WHEN THE WHOLE EXCEEDS THE PARTS: CONSTRUCTING MULTI-ABILITY LLMS VIA COMPONENT-WISE WEIGHTED FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Specialized language models often perform poorly on tasks that require capabilities from different domains. We introduce **Component-wise Weighted Fusion (CWF)**, a model-merging method that learns how much each expert contributes to individual feed-forward network (FFN) neurons and attention heads. An FFN coefficient is shared across the neuron's Gate, Up, and Down projections. Attention coefficients are shared by corresponding query and output slices, with group-averaged coefficients for shared key/value heads. The expert weights remain frozen while the coefficients are trained on mixed-domain data; the resulting weights form one dense model with no additional inference-time experts. We evaluate mathematics in Indonesian and Thai, coding with mathematics, and coding with multilingual knowledge using Llama and Qwen models at the 8B scale. Without target-language mathematics examples in mask training, CWF improves Indonesian and Thai MATH accuracy over the stronger input expert by 13.98 and 15.70 percentage points. On Qwen Code+Math, it matches the best reported merging baseline on each coding benchmark while retaining strong mathematical performance. We jointly evaluate target-language gains and source-domain retention. Component and learning-rate ablations favor joint FFN/attention fusion with separate learning rates, while mask and activation analyses expose heterogeneous expert contributions.

1 INTRODUCTION

Specialized LLMs target domains such as chemistry (Zhao et al., 2025) and medicine (Zhang et al., 2024). Cross-lingual transfer introduces a complementary challenge (Ruder et al., 2019; Hu et al., 2020): solving mathematics in a low-resource language may require abilities distributed across different experts. Sequential fine-tuning can introduce interference and forgetting (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017), motivating ways to combine experts without training a replacement model from scratch.

Multi-agent systems coordinate specialized models (Guo et al., 2024; Du et al., 2024; Chen et al., 2024), whereas model merging combines their parameters into one checkpoint (Matena & Raffel, 2022; Ilharco et al., 2023). Weight averaging (Wortsman et al., 2022), task-vector merging, and conflict-reduction methods such as TIES-Merging (Yadav et al., 2023) and DARE (Yu et al., 2023) provide different rules for combining expert updates. Adaptive and structured approaches also learn coefficients or exploit functional groups (Yang et al., 2024; Gu et al., 2025). The central design question in this work is therefore not whether to learn a mixture, but *which parameter slices should share its coefficients*.

We propose **Component-wise Weighted Fusion (CWF)** (Figure 1). Each FFN mixing distribution spans a neuron's complete Gate–Up–Down pathway. Attention ties Q/O coefficients within each query head and averages these mixtures for shared K/V heads. Mixtures vary across components and layers. Only mixing logits are trained; expert weights remain frozen. The result is folded into a standard dense checkpoint, so inference requires neither routing between experts nor a second model forward pass.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

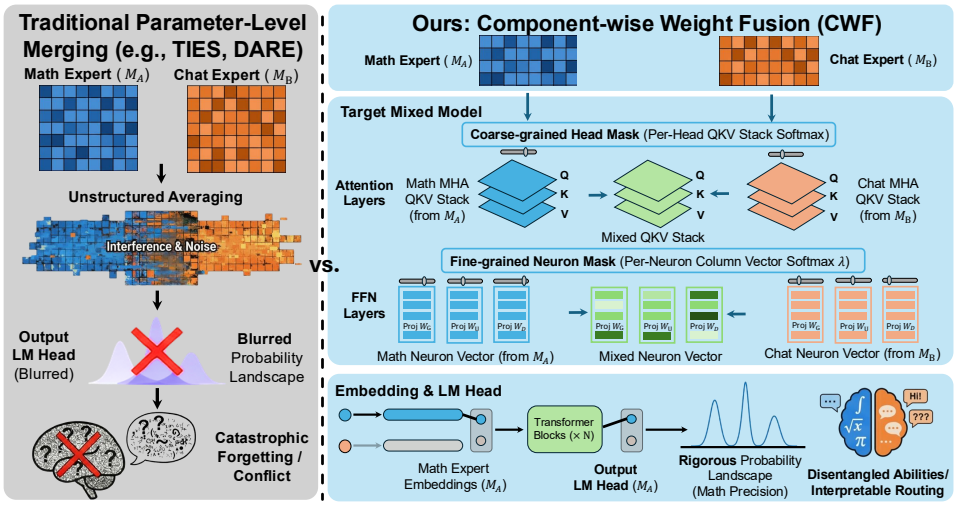

Figure 1: **Component-wise Weighted Fusion.** CWF shares mixing coefficients across the Gate, Up, and Down slices of an FFN neuron and across attention components, rather than assigning unrelated coefficients to individual parameters. A primary expert supplies the fixed embedding and output interface. Learned mixtures are materialized into the model weights before inference. Attention is schematic: Q/O share per-query-head coefficients; shared K/V use group averages (Section 3.1). Tables 1–4 evaluate retention and transfer.

Studies of FFN features and attention components motivate this parameterization (Geva et al., 2021; Tan et al., 2024; Zheng et al., 2024; Han et al., 2025). It permits different mixtures within a layer without per-weight learned variables. The learned coefficients also reveal which expert contributes more strongly to each component.

Our experiments address two complementary objectives. Cross-lingual mathematics tests capability composition: masks are trained on disjoint English mathematics and target-language conversation data, then evaluated on translated mathematics. Code+Math and Code+Multilingual test separate abilities within a single checkpoint. We evaluate source-expert retention alongside complementary-task gains.

CWF exceeds both input experts on target-language MATH and matches the strongest reported Qwen merging baselines on both coding benchmarks. Across the four expert pairs, we report capability gains together with source-expert retention and general-reasoning performance.

The paper makes three contributions: a coefficient-sharing scheme tied to FFN pathways and shared-KV attention; an evaluation of both composite tasks and separate abilities in one model; and an analysis of mask optimization and the resulting expert mixtures.

2 Related Work

Weight-space and adaptive merging. Model soups average fine-tuned weights (Wortsman et al., 2022); Task Arithmetic combines expert updates (Ilharco et al., 2023). TIES selects consistent update signs (Yadav et al., 2023); DARE sparsifies and rescales updates (Yu et al., 2023). ACT-Mat estimates weight-derived covariances (Abdel Hameed et al., 2026); Localize-and-Stitch identifies sparse task regions (He et al., 2025a). CWF instead optimizes convex mixtures of complete FFN pathways and attention components through mixed-domain training.

AdaMerging learns task- or layer-wise coefficients through entropy minimization (Yang et al., 2024). Differentiable Adaptive Merging (DAM) addresses language–mathematics and coding–multilingual fusion using column-wise scaling learned through expert-distribution matching and

regularization (Gauthier-Caron et al., 2024). CWF learns within-layer mixtures, while coupling each neuron's Gate–Up–Down slices rather than scaling columns independently. Its shared Q/O coefficients and shared-KV group averages extend this binding to attention. CWF jointly trains these mixtures with mixed-domain language-model cross-entropy.

Structured fusion and capability transfer. NeuronMerge classifies neurons into functional groups for merging (Gu et al., 2025). CWF instead learns each neuron's mixture jointly with attention coefficients, permitting different expert contributions within a functional group. We evaluate a paper-guided NeuronMerge$_1$ adaptation (Appendix A.6). Layer Swapping selects expert layers, with optional interpolation at transition layers, for cross-lingual transfer (Bandarkar et al., 2025). CWF addresses the same composition goal through continuous within-layer mixtures, allowing FFN and attention components at the same depth to favor different experts.

Functional analyses of LLMs. FFN neurons can encode task-relevant features (Geva et al., 2021; Song et al., 2024), and attention heads contribute to different computations (Clark et al., 2019; Todd et al., 2024); circuits may span components and layers (Yao et al., 2024; Ameisen et al., 2025). These findings motivate learning expert mixtures at the level of computational components. We analyze expert-mixture coefficients, selected activations, and neuron-removal results to characterize the fused model.

3 METHODOLOGY

CWF learns mixtures of corresponding expert components and materializes their parameters into one model. We suppress layer indices below. The notation permits N experts; experiments use $N = 2$.

3.1 FINE-GRAINED MODULAR FUSION

We consider experts fine-tuned from a shared backbone with corresponding parameter keys, tensor shapes, hidden dimension d, and FFN width d_{inter}. CWF matches components using the indices inherited from the shared backbone. The non-fused parameter owner is fixed for each expert pair, as described below.

Module 1: FFN neuron pathways. Use the row-input convention, with $\mathbf{X} \in \mathbb{R}^{s\times d}$. For expert i, let $\mathbf{W}_{G,i}, \mathbf{W}_{U,i} \in \mathbb{R}^{d\times d_{\text{inter}}}$ and $\mathbf{W}_{D,i} \in \mathbb{R}^{d_{\text{inter}}\times d}$. The kth neuron consists of the kth columns $\mathbf{w}_{G,i}^{(k)}, \mathbf{w}_{U,i}^{(k)} \in \mathbb{R}^d$ and the transpose $\mathbf{w}_{D,i}^{(k)} \in \mathbb{R}^d$ of the kth Down row. One coefficient is shared across all three slices:

$$\tilde{\mathbf{w}}_*^{(k)} = \sum_{i=1}^{N} \lambda_{F,i}^{(k)} \mathbf{w}_{*,i}^{(k)}, \qquad * \in \{G, U, D\}. \tag{1}$$

The fused neuron's output is

$$\mathbf{Y}_F^{(k)}(\mathbf{X}) = \left[\sigma(\mathbf{X}\tilde{\mathbf{w}}_G^{(k)}) \odot (\mathbf{X}\tilde{\mathbf{w}}_U^{(k)})\right] (\tilde{\mathbf{w}}_D^{(k)})^\top \in \mathbb{R}^{s\times d}, \tag{2}$$

where σ is the backbone's gating nonlinearity, and $\mathrm{FFN}(\mathbf{X}) = \sum_k \mathbf{Y}_F^{(k)}(\mathbf{X})$. The shared coefficient controls a complete input–output pathway. Because gating and multiplication are nonlinear, averaging these parameter slices is not equivalent to averaging the experts' FFN outputs.

Module 2: attention with shared key/value heads. The evaluated backbones use grouped-query attention (GQA), with $H_Q = 32$ query heads and $H_{KV} = 8$ key/value heads per layer. Let $\mathcal{H}_g$ be the four query heads sharing KV head g, and let $g(h)$ map query head h to its KV group. CWF learns one expert-mixture vector $\boldsymbol{\lambda}_A^{(h)}$ per *query* head. The Q and O slices for head h use this same vector. The shared K/V slices use the arithmetic mean of their group's four vectors:

$$\bar{\lambda}_{A,i}^{(g)} = \frac{1}{|\mathcal{H}_g|} \sum_{h\in\mathcal{H}_g} \lambda_{A,i}^{(h)}, \qquad |\mathcal{H}_g| = 4. \tag{3}$$

This average is taken *after* softmax normalization (Eq. 6), not over the logits. For expert i, Q/K/V slices have shape $d \times d_h$ and the O slice has shape $d_h \times d$, where d_h is the native head dimension. We construct

$$\tilde{\mathbf{W}}_*^{(h)} = \sum_{i=1}^{N} \lambda_{A,i}^{(h)} \mathbf{W}_{*,i}^{(h)}, \quad * \in \{Q, O\},$$
$$\tilde{\mathbf{W}}_*^{(g)} = \sum_{i=1}^{N} \bar{\lambda}_{A,i}^{(g)} \mathbf{W}_{*,i}^{(g)}, \quad * \in \{K, V\}. \tag{4}$$

There are no independently trained K, V, or O coefficients. Each K/V tensor is fused once and shared across its query heads, preserving the native GQA architecture. Suppressing positional transformations, native normalization, and the causal mask, head h contributes

$$\mathbf{A}^{(h)} = \mathrm{softmax}\left(\frac{(\mathbf{X}\tilde{\mathbf{W}}_Q^{(h)})(\mathbf{X}\tilde{\mathbf{W}}_K^{(g(h))})^\top}{\sqrt{d_h}} \right),$$
$$\mathbf{Y}_A^{(h)}(\mathbf{X}) = \mathbf{A}^{(h)}(\mathbf{X}\tilde{\mathbf{W}}_V^{(g(h))})\tilde{\mathbf{W}}_O^{(h)}. \tag{5}$$

The attention output sums these contributions. Llama and Qwen use the same coefficient-binding rule; Qwen3's additional Q/K normalization parameters remain fixed. For ordinary MHA, each KV group has one query head, reducing this rule to one shared Q/K/V/O mixture per head.

Convex mixing coefficients. For component k of type $* \in \{F, A\}$, CWF learns a vector of logits $\boldsymbol{\alpha}_*^{(k)} \in \mathbb{R}^N$:

$$\lambda_{*,i}^{(k)} = \left[\mathrm{softmax}(\boldsymbol{\alpha}_*^{(k)})\right]_i = \frac{\exp(\alpha_{*,i}^{(k)})}{\sum_{j=1}^{N} \exp(\alpha_{*,j}^{(k)})}. \tag{6}$$

Each slice is a convex combination of its expert counterparts. The group mean in Eq. 3 preserves nonnegativity and unit sum. For two experts, the coefficient selects a point on the line segment between the source components.

Non-fused parameters. Only seven projection families are fused: Q, K, V, O, Gate, Up, and Down. Embeddings, the LM head, and all normalization parameters—including Qwen3's Q/K norms—come from a designated primary expert. This is the math expert for Math+Indonesian/Thai and the code expert for both coding settings. The primary expert is fixed before mask optimization; its output weights receive the hidden states of the fused decoder.

3.2 Mask Training and Materialization

Let $\mathcal{D}$ mix examples from the experts' domains and let $\Phi = \{\mathcal{M}_F, \mathcal{M}_A\}$ contain the mixing logits. With all expert parameters Θ_{orig} frozen, training minimizes autoregressive cross-entropy:

$$\min_{\Phi} \ \mathbb{E}_{x \sim \mathcal{D}}\left[\ell_{\mathrm{CE}}(x; \Phi, \Theta_{\mathrm{orig}})\right]. \tag{7}$$

Gradients pass through the fused network to the coefficients, so its loss evaluates interactions among component mixtures rather than independently scoring isolated experts. After optimization, weighted slices are stored as ordinary model parameters. No input-dependent mask or additional expert forward pass is needed at inference. Training settings are summarized in Appendix A.3.

With N logits for every FFN neuron and query head, CWF stores

$$P_{\mathrm{logits}} = N \sum_l \left(d_{\mathrm{inter},l} + H_{Q,l}\right)$$

trainable scalars. For two experts this gives 919,552 logits for Llama-3.1-8B and 887,040 for Qwen3-8B-Base. Shared-KV averaging adds no parameters. The corresponding independent mixing degrees of freedom are half these counts; Appendix A.2 details the parameterization.

4 Experiments

We evaluate cross-lingual capability composition and joint expert performance across three ability-pair settings. Source models and merging baselines provide complementary reference points.

Table 1: **Mathematical reasoning across languages (%).** English and target-language scores jointly characterize retention and transfer. Bold denotes the highest reported score per column, including source models. NeuronMerge$_1$ denotes our paper-guided adaptation (Appendix A.6).

	Indonesian Fusion ($M_{math} + M_{id}$)				Thai Fusion ($M_{math} + M_{th}$)			
Method	**GSM En**	**GSM Id**	**MATH En**	**MATH Id**	**GSM En**	**GSM Th**	**MATH En**	**MATH Th**
M_{orig} (Base)	75.99	45.19	20.76	13.34	75.99	36.20	20.76	12.76
M_{math} (Expert)	**80.27**	23.11	**70.74**	29.14	**80.27**	28.54	**70.74**	11.98
M_{lang} (Target)	72.33	55.48	30.54	24.18	56.60	31.23	21.24	9.24
Seq-FT ($M_m \to L$)	74.32	27.56	62.08	22.62	73.52	28.82	64.32	13.36
Simple Avg.	61.50	21.62	21.76	15.54	50.22	22.75	20.21	12.42
TIES-Merging	63.05	48.15	26.52	19.14	60.52	**45.88**	28.94	18.96
DARE	77.91	28.23	62.16	38.66	77.17	37.15	53.24	20.78
ACT-Mat	71.19	41.17	47.20	25.40	68.99	34.38	58.99	18.10
L&S (dataless)	64.90	37.30	57.32	31.24	63.00	27.04	60.90	19.27
NeuronMerge$_1$	66.11	33.51	54.18	33.62	66.94	29.19	56.33	19.21
CWF (Ours)	79.35	**56.40**	64.40	**43.12**	79.53	39.47	60.52	**27.68**

Table 2: **General capability on English benchmarks (%).** Each cell reports the Math+Indonesian / Math+Thai configurations, both evaluated on the same English task. The pair denotes fusion settings, not evaluation languages.

Task	M_{orig}	M_{math}	M_{lang}	TIES	DARE	**CWF (Ours)**
WinoGrande	**63.9** / 63.9	47.2 / 47.2	61.0 / **65.0**	51.8 / 50.3	48.1 / 47.3	53.3 / 51.8
HellaSwag	64.3 / 64.3	25.1 / 25.1	**66.5** / **67.7**	51.6 / 51.2	26.2 / 24.8	47.8 / 45.4

4.1 Experimental Setup

Models and data. For cross-lingual mathematics, we use a custom Llama-3-8B math expert with Sahabat-AI and Typhoon Indonesian/Thai experts. GSM8K training examples (Cobbe et al., 2021) are mixed with sampled Indonesian conversation[1] or Thai instruction data (Phatthiyaphaibun, 2024). No target-language mathematics examples are used for mask training. We evaluate English GSM8K and MATH (Hendrycks et al., 2021) and their Indonesian/Thai translations produced with GPT-5 (OpenAI, 2025). Both fused models are evaluated on English WinoGrande (Sakaguchi et al., 2020) and HellaSwag (Zellers et al., 2019) to measure general capability outside the mask-training tasks. The translated-task prompts request step-by-step reasoning and a final answer in `\boxed{}`; OpenCompass performs mathematical answer postprocessing and evaluation.

For **Code+Math**, we merge the Code-1800steps and ODA-Math-460k experts of Qwen3-8B-Base (Yang et al., 2025). For **Code+Multilingual**, we use the MergeBench (He et al., 2025b) coding and multilingual experts of Llama-3.1-8B (Meta AI, 2024). Model identities are in Appendix A.1. Coding uses HumanEval+ (HE+) and MBPP+ enhanced-test pass@1 (Liu et al., 2023) on 164 and 378 problems. Qwen mathematics uses a fixed 256-problem GSM8K subset and 30 AIME 2026 problems. The multilingual score averages German, Spanish, French, and Russian across M-ARC, M-HellaSwag, and M-MMLU: 12 cells with equal weights, using normalized accuracy for ARC/HellaSwag and accuracy for MMLU. The coding settings test separate abilities within one checkpoint, whereas translated mathematics requires the two abilities within the same task.

Implementation and evaluation. In cross-lingual mathematics, logits $(0.5, -0.5)$ initialize the math/language mixture at $0.731 : 0.269$. Attention and FFN mask learning rates are 5×10^{-4} and 2.5×10^{-3}; the uniform-rate control uses 10^{-3}. The math expert supplies non-fused parameters

[1] https://huggingface.co/datasets/IzzulGod/indonesian-conversation

in these settings; the code expert supplies them in both coding settings. We use the GQA rule in Section 3.1 for Llama and Qwen.

Code and mathematical evaluations generate one response per problem: enhanced-test pass@1 is the fraction of code problems whose single response passes, while mathematics uses answer accuracy. For Qwen, generation limits are 8,192 tokens for HE+, MBPP+, and GSM, and 30,000 for AIME. All methods in a given main-table task use the same evaluation protocol. Appendix A.4 specifies task-level details and numerical reporting.

Baselines and comparability. We compare against source models, sequential fine-tuning (Seq-FT), Linear/Simple Avg., Task Arithmetic, TIES, DARE variants, ACT-Mat, dataless Localize-and-Stitch, and NeuronMerge. ACT-Mat and dataless L&S provide data-free comparisons. We evaluate a *paper-guided adaptation* of NeuronMerge$_1$ using the same expert pairs. ACT-Mat, dataless L&S, and NeuronMerge$_1$ are evaluated in all four expert pairs. The comparison covers each method's complete merging procedure. Qwen Task Arithmetic uses $\lambda = 1$. NeuronMerge$_1$ uses separate classification and coefficient-estimation examples; the cross-lingual and Qwen settings are detailed in Appendix A.6.

4.2 Main Results

Cross-lingual mathematical reasoning. Table 1 shows that CWF reaches 43.12% on MATH-Id and 27.68% on MATH-Th. Relative to the stronger input expert on each task (29.14% and 11.98%), these are gains of **13.98 and 15.70 percentage points**. On Indonesian fusion, CWF has the highest score among the compared merging and adaptation methods on all four metrics. Its 56.40% GSM8K-Id and 43.12% MATH-Id exceed the strongest competing method on each task by 8.25 and 4.46 points.

The structured baselines support this comparison beyond TIES and DARE. CWF exceeds ACT-Mat, dataless L&S, and NeuronMerge$_1$ on all four Indonesian metrics. Relative to the NeuronMerge$_1$ adaptation, its MATH-Id and MATH-Th gains are 9.50 and 8.47 points; its English and target-language GSM scores are also higher in both settings. Thai results remain a trade-off: TIES leads GSM8K-Th (45.88% versus 39.47%), while CWF leads MATH-Th by 6.90 points over DARE, the strongest competing method on that task. On English MATH, L&S and Seq-FT score 60.90% and 64.32%, above CWF's 60.52%.

Expert and general-ability retention. CWF's English GSM scores of 79.35% and 79.53% are within 0.92 and 0.74 points of the math expert. English MATH retention is less complete: 64.40%/60.52% versus 70.74%. General reasoning also decreases (Table 2); HellaSwag is 18.7/22.3 points below the respective language expert. Thus, better performance on translated mathematics is accompanied by reduced accuracy on some English control tasks.

Code+Math. On Qwen (Table 3), CWF ties TIES on HE+ at 84.76% and Linear on MBPP+ at 74.60%, the highest reported coding scores. It reaches 93.75% GSM and 63.33% AIME in the same checkpoint. The math expert scores 94.53% on GSM. On AIME, CWF solves 19/30 problems, compared with 18/30 for Linear and TIES. CWF exceeds ACT-Mat, dataless L&S, and NeuronMerge$_1$ on all four reported metrics. NeuronMerge$_1$ retains 93.36% GSM but obtains 79.88% HE+, 69.05% MBPP+, and 26.67% AIME, illustrating why both domains and both mathematical benchmarks are reported.

Task Arithmetic's Qwen coding scores at $\lambda = 1$ (1.83%/3.97%) contrast with 55.49%/55.29% on Llama. The Qwen math expert also scores higher than the code expert on HE+ (81.71% versus 79.27%).

Code+Multilingual. On Llama (Table 4), CWF obtains 59.76% HE+, 56.88% MBPP+, and 56.54% multilingual accuracy. The MBPP+ entry ties ACT-Mat. NeuronMerge$_1$ obtains 56.71%, 54.76%, and 53.32%, respectively; CWF's multilingual score is 3.22 points higher. The multilingual expert scores 55.98%. Relative to the stronger input expert on each metric, CWF gains 1.83, 2.12, and 0.56 points, respectively. On the multilingual macro-average, CWF scores 56.54% versus 56.46% for Linear.

Table 3: **Code + Math (Qwen3-8B).** Accuracy and pass@1 (%).

Method	HE+	MBPP+	GSM	AIME
Base	77.44	70.11	88.28	40.00
Code expert	79.27	73.28	91.41	46.67
Domain expert	81.71	68.52	**94.53**	56.67
Linear	79.27	**74.60**	92.19	60.00
Task Arithmetic	1.83	3.97	55.08	6.67
TIES	**84.76**	74.34	92.97	60.00
DARE-Linear	75.00	67.72	92.58	56.67
DARE-TIES	82.93	74.07	91.80	56.67
ACT-Mat	76.22	69.31	92.58	50.00
L&S (dataless)	81.71	71.69	85.55	40.00
NeuronMerge$_1$	79.88	69.05	93.36	26.67
CWF	**84.76**	**74.60**	93.75	**63.33**

Table 4: **Code + Multilingual (Llama-3.1-8B).** Accuracy and pass@1 (%).

Method	HE+	MBPP+	Multi.
Base	31.71	51.32	53.91
Code expert	57.93	54.76	53.14
Domain expert	36.59	53.44	55.98
Linear	52.44	54.76	56.46
Task Arithmetic	55.49	55.29	54.25
TIES	56.71	56.61	55.64
DARE-Linear	58.54	53.70	53.38
DARE-TIES	59.15	53.17	53.32
ACT-Mat	58.54	**56.88**	55.84
L&S (dataless)	56.71	53.97	55.24
NeuronMerge$_1$	56.71	54.76	53.32
CWF	**59.76**	**56.88**	**56.54**

HE+/MBPP+: 164/378 problems; GSM/AIME: 256/30 problems. Multi.: equal-weight average over 12 language–benchmark cells. Domain expert: Math (left), Multilingual (right). Bold marks column maxima, including ties. Evaluation and numerical reporting are detailed in Appendix A.4. NeuronMerge$_1$ is a paper-guided adaptation.

4.3 Ablation Studies

Output-interface selection. On Math+Thai, Copy Math Head obtains 60.52%/27.68% English/Thai MATH, versus 16.54%/12.08% for Copy Chat Head and 52.98%/24.02% for Weighted Fused Head (Table 5). Copy Math Head achieves the highest English and Thai accuracy among the three strategies. Its English MATH score is 43.98 points above Copy Chat Head. These results favor retaining the math expert's output interface in this setting (Appendix B.3).

Component-specific learning rates. Separate FFN/attention rates improve GSM8K-Id from 55.15% to 56.40% and MATH-Id from 39.56% to 43.12% (Table 6). Both mask families are optimized jointly, with the FFN learning rate five times the attention rate. Table 7 additionally compares the two component families.

Table 5: Output-interface strategies on Math+Thai (MATH accuracy, %).

LM Head Strategy	**MATH (En)**	**MATH (Thai)**
(I) Copy Math Head (Ours)	**60.52**	**27.68**
(II) Copy Chat Head	16.54	12.08
(III) Weighted Fused Head	52.98	24.02

Table 6: Uniform versus separate mask learning rates on Indonesian mathematics (accuracy, %).

Strategy	**GSM8K (Indo)**	**MATH (Indo)**
Uniform LR	55.15	39.56
Separate LR (Ours)	**56.40**	**43.12**

Uniform: $\eta_{ffn} = \eta_{attn} = 10^{-3}$. Separate: $\eta_{attn} = 5 \times 10^{-4}$, $\eta_{ffn} = 2.5 \times 10^{-3}$.

Joint component fusion. Table 7 compares FFN-only, attention-only, and joint fusion on Math+Indonesian. Joint CWF improves GSM8K-Id and MATH-Id by 12.04 and 25.74 points over the stronger restricted configuration on each task. It also has higher English accuracy than either restricted configuration.

Table 7: Component-only fusion on Math+Indonesian (accuracy, %).

Method	GSM-En	GSM-Id	MATH-En	MATH-Id
FFN only	67.34	44.36	54.60	14.72
Attention only	62.64	26.40	43.84	17.38
CWF	**79.35**	**56.40**	**64.40**	**43.12**

5 Discussion

5.1 Training Cost and Capability Composition

Method	# Models	Opt. states	VRAM
Full FT	1	∼8B params	≈78 GB
CWF	2	Masks only	≈60 GB

Table 8: Peak VRAM on Llama-3-8B: batch size 2, gradient accumulation 4. CWF loads two frozen experts and trains only fusion masks (Appendix B.4).

CWF uses approximately 18 GB less peak VRAM in this experiment (Table 8). It avoids optimizer states for expert parameters while storing both experts and differentiating through the fused network to train the masks.

Cross-lingual evaluation tests capability composition after mask learning on English mathematics and target-language conversation. We measure answer accuracy for mathematical questions written in the target language. Coding settings instead test separate abilities in one checkpoint.

5.2 Analysis of Learned Fusion Masks

The weights in Figures 2 and 3 denote the math expert's contribution w; the language expert contributes $1 - w$.

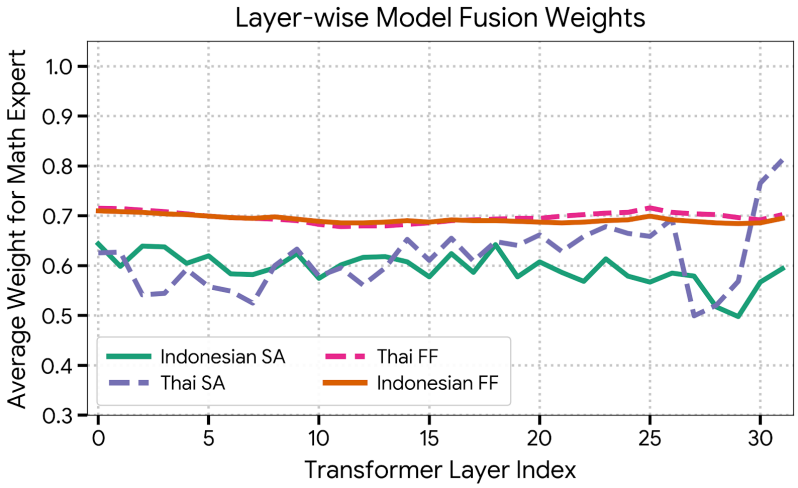

Figure 2: Layer-wise math-expert weights in Thai and Indonesian fusion. FFN averages remain near 0.7, while attention averages vary more across layers.

Component and layer variation. FFN averages are relatively stable, whereas attention weights show larger layer-to-layer variation. Math+Thai has larger differences between the two component types near the model's endpoints than Math+Indonesian. Component-specific learning rates control the adaptation of the two mask families under the shared training objective.

Within-layer heterogeneity. Most layer-wise FFN medians are near 0.7, below the 0.731 initialization; individual neurons show a broad spread of weights (Figure 3). This is heterogeneous parameter allocation. A near-one coefficient means the fused parameter slice is close to the math expert at that position.

What the granularity permits. CWF can represent within-layer allocations hidden by layer averages. Under identical fixed-parameter ownership, tying its coefficients within each layer recovers a layer-wise mixture. The nesting describes the choices the parameterization can represent.

5.3 Micro-Analysis of Token Activations

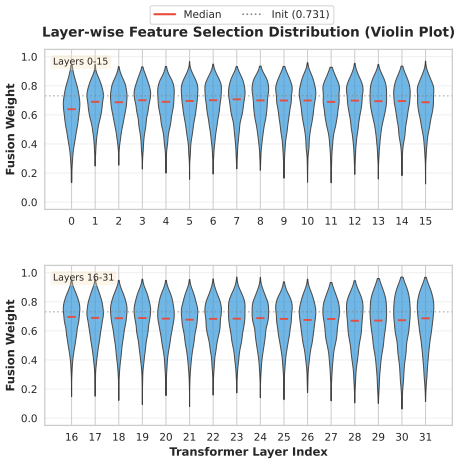

Figure 3: FFN fusion-weight distributions for the Math+Indonesian merger across all 32 layers. Red bars mark medians; dashed lines mark initialization (0.731). An enlarged view is in Appendix B.6.

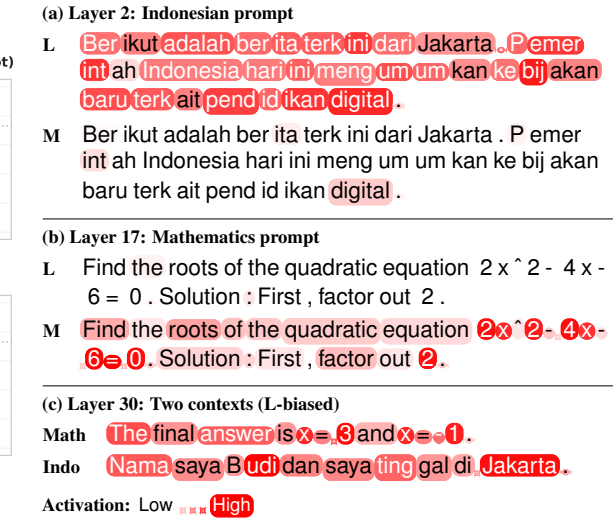

Figure 4: Token activations in the Math+Indonesian merger. Language-biased (L) and math-biased (M) neurons are shown for (a) Indonesian input at layer 2, (b) mathematical input at layer 17, and (c) both contexts for an L-biased neuron at layer 30. Darker red denotes higher displayed activation intensity.

Figures 3 and 4 analyze the Math+Indonesian merger. At each illustrated layer, the largest math-expert coefficient selects the math-biased neuron and the smallest selects the language-biased neuron. Activations are recorded at the input to `down_proj`, after the SwiGLU product. The layer-2 language-biased example responds across the Indonesian prompt. The layer-17 math-biased example has prominent responses on numbers and variables. At layer 30, the language-biased example responds in both contexts. The selected neurons therefore exhibit depth-dependent activation patterns alongside their different expert preferences.

Expert preference and input-dependent activity. An expert coefficient determines the contribution of that expert's parameter slices. The layer-30 example illustrates how a language-biased component can participate in both mathematical and Indonesian contexts. The learned mixtures are static after merging, while activations respond to the input. This distinction connects the visible allocation of expert parameters to the behavior of the resulting fused network.

Intervention evidence. Appendix B.5 compares targeted and random neuron removal on 500 GSM8K questions. Removing math-biased neurons yields mean accuracies 9.76 and 16.13 points below random removal at the 1% and 5% ratios, respectively. These interventions link coefficient-based neuron selection to mathematical performance: the targeted sets are more consequential than the reported random sets at the same removal ratios. The activation examples provide a complementary view of the selected components' responses.

6 Conclusion

CWF combines complementary experts through learned mixtures of FFN neuron pathways and GQA attention components. It trains only mixing coefficients and folds them into one dense checkpoint. The cross-lingual experiments improve target-language MATH accuracy over both input experts, while the coding experiments combine coding with mathematics or multilingual knowledge. Component and learning-rate ablations favor the joint configuration; mask distributions and neuron-removal experiments connect expert preferences with mathematical performance. The English control tasks reveal the accompanying retention trade-offs. CWF provides a parameter-efficient way to exploit Transformer structure for capability composition with ordinary single-model inference.

IMPACT STATEMENT

This work studies methods for combining specialized language models. A merged model can inherit capabilities, errors, and biases from its source experts. Improved benchmark accuracy does not establish suitability for high-stakes deployment. Cross-lingual evaluation additionally depends on translation fidelity and answer normalization. Applications and artifact releases should respect the licenses and use restrictions of the source models and datasets.

AI USE STATEMENT

Generative AI tools were used for related-work retrieval and discovery, translation of evaluation data, assistance with experiment implementation and debugging, feedback on experimental design and methodology, exploration and interpretation of fusion experiments, and drafting, editing, and formatting manuscript sections. Authors are responsible for reviewing the resulting text, claims, code, and results before submission.

REPRODUCIBILITY STATEMENT

Section 3 specifies the FFN and GQA coefficient-sharing rules, training objective, and materialization procedure. Appendix A describes the expert pairs, parameter ownership, training setup, and evaluation metrics. Appendix A.6 details the NeuronMerge$_1$ adaptation. Additional analyses report component-only fusion, output-interface strategies, initialization sensitivity, and neuron-removal results.

REFERENCES

Marawan Gamal Abdel Hameed, Derek Tam, Pascal Jr Tikeng Notsawo, Colin Raffel, and Guillaume Rabusseau. Model merging via data-free covariance estimation. *arXiv preprint arXiv:2604.01329*, 2026. URL https://arxiv.org/abs/2604.01329.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 6, 2025.

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. Layer swapping for zero-shot cross-lingual transfer in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vQhn4wrQ6j.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *ICLR*, 2024. URL https://openreview.net/forum?id=EHg5GDnyq1.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=zj7YuTE4t8.

Thomas Gauthier-Caron, Shamane Siriwardhana, Elliot Stein, Malikeh Ehghaghi, Charles Goddard, Mark McQuade, Jacob Solawetz, and Maxime Labonne. Merging in a bottle: Differentiable adaptive merging (DAM) and the path from averaging to automation. *arXiv preprint arXiv:2410.08371*, 2024. URL https://arxiv.org/abs/2410.08371.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Wangyun Gu, Qianghua Gao, Zhang Li-Xin, Xu Shen, and Jieping Ye. NeuronMerge: Merging models via functional neuron groups. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9015–9037, 2025. doi: 10.18653/v1/2025.findings-acl.471. URL https://aclanthology.org/2025.findings-acl.471/.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Senyu Han, Hongchuan Zeng, Kai Yu, and Lu Chen. Heads up! large language models can perform tasks without your instruction via selective attention head masking. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=x2Dw9aNbvw.

Yifei He, Yuzheng Hu, Yong Lin, Tong Zhang, and Han Zhao. Localize-and-stitch: Efficient model merging via sparse task arithmetic. *Transactions on Machine Learning Research*, 2025a. URL https://arxiv.org/abs/2408.13656.

Yifei He, Siqi Zeng, Yuzheng Hu, Rui Yang, Tong Zhang, and Han Zhao. MergeBench: A benchmark for merging domain-specialized LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025b. URL https://openreview.net/forum?id=rw50iUoyLu.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4411–4421. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hu20b.html.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.01210.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. URL https://arxiv.org/abs/1609.07843.

Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2025-09-24.

OpenCompass Contributors. OpenCompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Wannaphong Phatthiyaphaibun. Han instruct dataset, July 2024. URL https://doi.org/10.5281/zenodo.13145164.

Qwen Team. Qwen3-8B-Base: Official model configuration, 2025. URL https://huggingface.co/Qwen/Qwen3-8B-Base/blob/main/config.json.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. Unsupervised cross-lingual representation learning. In Preslav Nakov and Alexis Palmer (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 31–38, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-4007. URL https://aclanthology.org/P19-4007/.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. Does large language model contain task-specific neurons? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7101–7113, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.403. URL https://aclanthology.org/2024.emnlp-main.403/.

Shaomu Tan, Di Wu, and Christof Monz. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6506–6527, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.374. URL https://aclanthology.org/2024.emnlp-main.374/.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nZP6NgD3QY.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YVXzZNxcag.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *arXiv preprint arXiv:2311.03099*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Jin-Fang Hu, Zhiyuan Liu, and Bowen Zhou. Ultramedical: Building specialized generalists in biomedicine. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 26045–26081. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2dfc26ce9039f00eee4aba0c54931e46-Paper-Datasets_and_Benchmarks_Track.pdf.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, et al. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4), 2025.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*, 2024.

A Implementation and Evaluation

A.1 Expert Models

All settings combine two experts derived from a shared backbone. Table 9 lists the model identities used in the experiments.

Table 9: Backbones and expert models.

Role	Model
Cross-lingual base	Llama-3-8B
Cross-lingual math	Custom fine-tuned Llama-3-8B math expert
Indonesian / Thai expert	Sahabat-AI / Typhoon
Qwen base	`Qwen/Qwen3-8B-Base`
Qwen code	`LumosJiang/Qwen3-8B-Base-SFT-AM-Thinking-v1-Distilled-Code-1800steps`
Qwen math	`OpenDataArena/Qwen3-8B-ODA-Math-460k`
Llama base	`meta-llama/Llama-3.1-8B`
Llama code	`MergeBench/Llama-3.1-8B_coding`
Llama multilingual	`MergeBench/Llama-3.1-8B_multilingual`

A.2 Coefficient Binding and Non-Fused Parameters

FFN storage convention. The main text uses row inputs, so its Gate/Up columns correspond to rows of the stored `gate_proj`/`up_proj` weight matrices. Its Down rows correspond to columns of stored `down_proj`. Each neuron's Gate row, Up row, and Down column share the same expert coefficients.

Shared-KV construction. For every layer, CWF learns two logits per query head and applies softmax across the experts. Head h uses the resulting vector for both its Q output rows and its O input columns. For the four query heads in $\mathcal{H}_g$, the shared K and V rows both use $\frac{1}{4}\sum_{h\in\mathcal{H}_g}\boldsymbol{\lambda}_A^{(h)}$. The mean is formed from normalized coefficients, not logits:

$$\frac{1}{4}\sum_{h\in\mathcal{H}_g}\mathrm{softmax}(\boldsymbol{\alpha}_A^{(h)}) \quad \text{is the implemented rule, not} \quad \mathrm{softmax}\left(\frac{1}{4}\sum_{h\in\mathcal{H}_g}\boldsymbol{\alpha}_A^{(h)}\right).$$

These two expressions need not be equal. The implemented mean remains nonnegative and sums to one. Independent Q/O mixtures can differ within a KV group while K/V remain shared; no separate K, V, or O logits are learned. The same binding is used for the Llama and Qwen experiments.

Table 10: Backbone dimensions and mask-logit counts. Dimensions follow the model report/configuration (Meta AI, 2024; Qwen Team, 2025). Counts follow the two-expert coefficient-binding rule.

Field	Llama-3.1-8B	Qwen3-8B-Base
Transformer layers	32	36
Hidden dimension	4,096	4,096
FFN width	14,336	12,288
Query heads / KV heads	32 / 8	32 / 8
Head dimension	128	128
FFN mixing logits	917,504	884,736
Attention mixing logits	2,048	2,304
Total mixing logits	919,552	887,040
Independent mixture degrees of freedom	459,776	443,520

There is one redundant common logit shift per expert-mixture vector, giving $N-1$ effective mixture degrees of freedom for N stored logits. KV-group averaging does not increase this count.

Non-fused parameters. Only the seven projection families Q, K, V, O, Gate, Up, and Down are merged. All other model parameters are copied from the primary expert, including embeddings, the LM head, decoder/final normalization, and Qwen3 Q/K normalization.

With fixed coefficients, materialization stores the same weighted projection slices used in the fused forward computation. It introduces neither extra attention heads nor dynamic expert routing. Rounding can introduce small numerical differences when weights are materialized in finite precision.

Table 11: Primary expert supplying all non-fused model parameters.

Setting	Primary expert
Math+Indonesian / Math+Thai	Math expert
Qwen Code+Math	Code expert
Llama Code+Multilingual	Code expert

A.3 Mask-Training Settings

For cross-lingual mathematics, mask training mixes GSM8K training examples with Indonesian conversation or Thai instruction examples. Target-language mathematics is excluded from this mixture. Expert weights remain frozen. Mask logits are initialized at $(0.5, -0.5)$, favoring the math expert after softmax. Attention and FFN learning rates are 5×10^{-4} and 2.5×10^{-3}, respectively; the uniform-rate comparison uses 10^{-3} for both. The primary expert supplies embeddings, the output head, and normalization parameters throughout the default CWF procedure.

The Code+Math and Code+Multilingual combinations use the same coefficient parameterization and mixed-domain cross-entropy objective. The code expert supplies their non-fused parameters, as specified in Section 3.1.

A.4 Evaluation and Score Reporting

Single-response evaluation. Code and mathematics benchmarks use one generated response per problem for each main-table method. Code pass@1 is the fraction of responses passing the enhanced test suite; mathematics accuracy is the fraction with a correct final answer. For c successful responses among n problems, the underlying percentage is $100c/n$. Scores are reported as percentages.

Llama coding uses the same 164/378-problem benchmarks and one response per problem. Its multilingual metric is the equal-weight average in Appendix A.7. The main tables report single-evaluation results.

Table 12: Qwen evaluation protocol. Limits refer to newly generated tokens.

Task	Questions	Max. generation	Metric
HumanEval+	164	8,192	Enhanced-test pass@1
MBPP+	378	8,192	Enhanced-test pass@1
GSM8K fixed subset	256	8,192	Answer accuracy
AIME 2026	30	30,000	Answer accuracy

General-capability evaluation. Both Math+Indonesian and Math+Thai models are evaluated on English WinoGrande and HellaSwag. Each pair in Table 2 lists the Math+Indonesian and Math+Thai models, in that order, evaluated on the same English benchmark. These tests measure general capability outside the English-math/target-language-conversation training mixture. Multilingual ARC and HellaSwag use normalized accuracy, and MMLU uses accuracy.

A.5 Translated Mathematical Evaluation

GSM8K and MATH test problems were translated into Indonesian and Thai using GPT-5. Evaluation prompts request step-by-step reasoning and a final answer enclosed in `\boxed{}`. OpenCompass (OpenCompass Contributors, 2023) performs mathematical answer postprocessing and evaluation. No target-language mathematics examples are included in mask training. Mathematical accuracy measures final-answer correctness for questions in the target language.

A.6 Baseline Settings and Comparison Scope

Weight-space baselines. We compare source models, Seq-FT, weight interpolation, Task Arithmetic, TIES, DARE variants, ACT-Mat, and dataless Localize-and-Stitch. Qwen Task Arithmetic uses task-vector scale $\lambda = 1$. The methods can differ in data access and treatment of embeddings, output heads, and normalization parameters.

NeuronMerge$_1$: paper-guided adaptation. We implement a *paper-guided adaptation* of NeuronMerge$_1$ (Gu et al., 2025) for our expert pairs. The adaptation is applied to the expert pairs listed in Table 9. Table 13 lists the functional categories for each pair.

Table 13: Functional categories used by the NeuronMerge$_1$ adaptation.

Setting	Categories
Math+Indonesian	General, Indonesian, Math
Math+Thai	General, Thai, Math
Qwen Code+Math	General, Code, Math
Llama Code+Multilingual	General, Code, Multilingual

Neuron classification. Each category supplies 300 examples. General examples are taken from WikiText-2 (Merity et al., 2016); task examples are taken from the corresponding task-training or independent calibration texts. Inputs are truncated to 512 tokens. For each neuron in every MLP layer of the base model, an example's score is its maximum absolute activation over tokens. The labels of the ten highest-scoring examples vote for the neuron's functional category.

Coefficient estimation and weight construction. A separate set contains 30 examples from each of the two task categories and does not overlap the classification set. Two coefficients are estimated for the General neurons in each layer. These coefficients are layer-specific rather than a single manually chosen global ratio. In Math+Indonesian and Math+Thai, the General-group update uses the two expert weight differences. Writing $\theta_{0,l,G}$ for the base-model slices of the General group in layer l, the construction is

$$\theta_{\mathrm{merged},l,G} = \theta_{0,l,G} + \sum_{i=1}^{2} \beta_{l,i}\big(\theta_{i,l,G} - \theta_{0,l,G}\big). \tag{8}$$

The corresponding Gate, Up, and Down slices are fused using these group coefficients. Task-specific MLP neurons retain the base-model weights, discarding their expert updates. For Qwen, the General group likewise uses layer-specific estimated coefficients, while task-specific neurons retain the base weights. The coefficients are estimated from the separate calibration examples rather than by CWF's joint cross-entropy optimization.

DARE-Linear within NeuronMerge$_1$. For Math+Indonesian, Math+Thai, and Qwen Code+Math, remaining parameters use DARE-Linear with drop probability 0.8, scale $\lambda = 0.5$, and seed 20260924. NeuronMerge$_1$ uses DARE-Linear outside the MLP in these settings, while CWF copies non-fused parameters from its primary expert. Classification texts and expert pairs differ from those in the reference paper, hence the adaptation label.

Code+Multilingual result. The Llama merger uses General, Code, and Multilingual categories. Table 4 reports its HE+, MBPP+, and multilingual macro-average results under the common evaluation protocol.

A.7 Multilingual Aggregation

The multilingual score is

$$\text{Multi} = \frac{1}{12} \sum_{\ell \in \{\text{de,es,fr,ru}\}} (\text{ARC}_\ell^{\text{norm}} + \text{HellaSwag}_\ell^{\text{norm}} + \text{MMLU}_\ell) .$$

The four languages and three benchmark families receive equal weights regardless of their individual test-set sizes. Table 4 reports the resulting macro-averages for the base, experts, and merged models.

B Additional Analyses

B.1 Initialization-Ratio Analysis

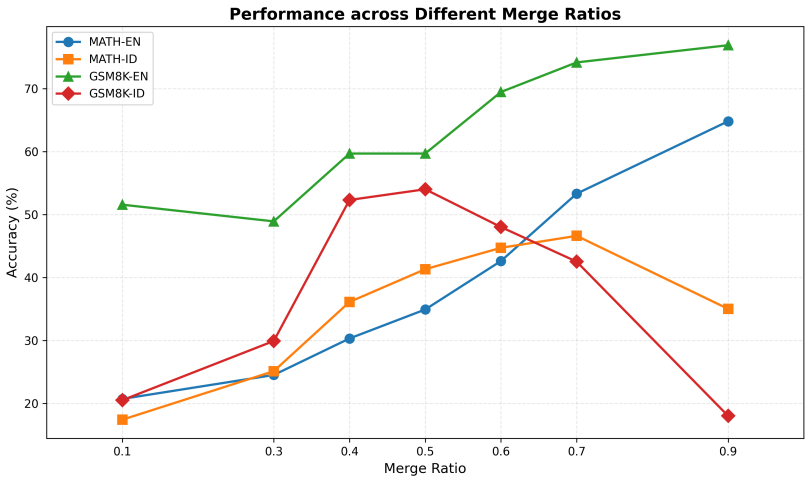

Figure 5: Sensitivity to the math expert's initialization ratio. The horizontal merge-ratio label denotes the initial contribution of the math expert. Indonesian GSM and MATH peak at different plotted ratios, approximately 0.5 and 0.7.

We study the sensitivity to initialization by varying the math expert's initial contribution from 0.1 to 0.9. Figure 5 shows increasing English MATH accuracy across the plotted values, while English GSM first decreases between 0.1 and 0.3 before increasing. The Indonesian curves peak at different ratios, approximately 0.5 for GSM8K-Id and 0.7 for MATH-Id, showing a task-dependent balance between source-language retention and target-language transfer. The default $0.731 : 0.269$ setting is given in Appendix A.3.

B.2 Component-Only Fusion

Table 7 compares FFN-only fusion, attention-only fusion, and joint CWF on Math+Indonesian. Joint CWF obtains the highest accuracy on all four benchmarks, supporting the combination of both component families.

B.3 Output-Interface Comparison

The default CWF merger retains the math expert's embeddings and LM head. Table 5 compares Copy Math Head, Copy Chat Head, and Weighted Fused Head on Math+Thai. Copy Math Head obtains 60.52% on English MATH and 27.68% on Thai MATH, compared with 16.54%/12.08% for Copy Chat Head and 52.98%/24.02% for Weighted Fused Head. These results support the selected output-interface strategy.

B.4 Memory Comparison

Table 8 compares peak VRAM on Llama-3-8B with batch size 2 and gradient accumulation 4. Full fine-tuning uses approximately 78 GB, whereas CWF uses approximately 60 GB, a reduction of about 18 GB in this experiment. CWF loads two frozen experts while maintaining optimizer states only for the fusion masks.

B.5 Activation Selection and Recording

For Math+Indonesian, the largest and smallest learned math-expert FFN coefficients select the math-biased and language-biased neurons at each illustrated layer. Figure 4 records their token-level SwiGLU outputs at the input to `down_proj`, after gating and multiplication.

Auxiliary neuron-removal experiment. We compare targeted and random neuron removal on a 500-question GSM8K subset by zeroing selected SwiGLU outputs before the Down projection. Table 14 reports mean accuracies for the two removal ratios.

Table 14: Mean accuracy after neuron removal on a 500-question GSM8K subset (%).

Ablation ratio	Neurons/layer	Random	Math-biased
1%	143	77.38	67.62
5%	717	76.56	60.43

At the 1% and 5% removal ratios, math-biased removal scores 9.76 and 16.13 percentage points below random removal, respectively.

B.6 Enlarged Fusion-Weight Distribution

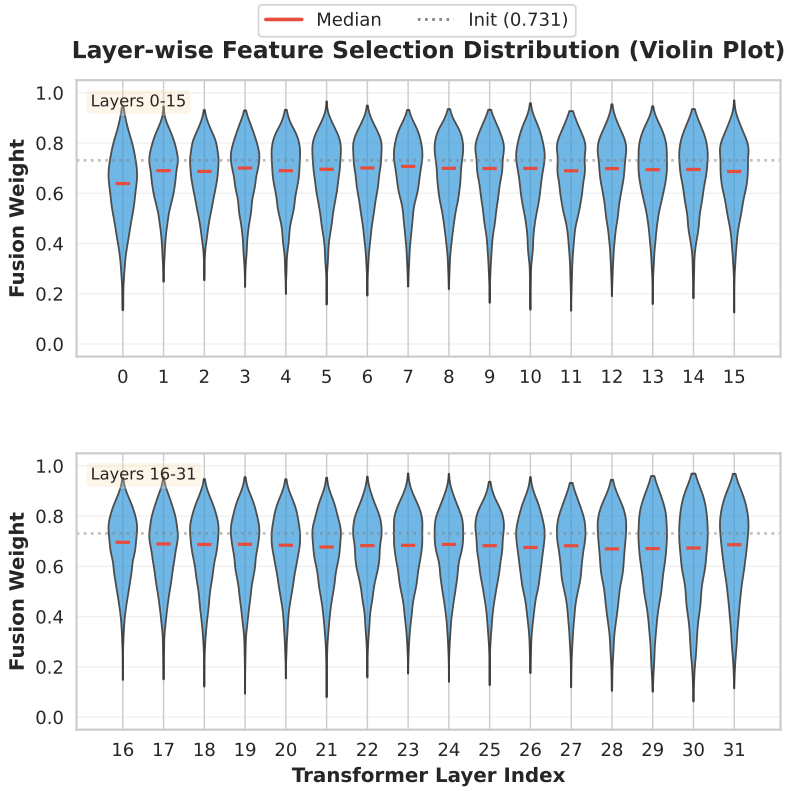

Figure 6: Enlarged view of the Math+Indonesian distributions in Figure 3. The 32 layer distributions and initialization reference are identical to the main-text figure.

B.7 Training Pipeline Illustration

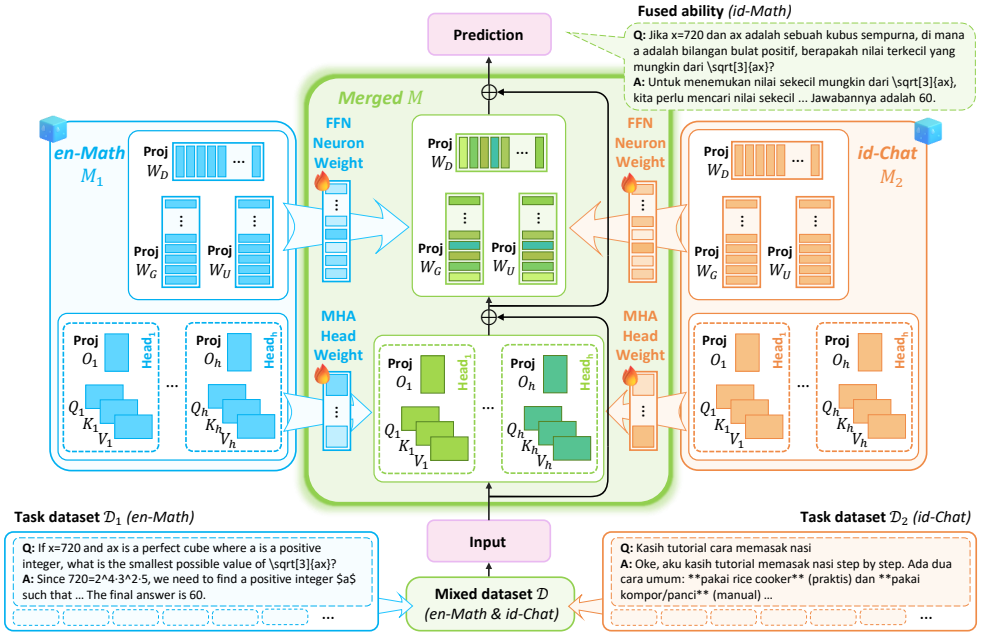

Figure 7: CWF training pipeline. Mixed-domain examples are evaluated through one fused network; gradients update component-specific mixing logits while source expert weights remain frozen. The FFN and attention blocks illustrate weighted parameter construction before single-model inference. The figure includes illustrative training inputs and a composite-task response. The attention drawing is schematic; shared-KV fusion follows the post-softmax averaging rule in Section 3.1.

The English mathematics and Indonesian conversation examples illustrate the two sources of mask-training supervision. The translated mathematics example illustrates the composite evaluation task. The shared training objective optimizes component-specific expert mixtures in the fused network. Once the coefficients are fixed, inference uses only the materialized parameters of the merged model.

Training and materialization sequence. For each training batch, first normalize the FFN and query-head logits across experts. Compute each shared K/V mixture by averaging its group's normalized query-head coefficients. Form the seven weighted projection families and combine them with the primary expert's fixed embedding, output, and normalization parameters. Evaluate the mixed-domain examples through this fused network and backpropagate the cross-entropy loss to the mixing logits; expert parameters remain fixed. After optimization, repeat the same weighted construction once to write the fused checkpoint. The saved model uses the ordinary backbone forward computation with those parameters.

C Limitations

The study evaluates two-expert mergers at the 8B scale, with different ability pairs on Llama and Qwen backbones. The results assess CWF as a complete merging procedure alongside baselines with their respective data access, coefficient parameterization, and fixed-parameter treatment. Controlled granularity and matched-data fine-tuning comparisons remain outside this evaluation.

The cross-lingual setting uses English mathematics and target-language conversation for mask training, with final-answer accuracy on target-language questions as the evaluation metric. The activation analysis examines selected neurons and prompts; the intervention compares targeted and random removal.