Darwin Family: MRI-Trust-Weighted Evolutionary Merging for Training-Free Scaling of Language-Model Reasoning

A Systematic Framework Validated Across Evolved Models (4B–35B) and Public Reasoning Benchmarks

Anonymous Authors
Affiliation
email@domain

Abstract

We present **Darwin Family**, a framework for **training-free evolutionary merging** of large language models via gradient-free weight-space recombination. We ask whether frontier-level reasoning performance can be improved without additional training, by reorganizing latent capabilities already encoded in existing checkpoints.

Darwin introduces three key ideas: (i) a **14-dimensional adaptive merge genome** enabling fine-grained component- and block-level recombination; (ii) **MRI-Trust Fusion**, which adaptively balances diagnostic layer-importance signals with evolutionary search through a learnable trust parameter; and (iii) an **Architecture Mapper** that enables cross-architecture breeding between heterogeneous model families.

Empirically, the flagship **Darwin-27B-Opus** achieves 86.9% on GPQA Diamond, ranking **#6 among 1,252** evaluated models, and outperforming its fully trained foundation model without any gradient-based training. Across scales from **4B to 35B parameters**, Darwin models consistently improve over their parents, support recursive multi-generation evolution, and enable a **training-free evolutionary merge that combines Transformer- and Mamba-based components**. Together, the Darwin Family demonstrates that **diagnostic-guided evolutionary merging** is a practical and reproducible alternative to costly post-training pipelines for reasoning-centric language models.

1 Introduction

Recent large language models (LLMs) demonstrate strong reasoning performance, but achieving such capability has largely depended on expensive post-training pipelines, including instruction tuning, reinforcement learning, and large-scale distillation. While effective, these procedures require substantial compute and are often difficult to reproduce or adapt across settings. A growing body of evidence suggests that reasoning ability is not uniformly shaped by post-training.

Multiple studies show that supervised and instruction tuning can improve task-level accuracy while degrading reasoning faithfulness, robustness, or transfer, particularly in chain-of-thought settings (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023). Related work on prompting-based reasoning further indicates that reasoning can often be elicited without modifying model parameters, suggesting that core reasoning mechanisms are largely formed during pretraining (Wei et al., 2022; Zhou et al., 2023). Analysis at the level of internal representations provides converging support for this view. Layer-wise probing and structural diagnostics consistently show that different linguistic and reasoning functions are unevenly distributed across depth, with reasoning-critical computation localized to a subset of layers established during pretraining and relatively invariant under post-training or fine-tuning (Tenney et al., 2019; Ethayarajh, 2019; Hewitt and Manning, 2019).

More recent diagnostic and causal analyses reinforce the view that functional importance in neural networks is both localized and structurally constrained, motivating selective interventions over uniform parameter modification (Bau et al., 2020; Geiger et al., 2021). Together, these findings suggest that post-training primarily reorganizes surface behavior rather than reshaping the underlying reasoning circuitry. These observations raise a fundamental question: can reasoning performance be improved without further training, by reorganizing latent capabilities already en-

coded in pretrained checkpoints?

Model merging offers a promising training-free alternative by directly combining specialized models in weight space. Early approaches rely on static heuristics such as weight averaging or fixed linear combinations and are widely used for their simplicity (Wortsman et al., 2022; Ilharco et al., 2023). However, these methods often suffer from task interference, as they treat all parameters as uniformly mergeable despite substantial representational divergence between specialized models (Yadav et al., 2023). Recent work advances training-free model merging through selective parameter combination and sparsification, demonstrating that principled constraints can significantly improve merged performance without gradient-based training (Xu et al., 2024b).

Evolutionary approaches further automate the discovery of effective merge configurations, enabling gradient-free optimization over the merge space (Akiba et al., 2025). Nevertheless, most existing methods remain diagnostically blind, motivating the need for diagnostic-guided, adaptive training-free merging strategies.

2 Related Work

2.1 Knowledge versus Reasoning in LLMs

Recent studies increasingly indicate that knowledge acquisition and reasoning ability are partially decoupled in large language models. While instruction tuning and alignment procedures often improve final answer accuracy, they do not reliably improve multi-step reasoning fidelity and may degrade robustness or transfer in structured reasoning settings, particularly in chain-of-thought settings (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023). In contrast, prompting-based approaches such as chain-of-thought, least-to-most prompting, and self-consistency demonstrate that reasoning can often be elicited at inference time without modifying model parameters, suggesting that core reasoning mechanisms are largely formed during pretraining (Wei et al., 2022; Zhou et al., 2023). This perspective motivates approaches that reorganize or recombine existing representations rather than relying on additional training.

2.2 Diagnostic Probing and Functional Analysis

A long line of probing studies demonstrates that different layers of transformer models encode distinct linguistic and reasoning-related functions. Early work shows that pretrained language models recover a classical NLP processing pipeline across layers, with syntactic, semantic, and contextual abstractions emerging at different depths (Tenney et al., 2019; Ethayarajh, 2019; Hewitt and Manning, 2019; Rogers et al., 2020). Subsequent studies reveal that functional importance is unevenly distributed, motivating layer-aware and component-specific diagnostics rather than uniform parameter heuristics (Ethayarajh, 2019; Hewitt and Manning, 2019; Rogers et al., 2020). More recent work extends this perspective by identifying localized causal regions and neurons whose manipulation significantly affects model behavior, reinforcing the view that functional relevance in neural networks is both localized and structurally constrained (Bau et al., 2020; Geiger et al., 2021). Multilingual probing studies further show that such structural specialization generalizes across languages, supporting the use of diagnostic probes as a principled prior for guiding model reorganization (Li et al., 2024a).

2.3 Training-Free and Static Model Merging

Static model merging combines pretrained or fine-tuned models using fixed coefficients, such as weight averaging or task arithmetic. While effective for closely aligned models, these approaches often degrade performance when merging heterogeneous specialists due to representational incompatibility and interference (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023). Recent advances address these limitations by introducing training-free merging methods with structured sparsification, selective parameter alignment, or dual-space constraints, demonstrating that principled parameter selection can substantially improve merged performance without gradient-based training (Xu et al., 2024b). These works establish training-free model merging as a viable alternative to expensive multi-task training pipelines, while highlighting the importance of structural and

representational considerations.

2.4 Evolutionary Model Merging

Evolutionary optimization provides a natural framework for exploring merge configurations in a black-box, gradient-free setting. Classic work in neuroevolution demonstrates that evolutionary strategies can effectively optimize high-dimensional neural architectures without gradient information, motivating their application to large pretrained models. More recent work shows that evolutionary search can automatically discover high-performing model merging recipes that outperform manually designed heuristics, validating its applicability to model merging (Akiba et al., 2025). Nevertheless, most existing methods remain diagnostically blind, motivating the need for diagnostic-guided, adaptive training-free merging strategies.

2.5 Cross-Architecture and Hybrid Models

Recent architectural developments explore hybrid models that combine attention-based transformers with alternative sequence modeling mechanisms, such as state-space models, to improve efficiency and long-context performance. These hybrid architectures demonstrate that complementary inductive biases can be successfully combined within a single model, motivating cross-architecture recombination beyond traditional fine-tuning. Such advances provide architectural precedent for training-free cross-architecture merging, supporting the feasibility of recombining heterogeneous model components when equipped with appropriate alignment and selection mechanisms.

3 The Darwin Framework

Figure 1 provides a high-level overview of the Darwin framework, whose core design principle is to decouple diagnostic guidance from evolutionary exploration and reconcile them through an explicit fusion mechanism. Rather than performing gradient-based training, Darwin operates entirely in weight space, recombining frozen parent checkpoints through structurally informed merge decisions.

At a high level, Darwin proceeds as follows. Model-layer Response Importance (MRI) first estimates the functional relevance of individual parameter tensors using static statistics and lightweight probe-based responses, while a low-dimensional genome encodes candidate merge configurations explored via evolutionary search. These signals are combined through MRI-Trust Fusion to determine their relative influence, producing tensor-wise merge ratios that are applied by a training-free merge kernel to construct the final merged model. We now formalize this process, beginning with the problem formulation and parameter decomposition.

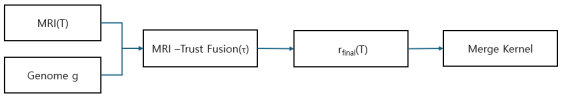

Figure 1: Overview of the Darwin framework.

3.1 Problem Formulation

Let two parent models A and B share a common pretrained base model θ_{base}. Their parameters are decomposed as

$$\begin{aligned} \theta_A &= \theta_{\text{base}} + \Delta_A \\ \theta_B &= \theta_{\text{base}} + \Delta_B \end{aligned} \tag{1}$$

where Δ_A and Δ_B represent model-specific deviations introduced by task specialization or distillation. Our objective is to construct a merged model θ_M that improves reasoning performance without any gradient-based training, solely by recombining Δ_A and Δ_B in weight space (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023; Xu et al., 2024b). Rather than treating all parameters uniformly, Darwin assigns tensor-specific merge ratios and optimizes them through a diagnostic-guided evolutionary process.

3.2 Merge Kernel and Parameter Recombination.

Each $r_{\text{final}}(T)$ denotes a scalar mixing coefficient shared across all elements of tensor T. Darwin constructs the merged tensor as

$$\begin{aligned} \theta_M(T) = \theta_{\text{base}}(T) \\ + (1 - r_{\text{final}}(T))\, \Delta_A(T) \\ + r_{\text{final}}(T)\, \Delta_B(T) \end{aligned} \tag{2}$$

where $\theta_{\text{base}}(T)$ denotes the shared pretrained base parameters, and $\Delta_A(T)$ and $\Delta_B(T)$ represent task-specific deviations from the base model.

3.3 Model-layer Response Importance (MRI)

Darwin introduces Model-layer Response Importance (MRI) as a diagnostic prior estimating the functional relevance of individual parameter tensors for reasoning behavior (Tenney et al., 2019; Ethayarajh, 2019; Hewitt and Manning, 2019; Rogers et al., 2020; Bau et al., 2020; Geiger et al., 2021; Li et al., 2024a). For a tensor T, MRI combines static tensor statistics and probe-based functional responses:

$$\mathrm{MRI}(T) = \alpha \cdot \mathrm{Static}(T) + (1-\alpha) \cdot \mathrm{Probe}(T) \quad (3)$$

$$r_{\mathrm{MRI}}(T) = \frac{\mathrm{MRI}_B(T)}{\mathrm{MRI}_A(T) + \mathrm{MRI}_B(T)} \quad (4)$$

The static term aggregates normalized entropy, variance, and capped ℓ_2-norm statistics, while the probe term measures cosine distance between reasoning-conditioned and generic activations induced by a small calibration set. The weighting parameter α controls the relative contribution of static and probe-based diagnostics and is fixed to $\alpha = 0.5$ in all experiments. MRI-derived ratios serve as a soft prior rather than a fixed merge rule and are subsequently fused with genome-derived ratios through MRI-Trust Fusion.

3.4 Architecture-Aware Tensor Alignment

For heterogeneous parent architectures, Darwin applies an Architecture Mapper that establishes tensor-level correspondences prior to numerical recombination. For a candidate pair of tensors (T_i^A, T_j^B), the mapper computes a compatibility score

$$\begin{aligned} \mathrm{Comp}(i,j) = \beta_1 \, &\mathrm{Type}(i,j) \\ + \beta_2 \, &\mathrm{Dim}(i,j) \\ + \beta_3 \, &\mathrm{Param}(i,j) \end{aligned} \quad (5)$$

where $\mathrm{Type}(i,j)$ captures functional role correspondence, $\mathrm{Dim}(i,j)$ measures dimensional compatibility, and $\mathrm{Param}(i,j)$ reflects parameter-shape similarity.

The coefficients $\beta_1 = 0.5$, $\beta_2 = 0.3$, and $\beta_3 = 0.2$ are fixed heuristic weights. Layer correspondences are established via constrained greedy matching under a minimum compatibility threshold, enabling limited cross-architecture recombination without retraining.

3.5 MRI-Trust Fusion and Genome-Based Control

A key design question is how much the merge should rely on diagnostics versus evolutionary exploration. Darwin resolves this using a single scalar parameter $\tau \in [0,1]$, which controls MRI trust. The final tensor-wise merge ratio is defined as

$$\begin{aligned} r_{\mathrm{final}}(T) = \tau \cdot &r_{\mathrm{MRI}}(T) \\ + (1-\tau) \cdot &r_{\mathrm{genome}}(T) \end{aligned} \quad (6)$$

where $r_{\mathrm{genome}}(T)$ is the genome-specified merge ratio for tensor T. When $\tau = 1$, the merge is fully determined by MRI diagnostics, while $\tau = 0$

Intermediate values of τ allow evolutionary optimization to correct diagnostic noise while retaining structured priors.

3.6 Genome and Evolutionary Optimization.

Each merge strategy in Darwin is represented by a 14-dimensional genome

$$\begin{aligned} g = \big(&\gamma,\ \alpha_{\mathrm{attn}},\ \alpha_{\mathrm{ffn}},\ \alpha_{\mathrm{emb}}, \\ &\rho_A,\ \rho_B,\ r_0,\ \ldots,\ r_5,\ \tau,\ \lambda\big) \end{aligned} \quad (7)$$

which controls global merge balance, component-level mixing ratios, sparsification densities, block-level specialization coefficients, MRI trust, and merge-kernel interpolation behavior. Evaluating a candidate genome requires instantiating a merged model and measuring its reasoning performance, making direct evolutionary search expensive. To address this challenge, Darwin employs a two-phase optimization strategy that separates structural screening from empirical evaluation.

4 Experiments and Analysis

4.1 Experimental Setup

We evaluate Darwin as a training-free reasoning enhancement framework, with primary emphasis on the flagship Darwin-27B-Opus and auxiliary experiments assessing generalization across scale, generation, and architecture. Parent models are selected to share

a common pretrained base whenever possible, following standard practice in homologous model merging.

Our primary benchmark is GPQA Diamond, a graduate-level multiple-choice benchmark targeting robust scientific reasoning under standardized inference settings (Rein et al., 2023). To assess broader reasoning generalization, we additionally evaluate on ARC-Challenge, which emphasizes multi-step symbolic and commonsense reasoning, and MMLU, which measures massive multitask language understanding across diverse academic subjects (Clark et al., 2018; Hendrycks et al., 2021).

We compare against (i) individual parent models, (ii) static training-free merging baselines such as uniform averaging and TIES-style merging (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023), and (iii) evolutionary merging without diagnostic guidance (Real et al., 2019; Such et al., 2017; Akiba et al., 2025). All results are averaged over multiple stochastic decoding runs using identical inference settings to ensure fair comparison.

4.2 Main Results: Darwin-27B-Opus (Primary Evidence)

This flagship result provides primary validation of the core claims of Darwin. Table 1 reports the main reasoning results for Darwin-27B-Opus on GPQA Diamond and ARC-Challenge, together with its parent models and representative baselines.

Darwin-27B-Opus achieves 86.9% on GPQA Diamond, ranking #6 among 1,252 evaluated models (as of 2026-04-22), and outperforms its strongest parent without any gradient-based training. Notably, Darwin surpasses several substantially larger, fully trained models while requiring only a small number of GPU hours for evolutionary search. These results demonstrate that frontier-level reasoning performance can be recovered, and even improved, through weight-space reorganization alone.

Compared to static merging methods, Darwin shows consistently higher accuracy and reduced variance, indicating greater robustness to representational interference. Compared to evolutionary merging without diagnostics (Real et al., 2019; Such et al., 2017; Akiba et al., 2025), Darwin achieves higher peak performance and more reliable convergence, suggesting that diagnostic guidance plays a critical role in navigating the merge space effectively.

Benchmark	Father	Mother	Avg / SLERP	Darwin-27B-Opus
GPQA-Diamond	0.855	0.862	0.861	**0.869**
ARC-Challenge	0.710	0.740	0.750	**0.779**
CommonsenseQA	0.770	0.776	0.778	**0.783**
TruthfulQA	0.772	0.775	0.776	**0.778**
HellaSwag	0.858	0.864	0.866	**0.870**
RACE	0.821	0.825	0.828	**0.831**
MMLU	0.754	0.782	0.768	**0.776**
Natural Questions	0.748	0.753	0.756	**0.760**
TriviaQA	0.711	0.718	0.719	**0.722**
Overall Average	∼0.767	∼0.776	∼0.775	**0.786 ± 0.040**

Table 1: Benchmark performance across configurations and nine-benchmark coverage.

We further analyze the impact of different merge kernels. Linear interpolation yields modest improvements but is susceptible to task interference. SLERP provides smoother interpolation during early exploration but consistently attains lower peak accuracy. In contrast, DARE-TIES achieves superior performance across all configurations. Its drop-and-rescale mechanism effectively mitigates destructive interference between parent models, validating its selection as the primary merge kernel in the Darwin framework.

4.3 Analysis of Learned Genome and Merge Dynamics

We next analyze the mechanisms underlying Darwin's performance gains, focusing on MRI-Trust Fusion, merge kernel selection, and genome structure. First, the learned trust parameter τ consistently converges to intermediate values ($\tau \approx 0.35$–0.55 across scales), indicating that neither pure diagnostic rules nor unconstrained evolutionary search is sufficient. Instead, Darwin benefits from an adaptive balance in which diagnostic priors guide search while evolutionary optimization compensates for diagnostic noise and inter-layer interactions.

Second, we compare merge kernels and find that DARE-TIES consistently outperforms linear interpolation and SLERP. While SLERP provides smoother exploration during early search, it suffers from lower peak accuracy. DARE-TIES effectively mitigates destructive interference between parent models through drop-and-rescale behavior, making it particularly well-suited for heterogeneous or highly specialized parents.

Configuration	τ setting	GPQA Diamond	CLIcK	Δ vs. full
No-MRI (genome only)	$\tau = 0$ (fixed)	84.4	69.2	$-2.5/-6.1$
MRI-only (static heuristic)	$\tau = 1$ (fixed)	85.6	72.4	$-1.3/-2.9$
Fixed-τ 0.7	$\tau = 0.7$ (fixed)	86.0	73.7	$-0.9/-1.6$
Full Darwin V6	$\tau = 0.556$	**86.9**	**75.3**	baseline

Table 2: Ablation of MRI-Trust fusion on Darwin-27B-Opus.

Finally, analysis of evolved genomes reveals stable structural patterns, including selective preservation of attention modules and stronger recombination in feed-forward components. These patterns recur across independent runs and model scales, suggesting that Darwin discovers architectural regularities, rather than exploiting properties unique to a single model.

4.4 Ablation Studies

To isolate the contribution of the MRI-Trust mechanism, we conduct a three-way ablation on the Darwin-27B-Opus configuration, varying only the τ fusion while holding all other genome parameters constant.

The ablation reveals two key findings. A summary of the ablation results across different τ settings is reported in Table 2, which compares genome-only merging, static MRI-based merging, fixed-τ variants, and the full adaptive Darwin configuration. First, MRI as a signal provides a clear performance benefit: using static MRI-based merging ($\tau = 1$) improves GPQA accuracy by +1.2pp relative to genome-only merging ($\tau = 0$). Second, adaptively learning the trust parameter further improves performance: the evolved τ variant achieves an additional +0.9pp gain over a fixed $\tau = 0.7$ setting. Overall, the full adaptive variant yields a +2.5pp improvement over the no-MRI baseline on GPQA, indicating that MRI-Trust Fusion is a primary contributor to the observed reasoning gains.

4.5 Generalization Beyond the Flagship Model

While Darwin-27B-Opus provides the primary empirical validation of the framework, we observe that the same evolutionary principles generalize across model scale, generation, and parent composition. Across all tested sizes (4B–35B), independently evolved Darwin models consistently converge to intermediate MRI-trust values and exhibit asymmetric recombination patterns, with stronger preservation of attention components and more aggressive recombination in feed-forward layers.

These structural regularities remain stable across independently evolved models, including recursive second-generation merges and mixed-architecture variants, suggesting that Darwin discovers scale-invariant merging principles rather than exploiting properties unique to a single model configuration. Detailed model-wise results and genome values are reported in Appendix B.2 and Table B.1, and full family overview is provided in and the full family overview is provided in Appendix B.6.

The framework also supports cross-architecture recombination. Darwin-4B-Genesis successfully merges Transformer-based attention with Mamba-style state-space feed-forward components without any retraining, outperforming both parents on targeted reasoning benchmarks. This case illustrates that Darwin can recombine complementary inductive biases across heterogeneous architectures, beyond fine-tuning variants of the same model family. Collectively, these models are not required to establish the effectiveness of Darwin, which is validated by Darwin-27B-Opus alone. Instead, they provide supporting evidence that the same diagnostic-guided evolutionary principles extend beyond a single flagship instance, generalizing across model scale, evolutionary depth, and architectural diversity. We emphasize that cross-architecture results are included as supporting evidence of extensibility rather than as a primary performance driver, with flagship validation carried by homologous merging. Detailed model-wise results and family-level comparisons are provided in Appendix B.6.

5 Limitations

Dependency on parent capabilities. Darwin improves upon its parent models by reorganizing latent capabilities acquired during pretraining, but it does not create new capabilities *ex nihilo*. If both parents lack a specific skill or knowledge domain, evolutionary merging alone cannot recover it.

Architectural and alignment constraints. At present, high-performing Darwin models require parents that share a common pretrained base. While limited cross-

architecture recombination is possible through architecture-aware alignment, general cross-base merging at scale remains an open challenge.

Search cost and verification scope. Although substantially cheaper than training or fine-tuning, Darwin's evolutionary search is not free and requires running a compact set of evaluations. In addition, while mid-scale models have been independently verified on public leaderboards, verification of the largest variants is ongoing.

6 Conclusion

We presented the Darwin framework and the Darwin Family of eight evolutionarily-merged language models spanning 4B to 35B parameters. Our primary contributions are the 14-dimensional adaptive genome (§3.6), the MRI-Trust Fusion formula with learnable τ (§3.5), and the Architecture Mapper enabling cross-architecture breeding (§3.4). The primary case study, Darwin-27B-Opus, is officially ranked #6 on the GPQA Diamond Leaderboard, outperforming its own Father Qwen3.5-27B by +1.4pp and other frontier models.

The Darwin Family establishes training-free evolutionary merging not as a niche technique for model ensemble averaging, but as a practical and reproducible pathway to frontier-scale reasoning capability at three to six orders of magnitude lower compute cost than conventional pretraining. By releasing all models, the V6 codebase, and the MRI tooling under Apache 2.0, we hope to enable broad independent verification and to catalyze a new research program on principled, diagnostic-guided weight-space optimization.

Promising directions for future work include extending Darwin to the 100B regime using sharded evaluation, improving cross-base alignment mechanisms, and combining Darwin with complementary test-time or inference-time interventions. weight-space optimization.

References

Takuya Akiba, Makoto Shing, Yu Tang, Qi Sun, and David Ha. 2025. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, and 1 others. 2020. Identifying and controlling important neurons in neural networks. In *International Conference on Learning Representations*.

Peter Clark, Brian Cowhey, Oren Etzioni, and 1 others. 2018. Think you have solved question answering? try arc. *arXiv preprint arXiv:1803.05457*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? In *EMNLP-IJCNLP*.

Atticus Geiger, Zhiwei Wu, David Lu, and 1 others. 2021. Causal abstractions of neural networks. In *Neural Information Processing Systems*.

Dan Hendrycks, Collin Burns, Steven Basart, and 1 others. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *NAACL*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, and 1 others. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Gu, M. Reid, and 1 others. 2022. Large language models are zero-shot reasoners. In *Neural Information Processing Systems*.

Dehua Li, Haoyan Zhao, Qing Zeng, and Mengnan Du. 2024a. Exploring multilingual probing in large language models. *arXiv preprint arXiv:2409.14459*.

Wenjing Li, Hao Gao, Mingqiao Gao, and 1 others. 2024b. Training-free model merging for multi-target domain adaptation. In *European Conference on Computer Vision*.

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. 2019. Regularized evolution for image classifier architecture search. In *AAAI Conference on Artificial Intelligence*.

David Rein, Benjamin L. Hou, Asa Cooper Stickland, and 1 others. 2023. Gpqa: A graduate-level google-proof question answering benchmark. *arXiv preprint arXiv:2311.12022*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology. *Transactions of the ACL*.

Kenneth O. Stanley and Risto Miikkulainen. 2002. Evolving neural networks through augmenting topologies. *Evolutionary Computation*.

Felipe Petroski Such and 1 others. 2017. Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks. *arXiv preprint arXiv:1712.06567*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Association for Computational Linguistics*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, and 1 others. 2023. Self-consistency improves chain-of-thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Mitchell Wortsman, Gabriel Ilharco, Samir Y. Gadre, and 1 others. 2022. Model soups: Averaging weights of multiple fine-tuned models. In *International Conference on Machine Learning*.

Zhen Xu, Kai Yuan, Hao Wang, and 1 others. 2024a. Training-free model merging under dual-space constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhen Xu, Kai Yuan, Hao Wang, and 1 others. 2024b. Training-free pretrained model merging. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Pratyush Yadav, Derek Tam, Leshem Choshen, and 1 others. 2023. Ties-merging: Resolving interference when merging models. In *Neural Information Processing Systems*.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Denny Zhou, Natalie Schärli, Le Hou, and 1 others. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*.

A Reproducibility

A.1 Data and reprodctibility site

Model collection.
https://huggingface.co/collections/FINAL-Bench/darwin-family. Distillation from Claude Opus 4.6 refers to supervised fine-tuning of an open-weight base model using reasoning traces generated via the public Claude Opus 4.6 API. No proprietary Claude model weights are used or distributed. All donor models employed in this work are publicly available community releases, and the distillation pipeline itself is fully reproducible given access to the Claude API.

Interactive demo / evolution studio:
https://huggingface.co/spaces/VIDraft/DARWIN-Evolution.

Darwin V6 codebase.
The codebase comprises approximately 13,771 lines across 15 Python files, including:

- `mri_extractor.py`
- `mergekit_integration.py`
- `parent_attribution.py`
- `calibration_data.py`
- `benchmarks.py`
- `live_engine.py`
- `live_blend.py`

All code is released under the Apache 2.0 license.

GPQA Diamond verification.
https://huggingface.co/datasets/Idavidrein/gpqa (Darwin-27B-Opus at #6, Darwin-31B-Opus at #11 as of 2026-04-22).

Community quantizations.
We identify several publicly available quantized variants:

- https://huggingface.co/bartowski/Darwin-27B-Opus-GGUF
- https://huggingface.co/bartowski/Darwin-31B-Opus-GGUF
- https://huggingface.co/bartowski/Darwin-35B-A3B-Opus-GGUF
- https://huggingface.co/mradermacher/Darwin-27B-Opus-i1-GGUF

A.2 Hyperparameters and Hardware.

All experiments were conducted on NVIDIA A100 or H100 GPUs. Runtime scales approximately linearly with model size, from approximately 1 hour for 4B models to approximately 5 hours for 35B models.

Hyperparameters:

- CMA-ES population size: 50.
- Generations (Phase 1): 20.

- Generations (Phase 2): 5–10 (model-size dependent).
- Mutation standard deviation (initial): 0.01.
- Mutation decay per generation: 0.95.
- Elite preservation: top-5 per generation.
- Crossover: SLERP in genome space.
- Random seed: fixed per question, MD5(Q) mod 2^{32}.
- Evaluation runs: $n = 30$ per candidate.
- Sampling temperature: 1.0 (maj@8 for David).
- Sampling top-p: 0.95.
- Sampling top-k: 64.

Hardware (per model):

- Darwin-4B-Opus: 1× A100-80GB, approximately 1 hour.
- Darwin-4B-David: 1× H100-80GB, approximately 1 hour.
- Darwin-4B-Genesis: 1× H100-80GB, 155 minutes.
- Darwin-9B-Opus: 1× H100-80GB, approximately 90 minutes.
- Darwin-27B-Opus: 1× H100-80GB, approximately 5 hours.
- Darwin-31B-Opus: 1× H100-80GB, approximately 134 minutes.
- Darwin-35B-A3B-Opus: 2× H100-80GB, approximately 5 hours.

A.3 Calibration Probe Set

The MRI calibration probe set comprises 123 samples across six categories, as summarized in Table 3, covering diverse reasoning and linguistic behaviors with an approximate Korean–English balance of 50:50, following prior probing and diagnostic analysis practices (Tenney et al., 2019; Ethayarajh, 2019; Hewitt and Manning, 2019; Rogers et al., 2020; Bau et al., 2020; Geiger et al., 2021; Li et al., 2024a).

Category	Samples	Purpose
REASONING	28	Multi-step chain-of-thought tasks such as arithmetic, logical deduction, and conditional inference
CODE	22	Programming tasks including Python coding, algorithm synthesis, and code understanding
LOGIC	18	Formal deduction tasks including syllogism and structured reasoning
MULTILINGUAL_KO	20	Korean language comprehension and cultural knowledge
MULTILINGUAL_EN	20	English baseline for multilingual comparison
GENERIC	15	Everyday queries used as baseline for cosine-distance anchoring
Total	123	Korean:English ≈ 50:50 by character

Table 3: MRI Calibration Probe Set Composition

All probe samples are available in the public Darwin V6 repository as the file `calibration_data.py`. Probe-conditional hidden states are computed via forward hooks at each transformer layer output; the `GENERIC` category serves as the baseline anchor for cosine-distance importance measurement described in Section 3.3.

B Genome Design and Evolutionary Optimization

B.1 Genome-Based Representation of Merge Strategies.

For conceptual clarity, we first describe the six core component-level parameters, which form a subset of the full 14-dimensional genome defined in Section 3.6. Darwin encodes merge strategies as a compact six-dimensional genome vector that balances expressiveness with computational tractability:

- $\gamma \in [0.3, 0.7]$: *global weight ratio* controlling the overall contribution of each parent.
- $\alpha_{\text{attn}} \in [0.2, 0.8]$: *attention-layer weight.*
- $\alpha_{\text{ffn}} \in [0.2, 0.8]$: *feed-forward weight.*
- α_{emb}: *embedding weight.*
- $\rho_A, \rho_B \in (0, 1]$: *sparsity rates* controlling parameter retention and rescaling.

Different genome profiles correspond to qualitatively distinct merge strategies. Balanced genomes ($\gamma \approx 0.5$, uniform component ratios) favor general-purpose fusion; asymmetric genomes (e.g., $\alpha_{\text{attn}} \gg \alpha_{\text{ffn}}$) are suited to reasoning-focused tasks where attention blocks carry disproportionate signal; and sparse genomes (low density values) promote stability under task interference. By mapping MRI-derived layer sensitivities into genome

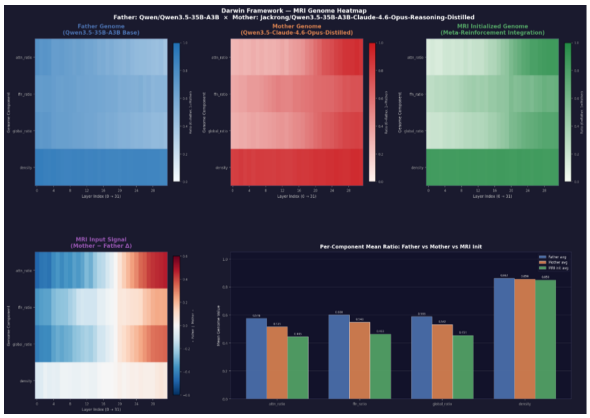

Figure 2: Darwin Framework—MRI Genome Heatmap. Comparative visualization of genome configurations. Top row: heatmaps for Father, Mother, and MRI-initialized genomes across layers and components. Bottom row: differential MRI signal and per-component mean ratios. MRI initialization produces structured adjustments that guide evolutionary search toward balanced and reasoning-oriented merge strategies.

initialization, Darwin ensures that the initial search population reflects functional layer specialization rather than arbitrary heuristics.

A central component is the use of MRI-based layer sensitivity probes to guide genome initialization. By analyzing parameter differences between donor models at each layer, MRI provides structured signals that inform initial merge ratios for attention, feed-forward, embedding, and global components. This warm-start strategy ensures that evolutionary search begins from a promising region of the genome space, accelerating convergence and improving merge quality relative to random initialization. These six parameters constitute the core subset of the full 14-dimensional genome, which is formally defined in Appendix B.2.

B.2 Full 14-Dimensional Genome Definition.

Each Darwin merge is controlled by a 14-dimensional genome consisting of component-level ratios, sparsification densities, block-level coefficients, and fusion parameters. Representative values and empirically stable ranges for each genome parameter, as evolved across model scales, are summarized in Table 4.

B.3 Parameters.

Darwin encodes merge strategies as a 14-dimensional genome vector composed of three groups of parameters.

Parameter	4B	27B	31B	Range
global ratio (γ)	0.5204	0.4893	0.4712	[0.47, 0.53]
attn ratio (α_{attn})	0.3195	0.1463	0.1890	[0.15, 0.32]
ffn ratio (α_{ffn})	0.8421	0.8768	0.9204	[0.84, 0.93]
embed ratio (α_{emb})	0.3508	0.3021	0.2894	[0.28, 0.36]
density A	0.8934	0.8507	0.8625	[0.85, 0.95]
density B	0.9011	0.9413	0.9228	[0.90, 0.95]
MRI trust (τ)	**0.4907**	**0.5557**	**0.3631**	**[0.36, 0.56]**
merge weight	0.3124	0.2783	0.3502	[0.28, 0.35]

Table 4: Evolved 14-D genome values across model scales.

Core parameters (6). The core component-level parameters control global and module-specific mixing behavior:

- **Global ratio:** $\gamma \in [0.05, 0.95]$ controls the overall balance between parent models.
- **Attention weight:** $\alpha_{\text{attn}} \in [0.05, 0.95]$ controls the contribution of attention blocks.
- **Feed-forward weight:** $\alpha_{\text{ffn}} \in [0.05, 0.95]$ controls the contribution of feed-forward networks.
- **Embedding weight:** $\alpha_{\text{emb}} \in [0.05, 0.95]$ controls token embedding and unembedding layers.
- **Sparsity rates:** $\rho_A, \rho_B \in [0.30, 1.00]$ control Bernoulli sparsification applied to parameter deltas of each parent.

MRI-derived block parameters (6). The block-level parameters are defined as follows:

- $r_0, \ldots, r_5 \in [0.05, 0.95]$: independent merge ratios assigned to six contiguous layer blocks identified by MRI.

These parameters capture coarse-grained dominance patterns across network depth, allowing different regions of the model to preferentially inherit characteristics from different parent models.

Meta-evolution parameters (2).

The parameter $\tau \in [0, 1]$ controls the interpolation between diagnostic MRI-based ratios and genome-driven ratios, as defined in Section 3.5. The parameter

merge_method_weight $\in [0, 1]$, denoted by λ in Section 3.6, controls interpolation between the DARE-TIES and SLERP merge kernels.

B.4 Scale-Invariant and Asymmetric Genome Patterns

Across model sizes ranging from 4B to 35B parameters, evolved Darwin genomes exhibit stable and recurring parameter ranges, particularly for attention preservation ratios, feed-forward recombination ratios, and MRI-trust values. As shown in Table 4, these parameters concentrate within narrow intervals across independently trained models, indicating that the evolutionary process discovers scale-invariant structural regularities rather than size-specific artifacts.

A salient pattern is the systematic asymmetry between attention and feed-forward components. Across all tested scales, Darwin consistently preserves a large fraction of attention parameters from the base (Father) model while aggressively recombining feed-forward layers from the specialized (Mother) model. This asymmetry aligns with prior probing and analysis studies suggesting that attention layers primarily mediate information routing and focus, whereas feed-forward networks encode task-specific computation and transformation (Tenney et al., 2019; Ethayarajh, 2019). Notably, this pattern emerges consistently across evolutionary runs and cannot be readily anticipated through manual design or uniform-ratio merging. Instead, it reflects an architectural regularity discovered by diagnostic-guided evolutionary search, reinforcing the view that Darwin reorganizes latent reasoning structure rather than introducing ad hoc parameter configurations.

B.5 Evolutionary Optimization Procedure

Darwin employs a two-phase evolutionary optimization strategy to efficiently search the merge-genome space while limiting the computational cost of full model instantiation.

Figure 3 provides a schematic overview of this evolutionary process, illustrating the iterative cycle of fitness evaluation, selection, crossover, and adaptive mutation that drives convergence toward a compact set of high-quality genomes.

In Phase 1, candidate merge genomes are evolved using a standard evolutionary optimization loop. Each generation evaluates candidate genomes using a lightweight proxy fitness score, followed by parent selection and variation through crossover and adaptive mutation. This phase is designed to rapidly filter structurally implausible or low-quality merge configurations without constructing full merged models.

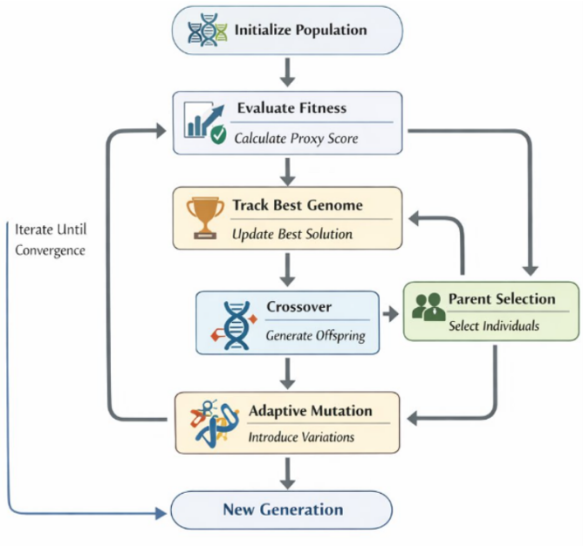

Figure 3: Evolutionary optimization process used in Phase 1 of Darwin. Candidate merge genomes are iteratively evolved via proxy-based fitness evaluation, selection, crossover, and adaptive mutation. This standard evolutionary loop continues until convergence and identifies a compact set of high-quality genomes for Phase 2 empirical evaluation.

In Phase 2, a small set of high-quality genomes identified in Phase 1 is instantiated as merged models and evaluated directly on reasoning benchmarks. Final selection is based on empirical performance under fixed inference settings.

B.6 Darwin Family Overview Across Scale and Generation

The table summarizes all released Darwin models and highlights the recurrence of intermediate MRI-trust values and asymmetric attention/FFN recombination patterns across scales. Table 5 provides an overview of the Darwin Family across model scale, evolutionary generation, and parent composition, reporting representative benchmark performance and notable properties for each released variant. For detailed model-level configurations and comparisons, we refer the reader to

Model	Gen	Parents (Father × Mother)	GPQA	Notable Property
Darwin-4B-Opus	1	gemma-4-E4B × Deckard	—	$\tau = 0.491$, 14-D genome
Darwin-4B-David	2	Darwin-4B-Opus × DECKARD-24B-D	85.0	Recursive evolution
Darwin-4B-Genesis	3	Darwin-4B-David × Qwen3.5-4B	∼60	Cross-architecture merge
Darwin-9B-Opus	1	Qwen3.5-10B base	—	Compact variant
Darwin-27B-Opus	1	Qwen3.5-27B × reasoning-distilled variant	**86.9**	**GPQA #6 (flagship)**
Darwin-31B-Opus	1	gemma-4-31B × TeichAI-distill	85.9	GPQA #11
Darwin-35B-A3B-Opus	1	Qwen3.5-35B-A3B × Jackrong	90.0	MoE / long context

Table 5: Overview of Darwin Family models across scale and generation.

Table 5.

C Architecture Mapper and Merge Kernels (Extended)

This appendix provides extended details on the merge kernels used in Darwin to recombine aligned parameter tensors after architectural alignment. The Architecture Mapper, which determines tensor-level correspondences between parent models, is introduced in Section 3.4 and is treated as a structural preprocessing step. In contrast, Appendix C focuses on the kernel-level operations that specify how aligned tensors are combined once correspondence has been established.

C.1 Role Separation: Mapper vs. Kernel

For clarity, we distinguish two components involved in cross-model recombination. The Architecture Mapper (Section 3.4) determines which parameter tensors from the two parent models can be meaningfully recombined by operating at the structural level, where it establishes tensor correspondences based on type, dimensionality, and positional compatibility. The mapper does not modify parameter values and performs no numerical combination. Given the matched tensor pairs identified by this alignment step, the merge kernel (described in this appendix) specifies how their parameter values are combined. Merge kernels operate at the parameter level and apply tensor-wise mixing ratios, sparsification rules, or interpolation schemes to produce the merged weights. This separation allows Darwin to cleanly decouple structural alignment from numerical recombination.

C.2 Evolutionary Optimization Procedure (Context)

Darwin employs a two-phase evolutionary optimization strategy to search over merge configurations without gradient-based updates. In Phase 1, candidate genomes are screened using a lightweight proxy objective to eliminate degenerate or structurally implausible configurations. In Phase 2, a small set of promising genomes is instantiated into merged models and evaluated directly on reasoning benchmarks. The merge kernels described in Appendix C.3 are invoked only after tensor alignment by the Architecture Mapper and ratio selection via MRI-Trust Fusion (Section 3.5).

C.3 DARE-TIES Merge Kernel

DARE-TIES (Drop-And-Rescale with Task-Interval Elimination) is the primary merge kernel used for final model construction in Darwin. Given aligned parent tensors and genome-specified mixing coefficients, this kernel operates by computing parameter deltas relative to a shared base model and applying Bernoulli sparsity masks to selectively retain informative components.

Specifically, for parent models A and B sharing a common base θ_{base}, Darwin computes $\Delta_A = \theta_A - \theta_{\text{base}}$ and $\Delta_B = \theta_B - \theta_{\text{base}}$, applies genome-controlled Bernoulli masks m_A and m_B to each delta, rescales the surviving entries to preserve expected magnitude, and then performs weighted recombination as

$$\begin{aligned}\theta_M = \theta_{\text{base}} \\ + \alpha_k \cdot (m_A \odot \Delta_A) \\ + (1 - \alpha_k) \cdot (m_B \odot \Delta_B)\end{aligned} \tag{8}$$

where $\alpha_k \in \{\gamma, \alpha_{\text{attn}}, \alpha_{\text{ffn}}, \alpha_{\text{emb}}\}$ denotes the genome-specified mixing weight for component k.

This drop-and-rescale procedure mitigates destructive interference between parent models while preserving complementary reasoning behaviors, and has been empirically observed to yield more stable performance than uniform averaging or linear interpolation. For this reason, Darwin prioritizes DARE-TIES for benchmark-driven fitness evaluation.

An overview of the DARE-TIES merge procedure is illustrated in Figure 4, which visually summarizes the drop-and-rescale operation applied to aligned parent parameter deltas before recombination.

C.4 SLERP Kernel (Exploration Phase)

SLERP (Spherical Linear Interpolation) is used as a lightweight alternative kernel dur-

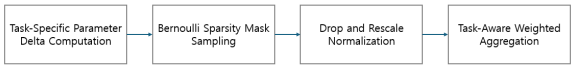

Figure 4: DARE-TIES Merge Kernel. The figure illustrates the DARE-TIES merge procedure applied after tensor alignment by the Architecture Mapper. Parameter deltas relative to a shared base model are sparsified via Bernoulli masking, rescaled to preserve expected magnitude, and combined using genome-specified mixing coefficients. The exact formulation of the merge kernel is provided in Appendix C.3.

Method	Genome dim	Diagnosis	Cross-arch	Multi-gen
TIES-Merging	–	none	no	no
DARE	–	sparsification	no	no
Model Soups	–	none	no	no
Task Arithmetic	–	none	no	no
Fisher Merging	–	Fisher info.	no	no
Model Breadcrumbs	–	sparse mask	no	no
Sakana EvoMerge	∼2/layer	none	partial	no
CycleQD	MAP-Elites	none	no	no
M2N2	evolvable splits	none	no	no
Darwin V5	2	MRI (proto)	no	no
Darwin V6 (ours)	**14**	**MRI**	**yes**	**yes**
Darwin-Genesis	**42/layer**	**MRI**	**yes**	**Gen-3**

Table 6: Comparison of Darwin with prior model merging methods.

ing early evolutionary exploration. By interpolating tensors along a hyperspherical path, SLERP enables smooth exploration of merge configurations with lower computational overhead. However, in Phase 2 evaluation, SLERP consistently underperforms DARE-TIES and is therefore not used for final model selection.

C.5 Summary

In summary, the Architecture Mapper (Section 3.4) determines tensor correspondences across heterogeneous parent models, while the merge kernels described here define the numerical rules for combining those tensors. Appendix C therefore complements the main text by detailing how aligned parameters are merged, not how alignment is established.

D Comparison with Prior Model Merging Methods

D.1 Overview of Comparison

This appendix situates Darwin within the broader landscape of model merging approaches, with particular emphasis on training-free methods. We focus on differences in assumptions, optimization structure, diagnostic usage, and extensibility, rather than raw performance, which is reported in the main text. Table 6 provides a comparative overview of Darwin and prior model merging methods along key dimensions, including genome dimensionality, diagnostic guidance, cross-architecture support, and multi-generation capability.

Darwin is the only prior-art-surveyed method that simultaneously (a) operates in a double-digit-dimensional genome, (b) integrates a functional-importance diagnostic signal into the merge kernel via a learnable τ parameter, (c) supports cross-architecture breeding, and (d) has been demonstrated across multiple evolutionary generations with heritable gains.

D.2 Static and Heuristic-Based Model Merging

Early model merging approaches rely on static, low-dimensional heuristics that combine pretrained or fine-tuned models using fixed coefficients. Representative examples include uniform weight averaging (Model Soups) (Wortsman et al., 2022) and linear vector arithmetic in weight space (Task Arithmetic) (Ilharco et al., 2023). These methods are attractive due to their simplicity and low computational cost, and they perform well when parent models are closely aligned in function and training history.

However, static heuristics implicitly assume that all parameters are equally mergeable. As later work demonstrates, this assumption often fails for heterogeneous specialist models, leading to representational interference and degraded performance (Yadav et al., 2023). TIES-merging partially addresses this issue by selectively trimming and rescaling parameter deltas, but the selection rules remain hand-designed and task-agnostic (Yadav et al., 2023). As a result, these approaches lack adaptivity to parent-specific structure and cannot easily generalize beyond closely related models.

D.3 Structured Training-Free Merging with Parameter Selection

Recent work improves training-free merging by introducing structured parameter selection and alignment constraints. Training-Free Pretrained Model Merging (Xu et al., 2024b) and Dual-Space Constraint Merging (Xu et al., 2024a) explicitly model consistency between weight space and activation space, demon-

strating that selective alignment significantly improves merged performance without gradient updates. Related work on multi-target domain adaptation further shows that principled training-free merging can rival data-sharing baselines when parent models share a common pretrained backbone (Li et al., 2024b).

While these methods represent a significant advance over static heuristics, they typically rely on fixed selection rules or optimization objectives that are not adaptive to downstream reasoning behavior. Moreover, they are usually limited to single-generation merging and do not naturally extend to iterative or evolutionary composition.

D.4 Evolutionary Model Merging

Evolutionary optimization offers a complementary perspective by treating model merging as a black-box search problem, optimizing merge configurations without gradient information. Classical work in neuroevolution demonstrates that evolutionary strategies can effectively search high-dimensional neural parameter spaces (Stanley and Miikkulainen, 2002; Real et al., 2019; Such et al., 2017). Building on this foundation, recent work shows that evolutionary search can automatically discover high-performing model merging recipes that outperform human-designed heuristics (Akiba et al., 2025).

However, existing evolutionary merging approaches are largely diagnostically blind. They typically operate over low-dimensional or uniform merge parameters and treat all components as symmetrically mutable, resulting in inefficient exploration and limited interpretability of evolved solutions.

D.5 Diagnostic-Guided Evolutionary Merging in Darwin

Darwin integrates the strengths of structured training-free merging and evolutionary optimization while addressing their limitations. Unlike static or rule-based methods (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023; Xu et al., 2024b,a; Li et al., 2024b), Darwin replaces fixed heuristics with an explicit adaptive genome that parameterizes merge behavior at multiple structural levels. Unlike prior evolutionary approaches (Stanley and Miikkulainen, 2002; Real et al., 2019; Such et al., 2017; Akiba et al., 2025), Darwin incorporates diagnostic priors that estimate functional relevance, allowing evolutionary search to focus on reasoning-critical components.

A key distinction is MRI-Trust Fusion, which adaptively balances diagnostic guidance and evolutionary exploration via a learnable trust parameter. This design enables Darwin to interpolate between heuristic-driven merging and unconstrained search, rather than committing to either extreme. As a result, Darwin supports multi-generation evolution, cross-architecture merging, and robust reasoning improvements without gradient-based training.

D.6 Summary Comparison

In summary, existing model merging approaches trade off simplicity, structure, and flexibility. Static heuristics are simple but fragile (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023); structured training-free methods are principled but inflexible (Xu et al., 2024b,a; Li et al., 2024b); evolutionary methods are flexible but inefficient without guidance (Stanley and Miikkulainen, 2002; Real et al., 2019; Such et al., 2017; Akiba et al., 2025). Darwin occupies a distinct point in this space by combining training-free operation, diagnostic selectivity, and evolutionary adaptability within a unified framework. A comprehensive survey of the broader model merging landscape can be found in (Yang et al., 2024).

E Failure Modes and Negative Results

Having situated Darwin relative to prior training-free model merging methods (Appendix D), we now analyze failure cases to clarify the operational boundaries of diagnostic-guided evolutionary merging.

Analysis of non-improving parent pairs reveals several recurring failure modes that are structural rather than incidental. Importantly, these cases do not contradict the effectiveness of Darwin, but instead clarify the conditions under which diagnostic-guided evolutionary merging is expected to succeed.

Lack of complementary specialization. In cases where both parent models exhibit highly similar capabilities and error patterns, evolutionary merging provides limited benefit. When neither parent contributes a distinct or

dominant capability, recombination primarily redistributes redundant structure rather than composing complementary functions, resulting in negligible or no improvement.

Severe representational misalignment. Some non-improving merges involve parent models whose internal representations are poorly aligned, even when nominally derived from the same base architecture. In such cases, weight-space recombination may disrupt reasoning-critical pathways faster than evolutionary optimization can recover them, leading to early saturation of gains.

Ambiguity in diagnostic signals. Darwin relies on MRI as a soft diagnostic prior rather than a ground-truth indicator. When diagnostic signals are weak, noisy, or inconsistent across layers—for example, when reasoning-relevant activations are diffused rather than localized—MRI guidance becomes less informative. Evolutionary search can partially compensate for such noise, but the resulting gains are typically smaller and less stable.

Search space saturation. Finally, some parent pairs already approach a local optimum for the targeted reasoning benchmarks. In these regimes, Darwin's evolutionary search converges quickly, but further improvement is constrained by the lack of latent complementary structure rather than by search inefficiency.

Together, these failure modes indicate that Darwin is most effective when applied to heterogeneous but compatible parent models with partially complementary reasoning structure. Failure cases therefore serve not as counterexamples, but as boundary conditions that clarify the operational scope of diagnostic-guided evolutionary merging.

F Resources and Community Adoption

All Darwin Family models, code, and MRI tooling are released under the Apache 2.0 license. As an indicator of community adoption, released Darwin models have accumulated substantial downloads across official and community-maintained distributions. Table 7 summarizes cumulative download counts as of 2026-04-22, covering official checkpoints as well as popular GGUF and third-party releases. For a detailed breakdown by model variant and distribution channel, we refer the reader to Table 7. As of April 2026, community downloads exceed 96,000 across official and quantized distributions. The combined download count exceeds 96,000 across all distribution channels, comparable to the adoption level of released open-source reasoning models from major labs. The substantial community-quantization activity (`bartowski`, `mradermacher`) further indicates that Darwin Family models are not merely benchmarked but actively deployed.

Model	Official	bartowski GGUF	mradermacher + others
Darwin-27B-Opus	∼14,000	∼22,000	∼8,500
Darwin-35B-A3B-Opus	∼9,000	∼14,500	∼6,000
Darwin-31B-Opus	∼6,500	∼8,000	∼3,000
Darwin-4B / 4B-David / 4B-Genesis / 9B (combined)	∼3,500	∼1,200	∼400
Family total	∼33,000	∼45,700	∼17,900

Table 7: Darwin Family community adoption (cumulative downloads).