DFPE: A Diverse Fingerprint Ensemble for Enhancing LLM Performance

Seffi Cohen[1], Nurit Cohen-Inger[2], Niv Goldshlager[2], Bracha Shapira[2], Lior Rokach[2]

[1]Harvard University [2]Ben Gurion University

Correspondence: seffi_cohen@hms.harvard.edu

Abstract

Large Language Models (LLMs) demonstrate impressive capabilities but exhibit inconsistent performance across diverse domains. We propose DFPE (Diverse Fingerprint Ensemble), a novel training-free method that systematically constructs subject-adaptive ensembles by balancing model diversity and competence. DFPE introduces three key innovations: (1) semantic fingerprinting using averaged response embeddings to capture distinct problem-solving patterns, (2) DBSCAN-based clustering with quantile-based competence filtering to ensure diverse yet capable model selection, and (3) exponentially-weighted aggregation adapted to subject-specific performance. Our method's effectiveness is highlighted on the challenging MMLU-pro benchmark, where DFPE achieves a striking 17.1 percentage point gain over the best single model, reaching 71.4% accuracy. This strong performance is consistent across other standard benchmarks, with significant accuracy improvements of 4.4 points on AGIEval and 2.7 points on MMLU. Our results underscore that a systematic approach to ensemble construction - one that balances diversity, subject-specific competence, and adaptive weighting, can substantially enhance the generalization and robustness of LLMs on multifaceted language understanding tasks.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks (Chang et al., 2024; Matarazzo and Torlone, 2025). However, even state-of-the-art models exhibit heterogeneous performance, excelling in some domains while underperforming in others. This performance inconsistency becomes particularly apparent on complex, multi-domain benchmarks like MMLU (Hendrycks et al., 2020), MMLU-pro (Wang et al., 2024), and AGIEval (Zhong et al., 2023), which span diverse subjects requiring specialized knowledge and reasoning capabilities. On these challenging testbeds, no single LLM can consistently deliver top-tier performance across the board, revealing critical gaps in their generalization abilities.

Ensembling multiple LLMs offers a powerful strategy to mitigate these weaknesses and enhance robustness (Lu et al., 2024a; Jiang et al., 2023a). By combining the complementary strengths of different models, an ensemble can theoretically achieve superior accuracy and reliability. However, naive approaches like simple majority voting often fail to outperform the best single model, as they can be diluted by the inclusion of weaker or redundant models. The central challenge, therefore, is not simply to combine models, but to do so intelligently: by preserving strategic diversity, ensuring a baseline of competence for each task, and adapting the ensemble composition to the specific demands of a subject, all without resorting to costly retraining or fine-tuning.

In this paper, we introduce the Diverse Fingerprint Ensemble (DFPE), a novel, training-free method designed to systematically construct high-performing LLM ensembles. DFPE addresses the core challenges of ensembling by integrating three key ideas: diversity-preserving clustering, adaptive competence filtering, and performance-based weighting. Our method first creates a "fingerprint" for each model based on its response patterns for a given subject, then clusters these fingerprints to identify groups of models with similar problem-solving approaches. It then filters out underperforming models using a dynamic, subject-specific quantile threshold and selects the best representative from each remaining cluster. Finally, it aggregates their predictions using an exponential weighting scheme that gives more influence to models with higher validation accuracy.

We demonstrate the effectiveness of DFPE across three demanding benchmarks. The results

Findings of the Association for Computational Linguistics: EACL 2026, pages 5326–5336

are particularly striking on the highly challenging MMLU-pro dataset, where DFPE achieves an accuracy of 71.4%, a remarkable 17.1 percentage point improvement over the best individual model. This success is mirrored on AGIEval, where it secures a 4.4 point gain, and on the standard MMLU benchmark, where it improves accuracy by 2.7 points over the strongest single model.

Our primary contributions are:

- **Diversity-Preserving Clustering:** We introduce a "fingerprinting" technique that captures each model's unique response patterns. By clustering these fingerprints, our method systematically constructs an ensemble with diverse problem-solving strategies, preventing it from collapsing into a single solution path and enhancing its robustness to varied question types.
- **Adaptive Competence Filtering:** To ensure only capable models contribute, we employ a subject-level quantile filter. This dynamic threshold automatically adapts to the difficulty of each topic, pruning underperformers while retaining valuable, specialized models that might otherwise be discarded by a fixed performance cut-off.
- **Training-Free Adaptive Weighting:** DFPE combines the selected models using an exponential weighting scheme based on their subject-specific validation accuracy. The entire framework is training-free, operating only on model outputs, which makes it a lightweight yet powerful method for significantly enhancing LLM performance without any fine-tuning.

These results demonstrate that a principled, data-driven approach to ensemble construction can unlock substantial performance gains, paving the way for more reliable and accurate LLM applications.

2 Related Work

Research on ensembling LLMs can be organised along three factors: (i) access to model internals, (ii) additional training cost, and (iii) how the method balances diversity against individual-model quality. We review prior work accordingly.

High-overhead or parameter-access methods. Token- or span-level fusion approaches such as LLM-TOPLA (Tekin et al., 2024), PackLLM (Mavromatis et al., 2024), and SpanLLM ("SweetSpan") (Xu et al., 2024a) coordinate generations by inspecting hidden states, while LoRA-based merging tunes shared adapters across experts (Wang et al., 2023). Query routers like SelectLLM (Maurya et al., 2024) and the reward-guided ZOOTER (Lu et al., 2024b) learn a dispatch function that chooses a single expert per query. Although these methods deliver strong accuracy, their dependence on gradient updates or hidden representations limits rapid, large-scale deployment.

Training-free output-level ensembles. Prompt-ensemble techniques improve calibration without touching model weights: CAPE (**C**alibration via **A**ugmented **P**rompt **E**nsembles) augments templates or answer options to obtain diverse views of the same input (Jiang et al., 2023b); Boosted-Prompts achieve a similar effect through paraphrasing (Pitis et al., 2023); BayesPE learns Bayesian weights over prompt variants for uncertainty quantification (Tonolini et al., 2024). To fuse token probabilities across heterogeneous vocabularies, DeePEn maps logits into a universal relative space (Huang et al., 2024), whereas EVA learns explicit vocabulary-alignment matrices and averages projected logits during decoding (Xu et al., 2024b). At the other extreme, Self-MoA shows that repeatedly sampling a single strong model can outperform mixtures when one model dominates the quality spectrum (Li et al., 2025).

Granularity & combination strategies. Token-level fusion (EVA, DeePEn) offers fine-grained control but hinges on accurate alignment; sample-level self-consistency voting is robust yet ignores internal structure; span-level ensemble (SpanLLM) strikes a middle ground. **DFPE** introduces subject-level adaptivity: fingerprint clustering groups models by behavioural similarity, and quantile-based filtering prunes under-performers before weighted voting, thereby retaining beneficial diversity while safeguarding quality.

Routing versus fusion. Routing methods (ZOOTER, SelectLLM) pick one expert to minimise cost, whereas fusion methods (EVA, DeePEn, SpanLLM) combine multiple outputs. DFPE is a hybrid: its filter routes out weak models and its vote fuses the survivors, achieving competitive accuracy without internal access or additional training.

Table 1 summarises how DFPE differs from prior art: it is the first *training-free* ensemble that (i) adapts at the subject level, (ii) explicitly optimises the diversity–quality trade-off, and (iii) requires only a small labelled validation shard.

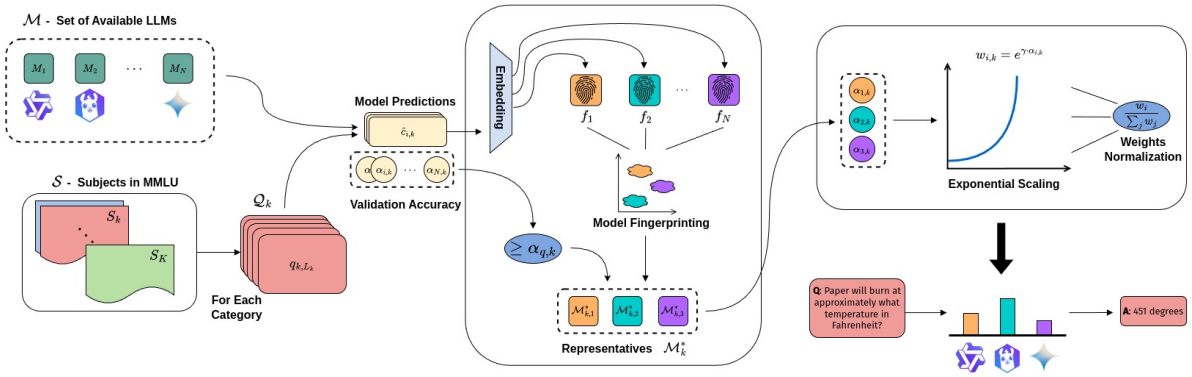

Figure 1: DFPE pipeline: LLMs are fingerprinted, clustered for diversity, filtered by quantile threshold, and weighted by performance for final aggregation.

Method	Few-Shot	Subject Adapt.	Diversity Opt.
LLM-TOPLA	×	✓	✓
SelectLLM	×	✓	×
PackLLM	×	✓	✓
SweetSpan	✓	×	✓
DeePEn	×	✓	✓
EVA	✓	×	✓
Boosted Prompts	✓	×	✓
CAPE	✓	✓	×
ZOOTER	×	✓	✓
LoRA Ensembles	×	×	✓
DFPE (Ours)	✓	✓	✓

Table 1: Comparison of related methods. "Zero/Few-Shot" indicates methods that do not require additional training or fine-tuning. DFPE requires no additional training, uses few-shot validation for subject adaptivity, and ensures diversity via clustering and quantile-based filtering.

3 Method

Our method, Diverse Fingerprint Ensemble (DFPE), is a training-free framework designed to construct a high-performing, subject-adaptive ensemble from a pool of existing LLMs. The core idea is to systematically select and weight models by balancing two critical factors: their performance on a specific subject and the uniqueness of their problem-solving approach. This is achieved through a four-step pipeline that integrates model fingerprinting, diversity-preserving clustering, adaptive competence filtering, and performance-based weighting.

3.1 Overview

As illustrated in Figure 1, for each subject within a benchmark, our method proceeds as follows:

1. **Model Fingerprinting and Clustering:** We first generate a "fingerprint" for each model by embedding its responses on a small, subject-specific validation set. These fingerprints capture the model's characteristic response patterns. We then use DBSCAN clustering to group models with similar fingerprints, identifying distinct problem-solving strategies present in the model pool.
2. **Adaptive Competence Filtering:** To ensure only proficient models contribute, we apply a dynamic, quantile-based accuracy filter. For each subject, we calculate a performance threshold based on the distribution of model accuracies on the validation set and discard any model that falls below this threshold.
3. **Representative Model Selection:** From each surviving cluster, we select only the single highest-performing model as the representative for that group. This step is crucial as it preserves strategic diversity (one model per approach) while simultaneously ensuring that each selected approach is represented by its most competent advocate.
4. **Adaptive Weighting and Prediction:** Finally, we assign weights to the selected representative models using an exponential scaling of their validation accuracy. This gives more influence to stronger models while still allowing for contributions from more specialized ones. The final ensemble prediction is determined by a weighted vote of these models.

This process ensures accurate, robust ensembles

with diverse yet competent strategies per subject.

3.2 Formal Description

Notation Let $\mathcal{M} = \{M_1, \ldots, M_N\}$ be the pool of N available LLMs. Let $\mathcal{S} = \{S_1, \ldots, S_K\}$ be the set of K subjects in a benchmark. For each subject S_k, we have a validation set of questions $\mathcal{Q}_k = \{q_1^{(k)}, \ldots, q_{m_k}^{(k)}\}$ with m_k samples. For a given question $q_j^{(k)} \in \mathcal{Q}_k$, let $\hat{c}_{i,k,j}$ be the prediction of model M_i, and let $\alpha_{i,k} = \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbb{I}[\hat{c}_{i,k,j} = c_j^{(k)}]$ be the validation accuracy of model M_i on subject S_k, where $c_j^{(k)}$ is the ground truth label.

3.3 Unsupervised Subject Induction (Optional Extension)

DFPE, as described above, assumes a partition of the validation set into subjects in order to compute per-group fingerprints and accuracies. When a benchmark does not provide a clean subject taxonomy, we can extend DFPE by inducing such groups directly from the question texts. This yields a training-free, unsupervised "pseudo-subject" layer that preserves DFPE's pipeline while removing reliance on human-defined disciplines.

Question fingerprinting Let $\mathcal{Q}^{\text{val}}$ be the validation questions. We embed each question text using the same sentence-embedding mechanism used for model fingerprints, but applied to the question instead of a model response. Concretely, for each question $q \in \mathcal{Q}^{\text{val}}$, compute

$$\mathbf{g}(q) = E(q) \in \mathbb{R}^d,$$

where $E(\cdot)$ is the encoder used.

Unsupervised grouping via DBSCAN. We cluster the question embeddings $\{\mathbf{g}(q)\}_{q \in \mathcal{Q}^{\text{val}}}$ using DBSCAN with cosine distance, producing cluster identifiers $\pi(q) \in \{1, \ldots, \tilde{K}\} \cup \{-1\}$, where -1 denotes outliers. Each cluster $c \in \{1, \ldots, \tilde{K}\}$ is treated as a pseudo-subject $\tilde{S}_c$. The outlier set $\pi(q) = -1$ naturally forms an "Other" pseudo-subject when present.

Plug-in to DFPE Given $\pi(\cdot)$, we define per-cluster validation shards $\tilde{\mathcal{Q}}_c^{\text{val}} = \{q \in \mathcal{Q}^{\text{val}} : \pi(q) = c\}$, and compute per-model accuracies $\alpha_{i,c}$ on each shard exactly as in Section 3. We then run the standard DFPE steps per pseudo-subject: (i) compute model fingerprints from responses restricted to $\tilde{\mathcal{Q}}_c^{\text{val}}$, (ii) cluster model fingerprints for diversity, (iii) apply quantile-based competence filtering within c, and (iv) weight the selected representatives using exponential accuracy weights. Thus, once pseudo-subjects are induced, Algorithm 1 applies without modification.

Assigning test questions to pseudo-subjects. At inference time, a test question q^{test} is assigned to a pseudo-subject by embedding $\mathbf{g}(q^{\text{test}})$ and mapping it to the nearest dense region discovered by DBSCAN. otherwise it falls back to the outlier/"Other" group. In settings where using unlabeled test questions for partitioning is acceptable, an even simpler alternative is to run DBSCAN on all question texts and use validation labels only for estimating $\alpha_{i,c}$ inside each cluster.

Practical notes and fallbacks. Very small clusters can yield noisy accuracy estimates; a practical safeguard is to merge clusters below a minimum size into "Other" or to revert to a coarser partition. In the extreme, setting $\tilde{K} = 1$ (one group containing all questions) recovers a non-subject-adaptive DFPE variant that still performs diversity-aware selection and weighting globally.

Model Fingerprinting and Clustering For each model M_i and subject S_k, we generate a fingerprint vector $\mathbf{f}_{i,k} \in \mathbb{R}^d$ that captures its response behavior patterns. Specifically, let $E(\cdot) : \text{String} \to \mathbb{R}^d$ be a pre-trained sentence embedding function (we use 'all-MiniLM-L6-v2' (Hugging Face, 2023)). For each question $q_j^{(k)} \in \mathcal{Q}_k$, let $r_{i,k,j}$ be the textual response of model M_i. The fingerprint is computed as:

$$\mathbf{f}_{i,k} = \frac{1}{m_k} \sum_{j=1}^{m_k} E(r_{i,k,j})$$

This averaging captures the centroid of the model's response patterns in semantic space. **Theoretical Justification:** By the Johnson-Lindenstrauss lemma (Achlioptas, 2003), high-dimensional embeddings preserve pairwise distances, making cosine similarity in this space a valid measure of semantic similarity. Models with similar problem-solving approaches will generate semantically similar responses, resulting in closer fingerprints. We apply DBSCAN clustering (Ester et al., 1996) with cosine distance $d_{\cos}(\mathbf{f}_{i,k}, \mathbf{f}_{i',k}) = 1 - \frac{\mathbf{f}_{i,k} \cdot \mathbf{f}_{i',k}}{||\mathbf{f}_{i,k}|| \cdot ||\mathbf{f}_{i',k}||}$ to partition models into clusters, where each cluster contains models with similar semantic response patterns.

Adaptive Competence Filtering and Selection To prune underperforming models, we define a quantile parameter q (e.g., $q = 0.1$). For each subject S_k, we compute the q-quantile accuracy threshold, $\alpha_{q,k}$, from the set of validation accuracies $\{\alpha_{i,k}\}_{i=1}^{N}$. Only models with $\alpha_{i,k} \geq \alpha_{q,k}$ are retained. **Theoretical Rationale:** This adaptive threshold automatically adjusts to subject difficulty, ensuring consistent relative performance requirements while preserving subject-specific expertise that fixed thresholds might eliminate.

From this filtered set of competent models, we select one representative model $M_{k,j}^*$ from each cluster C_j by choosing the one with the highest validation accuracy:

$$M_{k,j}^* = \arg \max_{M_i \in C_j \text{ and } \alpha_{i,k} \geq \alpha_{q,k}} \alpha_{i,k}.$$

This yields the final set of representative models for subject S_k, denoted as $\mathcal{M}_k^*$.

Adaptive Weighting and Ensemble Prediction For each selected model $M_i \in \mathcal{M}_k^*$, we assign a weight based on its validation accuracy, scaled by a factor γ to control the sharpness of the distribution:

$$w_{i,k} = \exp(\gamma \cdot \alpha_{i,k}).$$

Theoretical Justification: The exponential weighting follows from the principles of maximum entropy: maximize the diversity of the ensemble while respecting performance constraints. The parameter γ controls the bias-variance tradeoff: higher γ emphasizes top performers (lower variance) while lower γ includes more diverse opinions (higher variance but potential for better generalization). Weights are normalized: $\tilde{w}_{i,k} = w_{i,k} / \sum_{M_j \in \mathcal{M}_k^*} w_{j,k}$.

For any test question, the final ensemble prediction is obtained by aggregating the predictions of the models in $\mathcal{M}_k^*$ via a weighted vote. The choice c receiving the highest total weight is selected as the final answer:

$$\hat{c}_{\text{final}} = \arg \max_{c} \sum_{M_i \in \mathcal{M}_k^*} \tilde{w}_{i,k} \cdot \mathbb{I}[\hat{c}_{i,k} = c].$$

The complete process is summarized in Algorithm 1.

Algorithm 1 DFPE (Diverse Fingerprint Ensemble)

Input: Model pool $\mathcal{M}$, Subjects $\mathcal{S}$, Validation sets $\{\mathcal{Q}_k\}_{k=1}^{K}$
for each subject $S_k \in \mathcal{S}$ **do**
  Compute validation accuracies $\{\alpha_{i,k}\}_{i=1}^{N}$ on $\mathcal{Q}_k$.
  Generate fingerprints $\{\mathbf{f}_{i,k}\}_{i=1}^{N}$ from model responses.
  Cluster fingerprints using DBSCAN to get clusters $\{C_j\}$.
  Determine quantile accuracy threshold $\alpha_{q,k}$.
  Initialize empty set of representatives $\mathcal{M}_k^*$.
  for each cluster C_j **do**
    Select representative $M_{k,j}^* = \arg\max_{M_i \in C_j, \alpha_{i,k} \geq \alpha_{q,k}} \alpha_{i,k}$.
    Add $M_{k,j}^*$ to $\mathcal{M}_k^*$.
  end for
  Compute normalized weights $\{\tilde{w}_{i,k}\}$ for all $M_i \in \mathcal{M}_k^*$.
end for
Output: For any test question in subject S_k, predict using the weighted vote of models in $\mathcal{M}_k^*$.

3.4 Rationale

Our method strikes a balance between several key principles:

- **Diversity Preservation**: Clustering models by their response patterns ensures a variety of solution strategies are represented, reducing redundancy.
- **Quantile-based Competence**: The adaptive quantile threshold removes weak models on a per-subject basis, guaranteeing a baseline performance level.
- **Adaptive Weighting**: Higher-accuracy models naturally exert greater influence, but all retained models still contribute to the final decision.
- **Practical Simplicity**: Operating at the level of model outputs avoids complex token- or span-level computations, making the method scalable and broadly applicable.

By combining these elements, our approach effectively exploits the complementary strengths of multiple LLMs.

4 Experimental Setup

We evaluate our DFPE method on three diverse and challenging multitask benchmarks using a pool of ten publicly available Large Language Models (LLMs). Our setup is designed to be lightweight,

relying on few-shot validation to guide the ensemble construction without any model fine-tuning.

4.1 Datasets

To demonstrate the broad applicability of our method, we evaluate on three challenging benchmarks:

- **MMLU (Massive Multitask Language Understanding)** (Hendrycks et al., 2020): A widely used benchmark comprising 57 subjects across STEM, humanities, and social sciences. It contains 14,079 test samples and 1,540 validation samples.
- **MMLU-pro**: A more challenging, professionally-curated version of MMLU that features harder questions and has been shown to be a more reliable measure of model reasoning capabilities.
- **AGIEval**: A benchmark designed to evaluate foundation models on human-centric tasks, derived from standardized exams like the Gaokao (Chinese college entrance exams) and American college admission tests.

For each benchmark, we use the provided validation split for model fingerprinting and selection, and the official test set for final evaluation.

4.2 Base Models

Our model pool consists of ten open-source LLMs with up to 9 billion parameters, chosen to represent a variety of architectures and training methodologies. The models include GLM-4-9B, multiple versions of Qwen2.5, Mistral-v0.3, Phi-3.5-mini, Llama-3.1, Gemma-2, Apollo2, Starling-LM, and Yi-1.5. The performance of these models on the standard MMLU benchmark is detailed in Table 2. This diversity in both architecture and baseline performance (ranging from 61.5% to 70.8%) provides a rich foundation for constructing a powerful ensemble that can leverage complementary strengths.

4.3 Implementation and Evaluation Metrics

Implementation Details. Our implementation is entirely training-free and operates purely on model outputs. For clustering, we use DBSCAN with cosine similarity and empirically tuned ϵ values that adapt to the embedding space characteristics. For filtering and weighting, we set the quantile parameter $q = 0.05$ and the accuracy scaling factor $\gamma = 5.0$, as these values consistently yielded strong performance during validation across different subjects. The sentence embedding model (all-MiniLM-L6-v2) provides 384-dimensional representations that effectively capture semantic similarities in model responses. The source code and complete experimental setup are provided for reproducibility.

Model	Params	Accuracy
GLM-4	9B	0.7076
Qwen2.5	7B	0.6476
Qwen2.5	3B	0.6605
Mistral v0.3	7B	0.6316
Phi-3.5-mini	4B	0.7046
Llama-3.1	8B	0.6508
Gemma-2	9B	0.6804
Apollo2	7B	0.7034
Starling-LM-alpha	7B	0.6149
Yi-1.5	6B	0.6239

Table 2: Overall accuracies of the base LLMs on the MMLU benchmark. The diverse range of performance provides a strong foundation for our ensemble approach.

Evaluation Metrics and Baselines. We report overall accuracy with 95% confidence intervals computed using bootstrap sampling (n=1000 resamples). To evaluate our method comprehensively, we compare DFPE against multiple baselines:

- **Best Single Model (BSM)**: The performance of the single best-performing model from our pool on the test set.
- **Majority Voting (MVoting)**: A standard ensemble method where all models cast equal-weight votes.
- **Accuracy-Weighted Voting**: Models weighted by their overall validation accuracy.
- **Random Selection**: Randomly selecting 5 models from the pool (averaged over 10 runs).

All statistical comparisons use paired t-tests with Bonferroni correction for multiple comparisons.

5 Results

Our experiments demonstrate that DFPE consistently and significantly outperforms both the best single model and a standard majority voting ensemble across all three benchmarks. Furthermore, when compared against contemporary ensemble methods on the standard MMLU benchmark, DFPE achieves state-of-the-art performance,

Benchmark	BSM	Maj. Vote	Acc. Weight	DFPE (Ours)
MMLU-pro	54.3±1.2	55.7±1.1	58.2±1.3	**71.4±0.9**
AGIEval	50.2±1.8	48.4±1.6	51.1±1.7	**54.6±1.4**
MMLU	70.8±0.7	72.4±0.6	73.1±0.6	**73.5±0.5**

Table 3: Main performance comparison across three benchmarks. Accuracies reported as percentages with 95% confidence intervals. DFPE consistently and significantly outperforms all baselines ($p < 0.001$ for all comparisons).

highlighting the effectiveness of our principled approach to balancing model diversity and subject-specific competence.

As summarized in Table 3, DFPE achieves substantial accuracy gains on all evaluated datasets. The most dramatic improvement is on the challenging MMLU-pro benchmark, where DFPE achieves an accuracy of 71.4%, a remarkable 17.1 percentage point increase over the best single model (BSM). This demonstrates our method's ability to unlock significant performance on tasks where individual models struggle.

The trend continues on AGIEval, where DFPE delivers a 4.4 point gain, and on the standard MMLU benchmark, where it provides a solid 2.7 point improvement over an already strong BSM. Notably, DFPE also consistently surpasses the Majority Voting baseline, which itself often improves upon the BSM. This underscores that DFPE's intelligent selection and weighting strategy is far more effective than a naive ensemble.

5.1 Comparison with Contemporary Ensemble Methods

To situate DFPE within the current research landscape, we compare its performance on the standard MMLU benchmark against several recent ensemble methods. As shown in Table 4, DFPE achieves the highest accuracy among all compared techniques.

DFPE's superior performance can be attributed to its unique design. Unlike methods that rely on complex routing (SelectLLM) or internal model mechanics (DeePEn), DFPE operates on simple model outputs, making it lightweight and broadly applicable. The computational overhead of DFPE is modest: fingerprinting requires only embedding computation (384-dimensional vectors), clustering scales as $O(N^2)$ where $N \leq 10$, and inference time increases linearly with selected models. On average, DFPE takes ×1.8 the inference time of a single model when using 6 models (balanced mode) and ×2.7 with 9 models (optimal mode) on NVIDIA A100 GPUs, while achieving 17.1% accuracy gains that justify this overhead. Crucially, its fingerprint-based clustering directly addresses the diversity-quality trade-off. This allows it to outperform methods like Self-MoA, which questions the value of model diversity, and surpass even sophisticated fusion techniques like SpecFuse. By systematically ensuring that the ensemble is composed of a diverse set of competent models for each specific task, DFPE robustly integrates their complementary strengths, leading to state-of-the-art results.

Method	Ensemble Type	Overall Acc.
Self-MoA	Single-Model Sampling	0.691
LLM-TOPLA	Multi-granular Fusion	0.706
CAPE	Augmented Prompts	0.710
DeePEn	Deep Ensemble	0.712
Boosted Prompts	Boosted Prompts	0.712
PackLLM	Test-Time Coordination	0.715
SelectLLM	Query-Aware Routing	0.719
SpecFuse	Specialized Fusion	0.722
DFPE (Ours)	**Fingerprint**	**0.735**

Table 4: Comparison to recent ensemble methods on MMLU. DFPE achieves the highest overall accuracy, demonstrating its state-of-the-art performance.

5.2 Ablation and Sensitivity Analysis

A comprehensive ablation study, presented in Table 5, confirms that each component of DFPE - clustering, quantile filtering, and adaptive weighting - contributes positively to the final performance. The removal of any single component results in a noticeable drop in accuracy, with the full model achieving the best results. Sensitivity analyses for our key

Configuration	Overall Acc.	Discipline Acc.
Full DFPE	**0.735**	**0.740**
No Clustering	0.724	0.727
No Quantile Filtering	0.728	0.731
No Adaptive Weighting	0.731	0.734

Table 5: Ablation study results on MMLU. Each component contributes to the overall performance improvement.

hyperparameters (Figure 2) show that our method is robust across a reasonable range of parameter settings, simplifying tuning.

6 Further Analysis and Practical Considerations

Beyond the primary results, we analyze the internal dynamics of DFPE to provide deeper insights into its behavior and practical application.

6.1 Model Participation and Practical Trade-offs

In our optimal accuracy configuration, DFPE retains most of the available models (an average of 9 out of 10) per subject. This suggests that for achieving maximum performance, the diversity and complementary strengths of nearly all models contribute to the ensemble's success, even those with individually lower performance.

However, for applications where computational cost is a concern, DFPE can be tuned for efficiency. We identified a "Balanced Configuration" that offers a strong compromise between performance and cost. By using a higher quantile threshold ($q = 0.5$) and a larger DBSCAN ϵ ($\epsilon = 0.001$), this setup achieves an impressive 72.5% accuracy on MMLU (only 1% below optimal, and still 1.7% above best single model) while significantly reducing the mean number of models per subject to just 6. This demonstrates that near-optimal performance can be maintained with substantially reduced computational overhead.

6.2 Analysis of Ensemble Composition

A detailed, subject-level analysis of DFPE's composition is deferred to Appendix A (Sections A.2–A.3). In brief, the ensemble size adapts to subject difficulty, ranging from 1 to 10 models, typically 6–8 illustrating DFPE's per-subject flexibility (Fig. 3). Pairwise co-occurrence patterns further reveal stable complementarities among model families alongside specialist models that contribute niche strengths (Fig. 4). These observations inform practical pool design and support the balanced configuration discussed above.

7 Discussion

The strong and consistent performance of DFPE, especially on the difficult MMLU-pro benchmark, provides several key insights. The 17.1 point gain on MMLU-pro suggests that the value of a diversity-preserving ensemble is magnified on more challenging tasks where individual models are more likely to fail. By systematically retaining diverse yet competent problem-solving strategies, DFPE constructs a more resilient system that is less susceptible to the idiosyncratic weaknesses of any single model.

Performance Characteristics and Model Pool. DFPE's strength arises from its unique combination of clustering, filtering, and weighting. While massive ensembles like Mixtral-of-Experts (Jiang et al., 2024) (8x7B) can achieve higher absolute scores, they require substantially more computational resources. DFPE, in contrast, demonstrates the power of efficient ensembling, managing to surpass some much larger monolithic models with a pool of smaller LLMs. This highlights the value of a diversity-focused approach. The effectiveness of DFPE is, however, predicated on a diverse model pool; if the base models are too homogeneous, the benefits of clustering diminish. Given the rapid proliferation of new open-source LLMs, assembling a diverse pool is increasingly feasible.

Practical Trade-offs and Configurations. Our analysis reveals that DFPE can be flexibly tuned to balance performance and computational cost. For applications where accuracy is paramount, the optimal configuration uses most of the available models. However, for resource-constrained environments, a "Balanced Mode" can be employed. As shown in our analysis, by adjusting the quantile and clustering parameters, it is possible to reduce the average ensemble size to 6 models while maintaining performance within 1% of the optimal score. This flexibility, combined with insights from model co-occurrence patterns, provides a clear path for practitioners to optimize deployment based on their

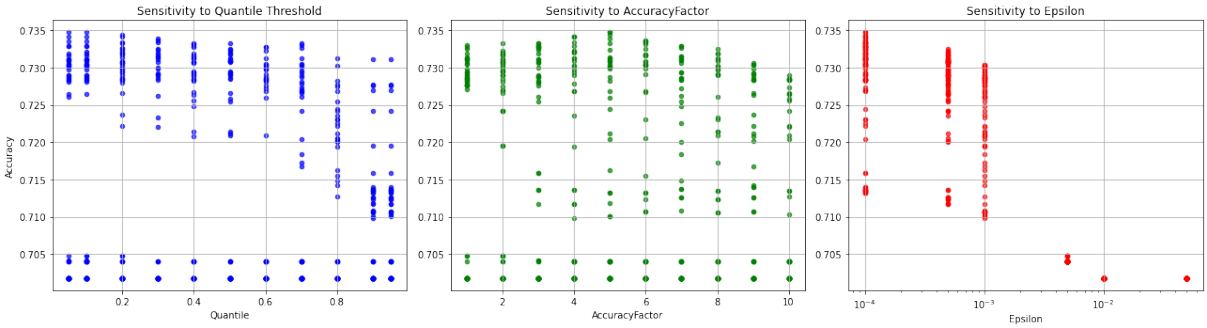

Figure 2: Sensitivity analysis. Left: Accuracy vs. Quantile Threshold; Middle: Accuracy vs. AccuracyFactor; Right: Accuracy vs. ϵ (log scale). Performance remains robust within moderate parameter ranges, easing the tuning process.

specific needs.

8 Conclusion

In this paper, we introduced DFPE, a training-free ensemble method that leverages diversity and adaptivity to improve LLM performance. DFPE clusters models via "fingerprint" patterns, filters underperformers with a quantile-based threshold, and applies exponential weighting that emphasizes top-performing models while preserving valuable secondary perspectives. In AGIEval benchmark, DFPE achieves a 4. 4% On the MMLU benchmark, DFPE achieves a 2.7% improvement over the best single model, and this gain grows to a remarkable 17.1% on the more challenging MMLU-pro benchmark. These results underscore the importance of diverse solution strategies, selective filtering, and dynamic weighting in ensemble construction.

Future work will focus on refining question-level adaptivity, exploring more scalable clustering for larger model pools, and extending the fingerprinting strategy to open-ended generation tasks. By balancing diversity, adaptivity, and efficiency, DFPE offers a practical and robust framework for building high-performing ensembles.

Limitations

DFPE has three primary limitations. First, it relies on a small set of labeled validation data for each subject to guide selection and weighting. Second, its effectiveness depends on the diversity of the initial model pool. Third, the current implementation is designed for multiple-choice question answering. Extending the fingerprinting concept to open-ended generation tasks, where defining "correctness" and "similarity" is more complex, remains a key direction for future work.

Acknowledgements

We used ChatGPT-5 for editing the language and refining the presentation of the text in this paper. The authors affirm that all research content and ideas are their own, and they take full responsibility for the final submitted manuscript.

References

Dimitris Achlioptas. 2003. Database-friendly random projections: Johnson-lindenstrauss with binary coins. In *Journal of computer and System Sciences*, volume 66, pages 671–687. Elsevier.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, and 1 others. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yichong Huang, Xiaocheng Feng, Baohang Li, Yang Xiang, Hui Wang, Bing Qin, and Ting Liu. 2024. Enabling ensemble learning for heterogeneous large language models with deep parallel collaboration. *arXiv preprint arXiv:2404.12715*.

Hugging Face. 2023. Sentence transformers all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas,

Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023a. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu I. Pitis, Roger B. Grosse, and Jimmy Ba. 2023b. Calibrating language models via augmented prompt ensembles. In *Proceedings of the ICML 2023 Workshop on Challenges in Deployable Generative AI*.

Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. 2025. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024a. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *arXiv preprint arXiv:2407.06089*.

Keming Lu, Hongyi Yuan, Runji Lin, Junyang Lin, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024b. Routing to the expert: Efficient reward-guided ensemble of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1964–1974.

Andrea Matarazzo and Riccardo Torlone. 2025. A survey on large language models with some insights on their capabilities and limitations. *arXiv preprint arXiv:2501.04040*.

Kaushal Kumar Maurya, KV Srivatsa, and Ekaterina Kochmar. 2024. Selectllm: Query-aware efficient selection algorithm for large language models. *arXiv preprint arXiv:2408.08545*.

Costas Mavromatis, Petros Karypis, and George Karypis. 2024. Pack of llms: Model fusion at test-time via perplexity optimization. *arXiv preprint arXiv:2404.11531*.

Silviu Pitis, Michael R Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *arXiv preprint arXiv:2304.05970*.

Selim Tekin, Fatih Ilhan, Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. Llm-topla: Efficient llm ensemble by maximising diversity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11951–11966.

Francesco Tonolini, Nikolaos Aletras, Jordan Massiah, and Gabriella Kazai. 2024. Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9280–9298.

Xi Wang, Laurence Aitchison, and Maja Rudolph. 2023. Lora ensembles for large language model fine-tuning. *arXiv preprint arXiv:2310.00035*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Yangyifan Xu, Jianghao Chen, Junhong Wu, and Jiajun Zhang. 2024a. Hit the sweet spot! span-level ensemble for large language models. *arXiv preprint arXiv:2409.18583*.

Yangyifan Xu, Jinliang Lu, and Jiajun Zhang. 2024b. Bridging the gap between different vocabularies for llm ensemble. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 7133–7145.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

A Appendix A

While our main results focus on the optimal accuracy configuration, practitioners often need to balance performance gains against computational costs. Here we present a detailed analysis of an alternative configuration that achieves near-optimal performance while significantly reducing computational overhead.

A.1 Balanced Configuration Parameters

We identified a balanced configuration with the following parameters:

- Quantile threshold: 0.5 (vs. 0.05 in optimal setting)
- AccuracyFactor: 7 (vs. 5 in optimal setting)
- DBSCAN Epsilon: 0.001 (vs. 0.0001 in optimal setting)

This configuration achieves an average accuracy of 72.5% (1% below optimal) while reducing the mean number of models per subject to 6 (compared to 9 in the optimal setting).

A.2 Model Selection Analysis

Figure 3 shows the distribution of selected models across subjects. Several key patterns emerge:

- Variation in model count: The number of selected models varies from 1 to 10 across subjects
- Subject-specific adaptation: Different subjects benefit from different ensemble sizes
- Consistent core: Most subjects maintain 6-8 models, suggesting a natural balance point

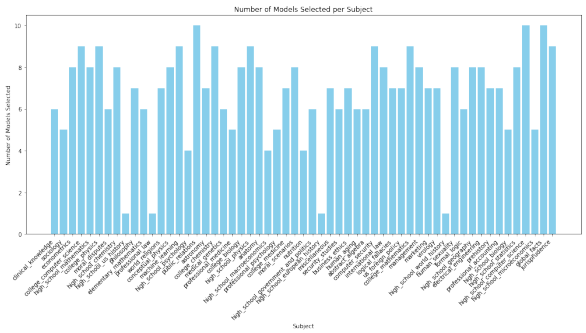

Figure 3: Distribution of selected models per subject. The variation in bar heights demonstrates how DFPE adapts its ensemble size to subject-specific requirements while maintaining efficiency.

A.3 Model Co-occurrence Analysis

To understand model relationships, we analyzed their co-occurrence patterns within clusters (Figure 4). Our analysis reveals several interesting patterns:

- **Strong Partnerships**: - Qwen family models (Qwen2.5-3B and Qwen2.5-7B-Instruct) show highest co-occurrence (52 instances) - Strong affinity between Qwen models and Llama-3.1-8B (51-52 co-occurrences) - Phi-3.5-mini frequently pairs with Qwen models (48-49 instances)
- **Complementary Groups**: - Models cluster into "specialists" and "generalists" - Lower co-occurrence patterns indicate complementary strengths
- **Model Independence**: - Some models show consistent independence - Suggests unique capabilities or specialization

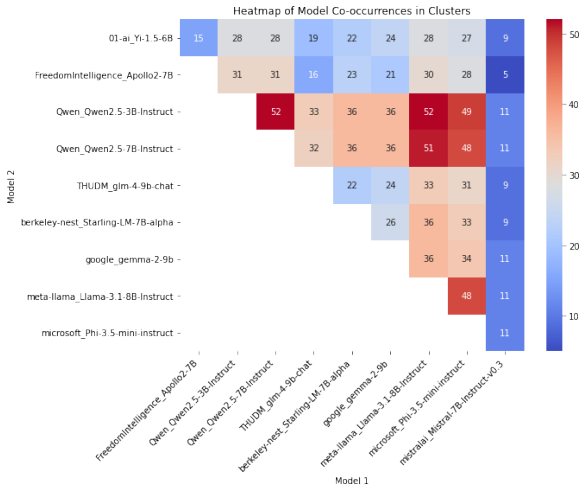

Figure 4: Heatmap of model co-occurrences within clusters. Cell values indicate frequency of model pairs being selected together. The diagonal is zero by definition. Higher values (darker colors) suggest stronger complementarity between models.

A.4 Practical Implications

These findings have several important implications for practitioners:

Resource Optimization: The balanced configuration offers a practical trade-off between performance and computational cost

Model Selection: Strong co-occurrence patterns can guide initial model selection when building new ensembles

Deployment Strategy: Subject-specific ensemble sizes suggest opportunities for dynamic resource allocation