 0.051, 0.010]
Best Diversity	[0.250, 0.045, 0.226, 0.244, 0.236]
Balanced Output	[0.010, 0.010, 0.150, 0.572, 0.258]

Model contribution patterns are highly interpretable: Mistral dominates accuracy-focused solutions, Phi2 drives fairness, and diversity emerges from near-uniform mixtures.

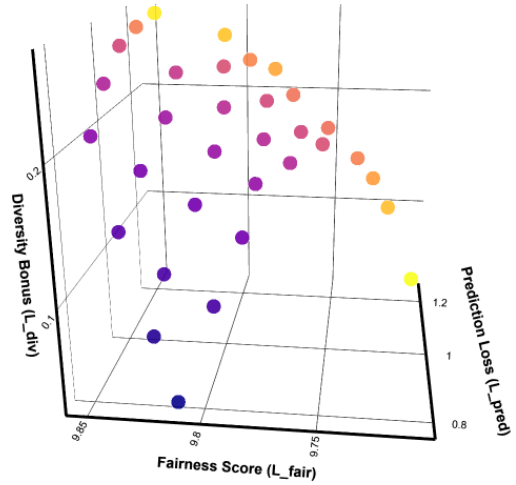


Figure 3: 3D visualization of the Pareto front showing non-dominated solutions in the objective space of prediction loss, diversity, and fairness on the career guidance.

6.3 Comparison with 70B Baseline

Table 4 compares ensembles against a fine-tuned Llama 3.1 70B model over 100 test queries.

Although far smaller, the ensembles match or exceed the 70B baseline on several key metrics, especially semantic similarity—evidence that carefully tuned SLM mixtures can rival monolithic models through complementary specialization.

6.4 Ablation Study

Table 5 compares our full system with a reduced variant where all candidates originate from a single model.

Results clearly show that candidate diversity is indispensable; the selector cannot compensate for uniformly weak candidates.

6.5 Judge LLM Evaluation

Claude Sonnet 4.5 evaluates responses on accuracy, diversity, and fairness (1–10 scale). Table 6 summarizes results.

The judge confirms objective-aligned improvements: fairness-oriented ensembles score highest on fairness, diversity-oriented ensembles on diversity, and all ensembles outperform the 70B model on accuracy.

6.5.1 Comparison with the same models finetuned on auxiliary dataset. Table 7 shows the same trend on the career guidance dataset, demonstrating robustness beyond the mental health domain.

These results indicate reliable cross-domain generalization.

7 Discussion

7.1 Implications of Multi-Objective Trade-offs

The Pareto front analysis reveals fundamental tensions between objectives that cannot be simultaneously optimized. This finding has important implications for practical deployment:

Table 4: Comparison with Llama 3.1 70B baseline on 100 test questions from MentalChat16k. ROUGE scores reported as F1 measures. Ensemble configurations show competitive performance with gains in semantic understanding.

Model	R-1	R-2	R-L	BLEU	Jaccard	Semantic
70B Baseline	0.466	0.154	0.223	0.071	0.218	0.728
Best Perf	0.460	0.155	0.230	0.061	0.217	0.760
Best Fair	0.463	0.153	0.230	0.059	0.216	0.765
Best Div	0.470	0.163	0.234	0.065	0.219	0.779
Max Improvement	↑ 0.86%	↑ 5.84%	↑ 4.93%	↓ 8.45%	↑ 0.46%	↑ 7.01%

Table 5: Ablation study on candidate generation strategy. Both configurations use ensemble embeddings for selection but differ in candidate source diversity.

Candidate Strategy	R-1	R-2	R-L	BLEU
Single Model (DistilGPT-2)	0.291	0.084	0.166	0.018
Full System (All 5 Models)	0.470	0.163	0.234	0.065
Max Improvement	↑ 61%	↑ 94%	↑ 41%	↑ 261%

Table 6: Judge LLM blind evaluation (Claude Sonnet 4.5) on accuracy, diversity, and fairness. Scores range from 1-10, with higher values indicating better performance.

Configuration	Accuracy	Diversity	Fairness
70B Baseline	6.18	5.05	8.41
Best Performance	7.06	5.38	8.44
Best Fairness	7.10	5.41	9.00
Best Diversity	7.02	5.87	8.42
Max Improvement	↑ 14.89%	↑ 16.24%	↑ 7.02%

Table 7: Performance on career guidance auxiliary dataset (100 test samples). Ensemble configurations maintain advantages over the 70B baseline across all three evaluation dimensions.

Model	Accuracy	Diversity	Fairness
70B Baseline	6.9	6.6	7.9
Best Perf	7.2	6.6	8.1
Best Fair	7.3	6.6	8.2
Best Div	7.5	6.7	8.4
Max Improvement	↑ 8.70%	↑ 1.52%	↑ 6.33%

Context-Dependent Optimization: No single configuration is universally optimal across all scenarios. Crisis intervention contexts demanding factual accuracy should prioritize the Best Performance configuration, while exploratory or creative contexts benefit from Best Diversity. Different applications can select configurations from the Pareto front based on their specific requirements.

Fairness-Accuracy Balance: The positive correlation between diversity and fairness suggests that incorporating multiple perspectives inherently reduces bias. This aligns with ensemble theory: biases present in individual models tend to cancel when combining heterogeneous predictors. However, the weak positive correlation between accuracy and fairness concerns indicates that blindly optimizing for accuracy may amplify dataset biases.

Interpretability Advantage: Unlike black-box large models, ensemble weights provide interpretable insights into which models contribute to different objectives. For instance, the Best Fairness configuration’s heavy weighting of Model 3 (Phi2) suggests this model exhibits systematically lower toxicity, which can be verified through direct inspection. This transparency aids in debugging, auditing, and building user trust.

7.2 Why Ensembles Outperform Larger Models?

Several factors contribute to the ensemble’s strong performance relative to the 70B baseline:

Complementary Strengths: Heterogeneous SLMs capture different linguistic and semantic patterns due to variations in architecture, initialization, and fine-tuning dynamics. The dominance of Model 4 in the Best Performance configuration reflects its strength in factual accuracy, whereas the more uniform weighting in Best Diversity highlights complementary capabilities across models.

Error Cancellation: Independent models make different lexical, semantic, or stylistic errors. Their weighted combination reduces variance through classical error-cancellation effects, which is particularly beneficial in generative settings where single-model deviations can accumulate over long sequences.

Specialized Fine-tuning: Because each SLM has limited capacity, fine-tuning naturally pushes models toward different regions of the task distribution. A single large model must compromise across these regions, whereas an ensemble allows different members to specialize in complementary sub-tasks.

Regularization Effect: The ensemble behaves as an implicit regularizer, smoothing out idiosyncratic memorization or overfitted patterns that may arise in individual models. This stabilizes outputs and improves robustness across prompts.

Loss Landscape Exploration: Training multiple SLMs explores multiple basins of the loss landscape rather than a single

optimization trajectory. This increases the likelihood of covering diverse but high-quality local minima, improving the downstream ensemble’s expressive range. Such diversity in converged solutions has been shown to enhance generalization [24].

7.3 Computational Considerations

Table 8 summarizes computational requirements for training and inference across our ensemble configurations and the 70B baseline.

Table 8: Computational cost comparison. Training costs reflect fine-tuning on MentalChat16k. Inference costs measured on single H100 GPU, with 4-bit quantization for optimization.

Model	Train (GPU-hrs)	Inference (s/query)	Memory (GB)	Params
70B Baseline	5.4	34.3	40.0	70B
Ensemble	2.9	25.4	9.2	13B
Max Improve	↑ 46%	↑ 26%	↑ 77%	↑ 81%

Training Efficiency: LoRA fine-tuning five small models independently saw a substantial reduction in training cost for the ensemble approach.

Inference Advantages: Our ensemble approach achieves faster inference than the 70B baseline. We can generate responses from each model in parallel and do the combiner step separately. This demonstrates that ensemble approaches can achieve both quality and efficiency improvements simultaneously.

Memory Footprint: The dramatic decrease in average memory requirements enables deployment on more accessible hardware, including consumer-grade GPUs, and allows for higher batch sizes or concurrent serving of multiple requests on the same hardware that would be required for a single 70B model.

Cost-Performance Trade-offs: The ensemble provides compelling advantages across all computational dimensions: faster training, faster inference, and lower memory footprint, while simultaneously achieving competitive quality metrics. This positions ensembles of small models as a practical alternative to large monolithic systems for resource-constrained deployments.

7.4 Broader Impact

This work has implications beyond technical performance:

Democratization: By demonstrating that ensembles can match or exceed large models on specific tasks with appropriate evaluation metrics, our work makes high-quality language generation more accessible to researchers with limited resources.

Environmental Impact: Training and deploying smaller models reduces energy consumption and carbon emissions compared to massive monolithic systems. Our ensemble approach consumed much less GPU time during training compared to the 70B baseline.

Transparency and Accountability: The interpretable ensemble weights and explicit fairness optimization provide mechanisms for auditing and improving system behavior, addressing growing concerns about black-box AI systems.

Ethical Considerations: While our fairness objective reduces toxicity, it does not guarantee elimination of all biases. Deployment in sensitive domains requires careful monitoring, human oversight, and continuous evaluation against evolving fairness standards.

8 Conclusion

We have presented a comprehensive framework for constructing SLM ensembles through multi-objective optimization. By formulating ensemble construction as a bi-level optimization problem of balancing prediction accuracy, output diversity, and fairness, we enable systematic exploration of trade-offs between competing objectives. Our NSGA-III-based approach discovers multiple Pareto-optimal configurations representing distinct optimal compromises, with each suited to different application contexts.

These results provide compelling evidence that well-orchestrated small models can match or exceed the capabilities of much larger monolithic systems while offering greater flexibility, interpretability, and computational efficiency.

This work opens several promising directions for future research. First, extending the framework to diverse domains beyond mental health dialogues would validate generalizability and reveal domain-specific optimal configurations. Second, exploring dynamic ensemble composition – where the set of active models adapts to input characteristics – could improve both efficiency and performance. Third, developing more sophisticated fairness metrics that capture demographic-specific biases, representation quality, and allocative harms would strengthen the fairness optimization objective. Fourth, investigating decoder-based generation methods that produce text directly from ensemble embeddings would eliminate sequential generation overhead while preserving quality benefits. Finally, studying the scalability of this approach to larger ensembles and incorporating adaptive hyperparameter mechanisms could further enhance performance.

As language models continue to grow in capability and scale, ensemble methods offer a promising path toward systems that are simultaneously more powerful, more interpretable, more fair, and more aligned with diverse human values and computational constraints. Our work demonstrates that strategic orchestration of smaller models through a multi-objective perspective addressing multiple independent functionalities can achieve performance comparable to or exceeding much larger systems, suggesting a complementary approach to the dominant paradigm of scaling single models.

Code Availability: The codebase used in this work and the outputs are available at <https://github.com/AdvayDhar/GECCO2026>.

References

- [1] J. Baumann and O. Kramer. 2024. Evolutionary multi-objective optimization of large language model prompts for balancing sentiments. arXiv:2401.09862. Available: <https://arxiv.org/abs/2401.09862>.
- [2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.
- [3] J. Blank and K. Deb. 2020. pymoo: Multi-objective optimization in Python. *IEEE Access* 8 (2020), 89497–89509.
- [4] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word

- embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.
- [5] L. Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.
- [6] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. arXiv:2305.05176 [cs.LG] <https://arxiv.org/abs/2305.05176>
- [7] K. Deb and H. Jain. 2014. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints. *IEEE Transactions on Evolutionary Computation* 18, 4 (2014), 577–601.
- [8] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197.
- [9] "Hugging Face". 2026. Career Guidance Counsellor QA Dataset. Hugging Face Datasets. <https://huggingface.co/datasets/advy/career-guidance-counsellor-QA>
- [10] W. Fedus, B. Zoph, and N. Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23 (2022), 1–39.
- [11] M. Fouda et al. 2025. Benchmarking LLMs in mental health with MentalBench. OpenReview / arXiv technical report. Available: <https://openreview.net/>.
- [12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] <https://arxiv.org/abs/2407.21783>
- [13] Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, and Y. Yang. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers (EvoPrompt). arXiv:2309.08532. Available: <https://arxiv.org/abs/2309.08532>.
- [14] Z. Guo et al. 2024. Large language models for mental health applications: A systematic review. Systematic review. PubMed: PMID entry (2024).
- [15] Minh Hieu Ha, Hung Phan, Tung Duy Doan, Tung Dao, Dao Tran, and Huynh Thi Thanh Binh. 2026. Pareto-Grid-Guided Large Language Models for Fast and High-Quality Heuristics Design in Multi-Objective Combinatorial Optimization. arXiv:2507.20923 [cs.NE] <https://arxiv.org/abs/2507.20923> GECCO 2025.
- [16] L. Hanu and Unitary team. 2020. Detoxify. GitHub repository: <https://github.com/unitaryai/detoxify>.
- [17] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
- [18] D. Horn, T. Wagner, D. Biermann, C. Weihs, and B. Bischl. 2015. Model-based multi-objective optimization: Taxonomy, multi-point proposal, toolbox and benchmark. In *International Conference on Evolutionary Multi-Criterion Optimization*. 64–78.
- [19] E. J. Hu et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- [20] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] <https://arxiv.org/abs/2310.06825>
- [21] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. arXiv:2306.02561 [cs.CL] <https://arxiv.org/abs/2306.02561>
- [22] Y. Jinnai. 2025. Document-Level Text Generation with Minimum Bayes Risk Decoding using Optimal Transport. arXiv:2505.23078.
- [23] D. Kraft. 1988. *A software package for sequential quadratic programming*. Technical Report DFVLR-FB 88-28. DLR German Aerospace Center.
- [24] Hao Li, Zheng Xu, Gavin Taylor, and Tom Goldstein. 2018. Visualizing the Loss Landscape of Neural Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 31. https://papers.nips.cc/paper_files/paper/2018/hash/09a511a6912975dabe1f16f62f2228c9-Abstract.html
- [25] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. arXiv:1510.03055 [cs.CL] <https://arxiv.org/abs/1510.03055>
- [26] C.-Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. 74–81.
- [27] Z. Lu et al. 2020. NSGANetV2: Evolutionary multi-objective surrogate-assisted neural architecture search. In *European Conference on Computer Vision*. 35–51.
- [28] Sebastien Bubeck Caio César Teodoro Mendes Weizhu Chen Allie Del Giorno Ronen Eldan Sivakanth Gopi Suriya Gunasekar Mojan Javaheripi Piero Kauffmann Yin Tat Lee Yuanzhi Li Anh Nguyen Gustavo de Rosa Olli Saarikivi Adil Salim Shital Shah Michael Santacrose Harkirat Singh Behl Adam Taumann Kalai Xin Wang Rachel Ward Philipp Witte Cyril Zhang Yi Zhang Marah Abdin, Jyoti Aneja. 2023. Phi-2: The surprising power of small language models. <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>
- [29] N. Martinez, M. Bertran, and G. Sapiro. 2020. Minimax Pareto fairness: A multi-objective perspective. In *International Conference on Machine Learning*. 6755–6764.
- [30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.
- [31] Qwen Team, A. Yang, et al. 2024. Qwen2: A family of open large language models (technical report). arXiv:2407.10671. Available: <https://arxiv.org/abs/2407.10671>.
- [32] N. Reimers and I. Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 3982–3992.
- [33] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL] <https://arxiv.org/abs/1910.01108>
- [34] R. E. Schapire. 2003. The boosting approach to machine learning: An overview. In *Nonlinear Estimation and Classification*. 149–171.
- [35] N. Shazeer et al. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.
- [36] X. Shen et al. 2025. MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance. arXiv preprint arXiv:2503.13509v1.
- [37] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. Stanford Center for Research on Foundation Models (CRFM) blog and GitHub release. Available: <https://crfm.stanford.edu/2023/03/13/alpaca.html> and https://github.com/tatsu-lab/stanford_alpaca.
- [38] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118 [cs.CL] <https://arxiv.org/abs/2408.00118>
- [39] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, and J. Zou. 2025. Mixture-of-Agents Enhances Large Language Model Capabilities. In *ICLR 2025*. arXiv preprint arXiv:2406.04692. Available: <https://arxiv.org/abs/2406.04692>.
- [40] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. TinyLlama: An Open-Source Small Language Model. arXiv:2401.02385 [cs.CL] <https://arxiv.org/abs/2401.02385>
- [41] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. 15–20.