# 進化的モデルマージ

M2 上野詩翔

# 進化的モデルマージ

- ✓ 進化的アルゴリズムを用いたモデル合成手法
  - 進化的アルゴリズム：　　生物の進化過程を模倣したアルゴリズム
  - モデル合成：　　複数の基板モデルを合成して新たなモデルを獲得

- ✓ 日本語LLM×数学特化英語LLM → 数学特化日本語LLM
- ✓ 日本語LLM×英語VLM → 日本語VLM

**既存のモデルを組み合わせて新しい強力なモデルを効率的に構築する**

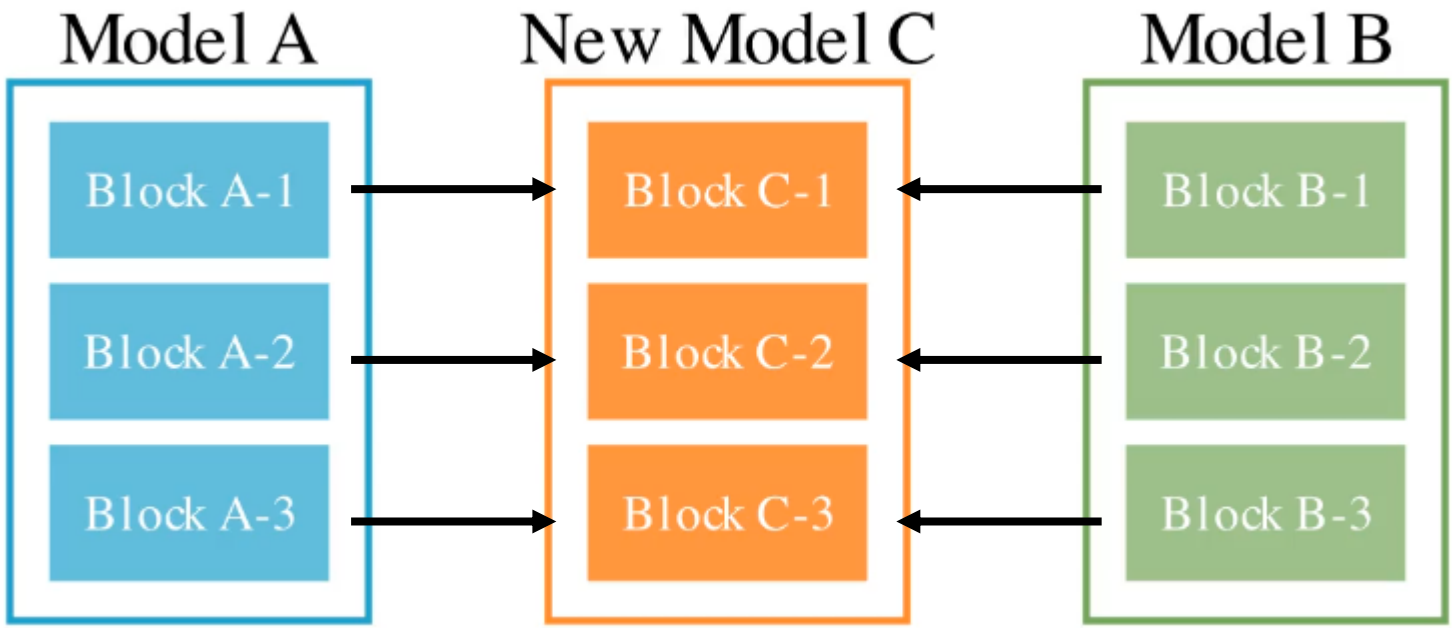
# モデルマージ

$$\mathrm{LLM}_{\mathrm{new}} = \mathrm{Merge}(\mathrm{LLM}_1, \mathrm{LLM}_2, \mathrm{LLM}_3, \ldots)$$

複数の既存モデルを元に1つの新しいモデルを作る

- ✓ 重みレベルのモデルマージ
- ✓ レイヤレベルのモデルマージ

# 重みレベルのモデルマージ




同じレイヤに相当する重みを混ぜ合わせる

# 重みレベルのモデルマージ｜線形補完による組み合わせ

$$\theta_{\text{new}} = \alpha\theta_1 + (1-\alpha)\theta_2$$

$\theta_1, \theta_2, \theta_{\text{new}}$：モデルの重み，$\alpha$：ハイパラ

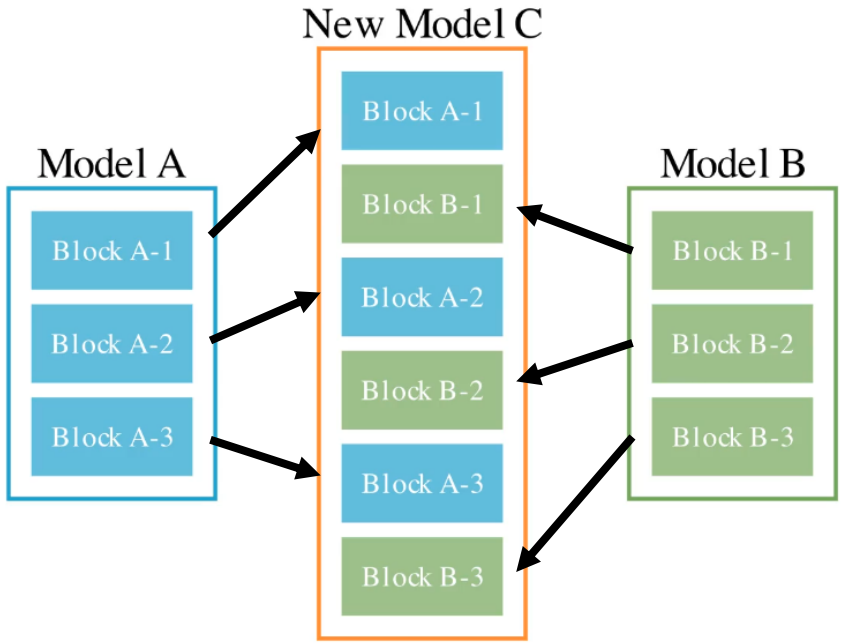
既存モデルの重みから新しいモデルの重みを合成

# 重みレベルのモデルマージ｜注意点

| | | | |
|---|---|---|---|
| | mlabonne/Beyonder-4x7B-v3 | 19.64 | false |
| | mistralai/Mixtral-8x7B-v0.1 | 19.56 | true |
| | mistralai/Mistral-7B-Instruct-v0.3 | 19.45 | true |
| | HuggingFaceH4/zephyr-7b-alpha | 18.85 | true |
| | mistralai/Mistral-7B-Instruct-v0.2 | 18.77 | true |
| | cognitivecomputations/dolphin-2.9-llama3-8b | 18.62 | true |

- ✓ベンチマークへのオーバーフィット
  - Open LLM Leaderboard2で解消気味？
- ✓マージのマージのマージ・・・
  - 中身は公開されているが・・・

# レイヤレベルのモデルマージ（1/2）




異なるモデルのレイヤを重ね合わせる

# レイヤレベルのモデルマージ（2/2）

- ✓ 重みの変更は行わなず，レイヤを組み合わせる
  - レイヤ：
    - Transformer Block
    - LLM Block

- ✓ パラメータ数が増減

# モデルマージの目的

## アンサンブル

- 多様な知識の融合
  - 新規タスクへの汎化性を獲得
  - 既存タスクへの性能向上
- 学習不要で新規モデルを構築

# モデルマージの課題

- ✓同じアーキテクチャのモデル同士でないと困難
- ✓同じベースモデルからのFTでないと困難
- ✓元のモデルサイズが大きくないと困難

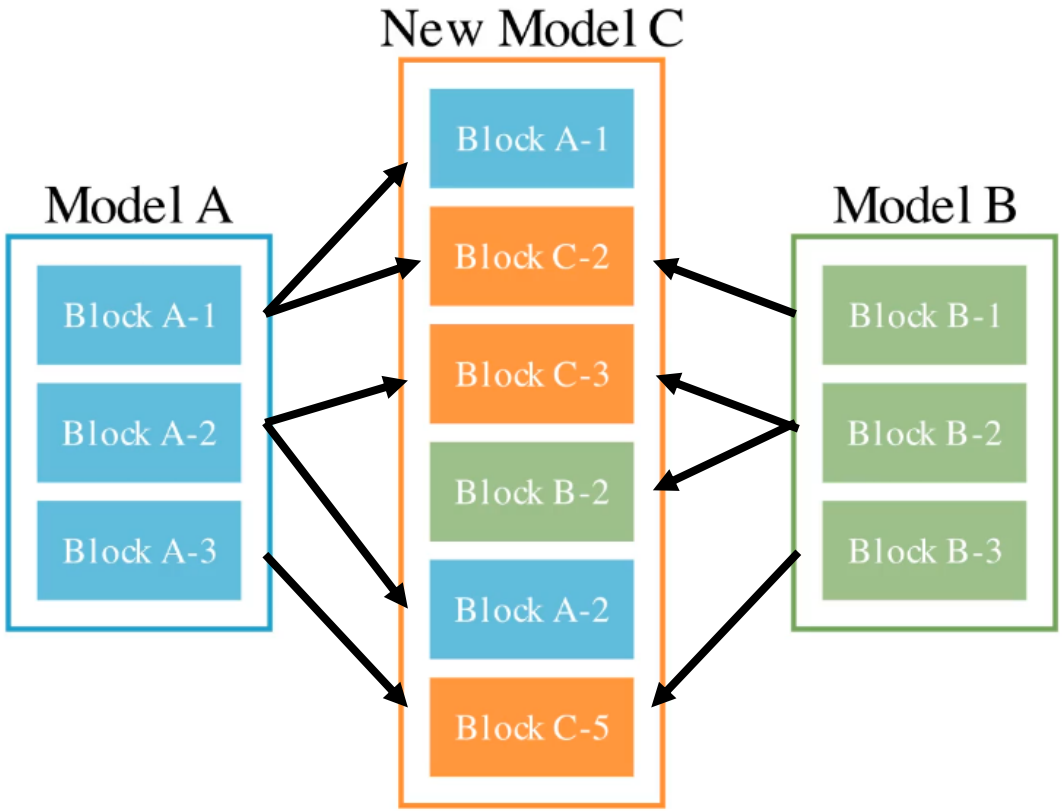
→ LLM，LVLMへの適用が現実的

# 日本語LLMマージの課題

- ✓ 日本語ベースLLMが少ない
  - – 既存の日本語LLMはLLaMA2などから継続事前学習
  - → 元のLLMの重みから大きく離れている

- ✓ 日本語ベースLLMは小さい

# 進化的モデルマージ



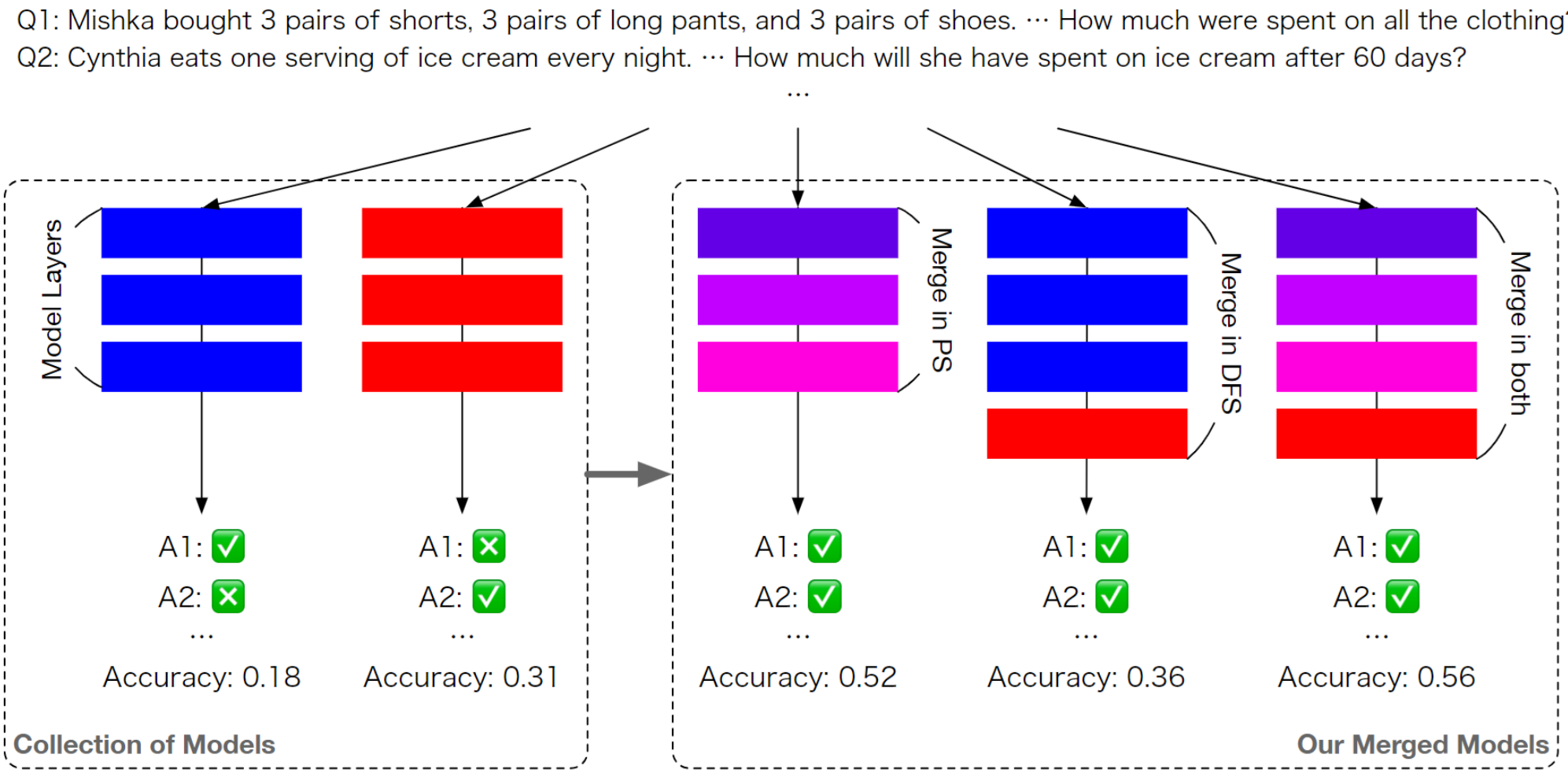

重みレベル・レイヤレベルのモデルマージを，
進化的アルゴリズムに基づいて探索

# 進化的アルゴリズム

生物の進化の過程を模倣した最適化アルゴリズム

1. （個体群）課題に対する回答パターンを1つ1つの遺伝子として生成
2. （適応度）各遺伝子の正解率をスコア化
3. （選択）正解率の高い遺伝子が生き残るようにする
4. （交叉）選ばれた遺伝子を混ぜ合わせて新しい遺伝子を生成
5. （変異）遺伝子にランダムな変更を加え，局所解を予防

# 進化的モデルマージ




パラメータ空間＝重みマージ　データフロー空間＝レイヤマージ　両方

# 重みレベルの進化的マージ｜ TIES-Merging

**Algorithm 1** TIES-MERGING Procedure.

**Input:** Fine-tuned models $\{\theta_t\}_{t=1}^{n}$, Initialization $\theta_{\text{init}}$, $k$, and $\lambda$.
**Output:** Merged Model $\theta_m$
**forall** $t$ ***in*** $1, ..., n$ **do**
  ▷ Create task vectors.
  $\tau_t = \theta_t - \theta_{\text{init}}$ ← 元の重みとFT後の差
  ▷ Step 1: Trim redundant parameters.
  $\hat{\tau}_t \leftarrow \text{keep\_topk\_reset\_rest\_to\_zero}(\tau_t, k)$
  $\hat{\gamma}_t \leftarrow sgn(\hat{\tau}_t)$ ← 閾値でフィルタ
  $\hat{\mu}_t \leftarrow |\hat{\tau}_t|$
**end**
▷ Step 2: Elect Final Signs.
$\gamma_m = sgn(\sum_{t=1}^{n} \hat{\tau}_t)$
▷ Step 3: Disjoint Merge.
**forall** $p$ ***in*** $1, ..., d$ **do**
  $\mathcal{A}^p = \{t \in [n] \mid \hat{\gamma}_t^p = \gamma_m^p\}$
  $\tau_m^p = \frac{1}{|\mathcal{A}^p|} \sum_{t \in \mathcal{A}^p} \hat{\tau}_t^p$ ← 重みを融合
**end**
▷ Obtain merged checkpoint
$\theta_m \leftarrow \theta_{\text{init}} + \lambda * \tau_m$
**return** $\theta_m$

# レイヤレベルの進化的マージ｜ CMA-ES

**Set parameters**

Set parameters $\lambda$, $w_{i=1\ldots\lambda}$, $c_\sigma$, $d_\sigma$, $c_\mathrm{c}$, $c_1$, and $c_\mu$ according to Table 1.

**Initialization**

Set evolution paths $\boldsymbol{p}_\sigma = \mathbf{0}$, $\boldsymbol{p}_\mathrm{c} = \mathbf{0}$, covariance matrix $\boldsymbol{C} = \mathbf{I}$, and $g = 0$.

Choose distribution mean $\boldsymbol{m} \in \mathbb{R}^n$ and step-size $\sigma \in \mathbb{R}_{>0}$ problem dependent.[1]

**Until termination criterion met**, $g \leftarrow g + 1$

Sample new population of search points, for $k = 1, \ldots, \lambda$

$$\boldsymbol{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{38}$$

$$\boldsymbol{y}_k = \boldsymbol{B}\boldsymbol{D}\boldsymbol{z}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C}) \tag{39}$$

$$\boldsymbol{x}_k = \boldsymbol{m} + \sigma \boldsymbol{y}_k \sim \mathcal{N}\left(\boldsymbol{m}, \sigma^2 \boldsymbol{C}\right) \tag{40}$$

← 多変量正規分布からサンプリング

Selection and recombination

$$\langle \boldsymbol{y} \rangle_\mathrm{w} = \sum_{i=1}^{\mu} w_i \, \boldsymbol{y}_{i:\lambda} \quad \text{where } \sum_{i=1}^{\mu} w_i = 1, \; w_i > 0 \text{ for } i = 1 \ldots \mu \tag{41}$$

$$\boldsymbol{m} \leftarrow \boldsymbol{m} + c_\mathrm{m} \sigma \langle \boldsymbol{y} \rangle_\mathrm{w} \quad \text{equals } \sum_{i=1}^{\mu} w_i \, \boldsymbol{x}_{i:\lambda} \text{ if } c_\mathrm{m} = 1 \tag{42}$$

← 分布の中心を更新

Step-size control

$$\boldsymbol{p}_\sigma \leftarrow (1 - c_\sigma)\boldsymbol{p}_\sigma + \sqrt{c_\sigma(2 - c_\sigma)\mu_\mathrm{eff}} \, \boldsymbol{C}^{-\frac{1}{2}} \langle \boldsymbol{y} \rangle_\mathrm{w} \tag{43}$$

$$\sigma \leftarrow \sigma \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma\|}{\mathsf{E}\|\mathcal{N}(\mathbf{0}, \mathbf{I})\|} - 1\right)\right) \tag{44}$$

← 探索範囲を更新

Covariance matrix adaptation

$$\boldsymbol{p}_\mathrm{c} \leftarrow (1 - c_\mathrm{c})\boldsymbol{p}_\mathrm{c} + h_\sigma \sqrt{c_\mathrm{c}(2 - c_\mathrm{c})\mu_\mathrm{eff}} \, \langle \boldsymbol{y} \rangle_\mathrm{w} \tag{45}$$

$$w_i^\circ = w_i \times (1 \text{ if } w_i \geq 0 \text{ else } n / \|\boldsymbol{C}^{-\frac{1}{2}} \boldsymbol{y}_{i:\lambda}\|^2) \tag{46}$$

$$\boldsymbol{C} \leftarrow (1 + \underbrace{c_1 \delta(h_\sigma) - c_1 - c_\mu \textstyle\sum w_j}_{\text{usually equals to 0}}) \, \boldsymbol{C} + c_1 \boldsymbol{p}_\mathrm{c} \boldsymbol{p}_\mathrm{c}^\mathsf{T} + c_\mu \sum_{i=1}^{\lambda} w_i^\circ \, \boldsymbol{y}_{i:\lambda} \boldsymbol{y}_{i:\lambda}^\mathsf{T} \tag{47}$$

← 共分散行列を更新

# 数学特化型日本語LLM

- ✓ 以下をマージ
  - shisa-gamma-7b-v1（日本語LLM）
  - WizardMath-7B-V1.1
  - Abel-7B-002
    - 全てMistral-7B-v0.1ベース

# 数学特化型日本語LLM｜結果

| Id. | Model | Type | Size | MGSM-JA (acc ↑) | JP-LMEH (avg ↑) |
|---|---|---|---|---|---|
| 1 | Shisa Gamma 7B v1 | JA general | 7B | 9.6 | 66.1 |
| 2 | WizardMath 7B v1.1 | EN math | 7B | 18.4 | 60.1 |
| 3 | Abel 7B 002 | EN math | 7B | 30.0 | 56.5 |
| 4 | **Ours (PS)** | 1 + 2 + 3 | 7B | **52.0** | **70.5** |
| 5 | **Ours (DFS)** | 3 + 1 | 10B | **36.4** | **53.2** |
| 6 | **Ours (PS+DFS)** | 4 + 1 | 10B | **55.2** | **66.2** |
| 7 | Llama 2 70B | EN general | 70B | 18.0 | 64.5 |
| 8 | Japanese StableLM 70B | JA general | 70B | 17.2 | 68.3 |
| 9 | Swallow 70B | JA general | 70B | 13.6 | 71.5 |
| 10 | GPT-3.5 | commercial | - | 50.4 | - |
| 11 | GPT-4 | commercial | - | 78.8 | - |

元の日本語能力を保持したまま，数学能力も向上

# 数学特化型日本語LLM｜結果詳細

| | | JP Language Model Evaluation Harness | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Size** | JComQA | JNLI | MARC | JSQuAD | JAQKET | XLSum | XWino | MGSM | JCoLA | **Avg** |
| Shisa Gamma 7b v1 | 7B | 91.2 | 72.1 | 94.6 | 73.9 | 68.0 | 25.9 | 80.5 | 29.6 | 58.7 | 66.1 |
| WizardMath 7B V1.1 | 7B | 74.7 | 42.7 | 90.4 | 84.6 | 68.5 | 22.3 | 69.8 | 38.8 | 48.9 | 60.1 |
| Abel 7B 002 | 7B | 70.3 | 51.8 | 62.3 | 83.8 | 69.0 | 22.5 | 68.2 | 28.0 | 52.7 | 56.5 |
| **Ours (PS)** | 7B | 89.1 | 65.7 | 95.4 | 89.5 | 77.7 | 25.5 | 81.2 | 50.0 | 60.5 | **70.5** |
| **Ours (DFS)** | 10B | 67.7 | 58.2 | 53.5 | 66.8 | 54.3 | 17.3 | 65.6 | 30.0 | 65.6 | **53.2** |
| **Ours (PS+DFS)** | 10B | 88.2 | 50.3 | 91.5 | 78.6 | 77.8 | 23.2 | 73.0 | 40.0 | 73.0 | **66.2** |
| Llama 2 70B | 70B | 80.2 | 53.4 | 94.4 | 91.6 | 80.1 | 21.8 | 73.6 | 30.4 | 54.6 | 64.5 |
| Japanese Stable LM 70B | 70B | 91.2 | 50.4 | 92.9 | 87.1 | 88.4 | 24.3 | 82.0 | 37.2 | 61.7 | 68.3 |
| Swallow 70B | 70B | 95.3 | 57.2 | 91.7 | 94.1 | 93.9 | 23.1 | 83.3 | 45.2 | 59.5 | 71.5 |

# 日本語VLM

- ✓ 以下をマージ
  - shisa-gamma-7b-v1
  - LLaVA-1.6-Mistral-7B
    - どちらもMistral-7B-v0.1ベース

- ✓ VLM：Vision-Language Model
  - 多くがViT＋LLMで画像と言語を扱う

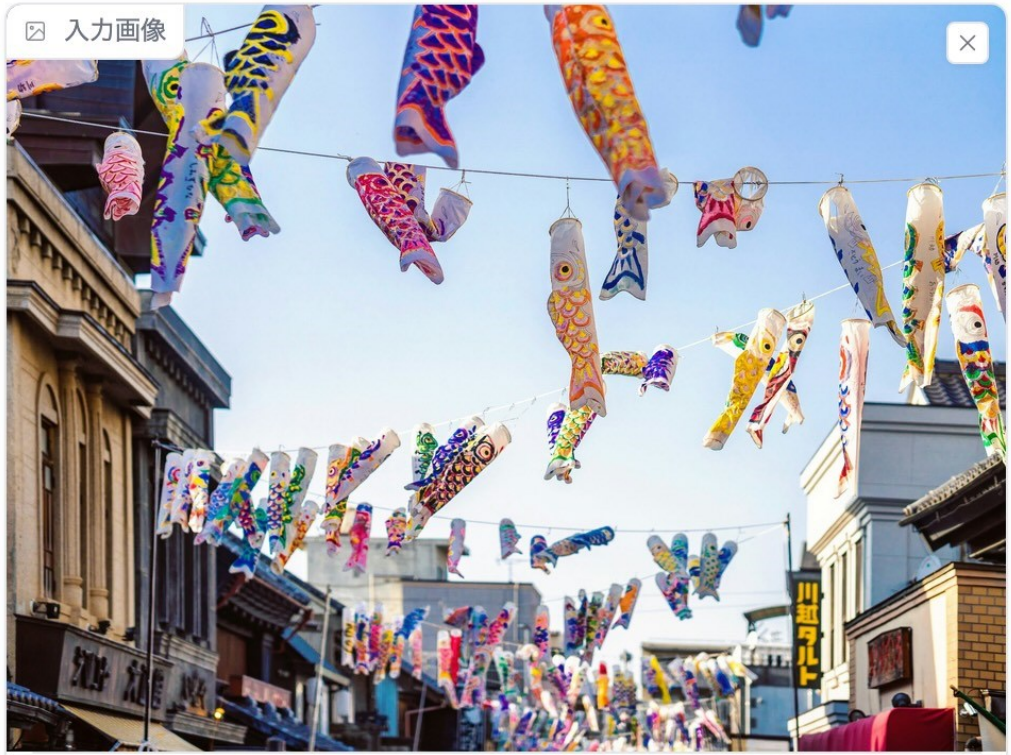
# 日本語VLM｜結果

| Model | JA-VG-VQA-500 (ROUGE-L ↑) | JA-VLM-Bench-In-the-Wild (ROUGE-L ↑) |
|---|---|---|
| LLaVA 1.6 Mistral 7B | 14.3 | 41.1 |
| Japanese Stable VLM | - | 40.5 |
| **Ours** | **19.7** | **51.2** |

日本語のVLベンチで性能向上

# 日本語VLM｜結果詳細

入力画像




質問

この画像に映っている魚の形をした飾りは、日本では何と呼ばれていますか？

モデルからの返答

この画像に映っている魚の形をした飾りは、日本では「鯉のぼり」と呼ばれています。

日本の文化も保有

# 日本語VLM｜デモ

https://huggingface.co/spaces/SakanaAI/EvoVLM-JP

# 進化的アルゴリズム×生成モデル

# 結論

- ✓進化的アルゴリズムを用いたマージの探索手法を提案

- ✓異なるモデル同士の知識を融合した新しいモデルを構築可能
  - 日本語LLM×数学特化英語LLM → 数学特化日本語LLM
  - 日本語LLM×英語VLM → 日本語VLM

- ✓モデルの訓練などが一切不要
  - ただし，元々のサイズが大きいモデルでないとマージが成功しない

  → 推論時にはある程度の計算機が必要

- ✓※ 元のモデルの選び方は記載なし