Mohamed bin Zayed University of Artificial Intelligence

30th ANNIVERSARY EMNLP 2025 Suzhou, China November 4-9 | 中国苏州 11月4日–9日

How a Bilingual LM Becomes Bilingual: Tracing Internal Representations with Sparse Autoencoders

Tatsuro Inaba, Go Kamoda, Kentaro Inui, Masaru Isonuma,
Yusuke Miyao, Yohei Oseki, Yu Takagi, Benjamin Heinzerling**

EMNLP 2025 findings

Presenter: Tatsuro Inaba, PhD student @MBZUAI

2025/10/28, 第24回 LLM 勉強会

まとめ

Mohamed bin Zayed University of Artificial Intelligence

- 学習ステージ・層・モデルサイズの変化に伴い LM 内部表現がどう変化するか

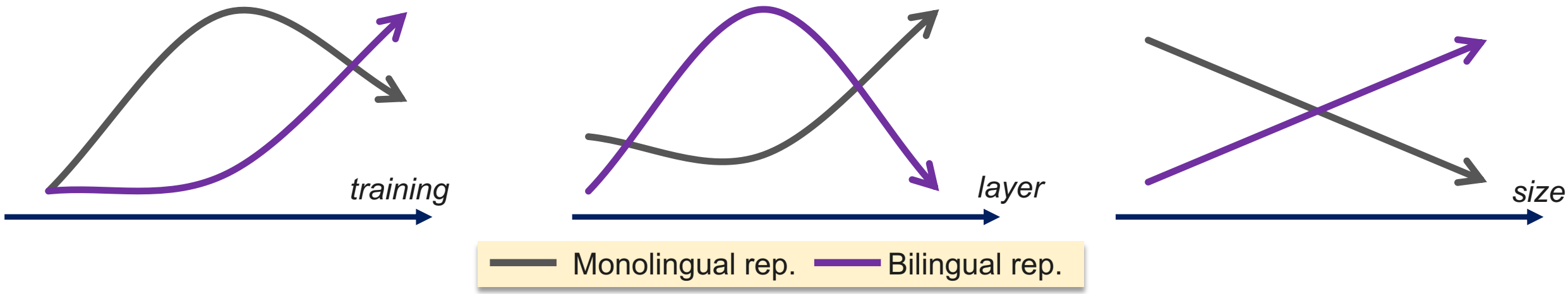
Findings

(i) 言語を別々に習得 → 言語間の対応を学習

(ii) 中間層付近が言語間の対応をより獲得

(iii) 大きいモデルほどバイリンガルで文脈の意味をとらえた表現を獲得

(iv) 言語間の対応に関する表現は性能に大きな影響



Background: Superposition

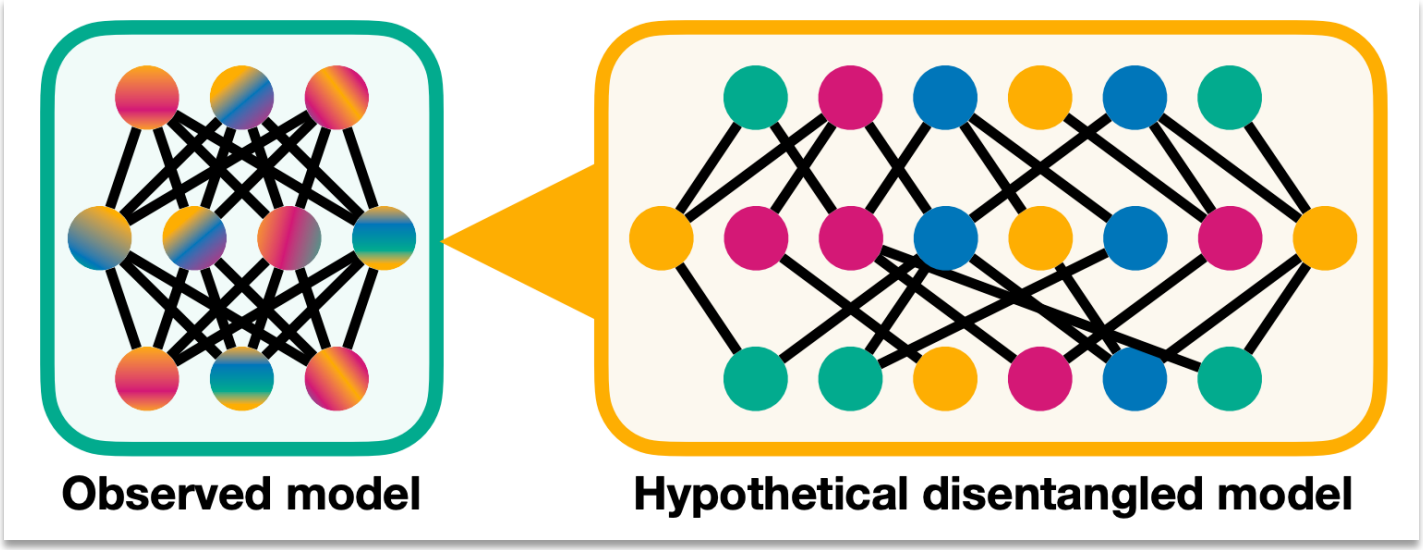
Mohamed bin Zayed University of Artificial Intelligence

- 概念数 > 次元数のとき、

  → 1つの次元に複数の概念がエンコードされる (Superposition)

  → 解釈するのが難しい

- 絡まった表現（次元）を解きほぐして Monosemantic にしたい



Background: スパースオートエンコーダ (SAE)

Mohamed bin Zayed University of Artificial Intelligence

- 中間層が高次元で**疎**なオートエンコーダにより解きほぐす
  - オートエンコーダ: 入力を**再構成**するネットワーク
  - 入力/出力次元数 << 中間層
- （本研究では）中間層の各次元を特徴量 (Features) と呼ぶ

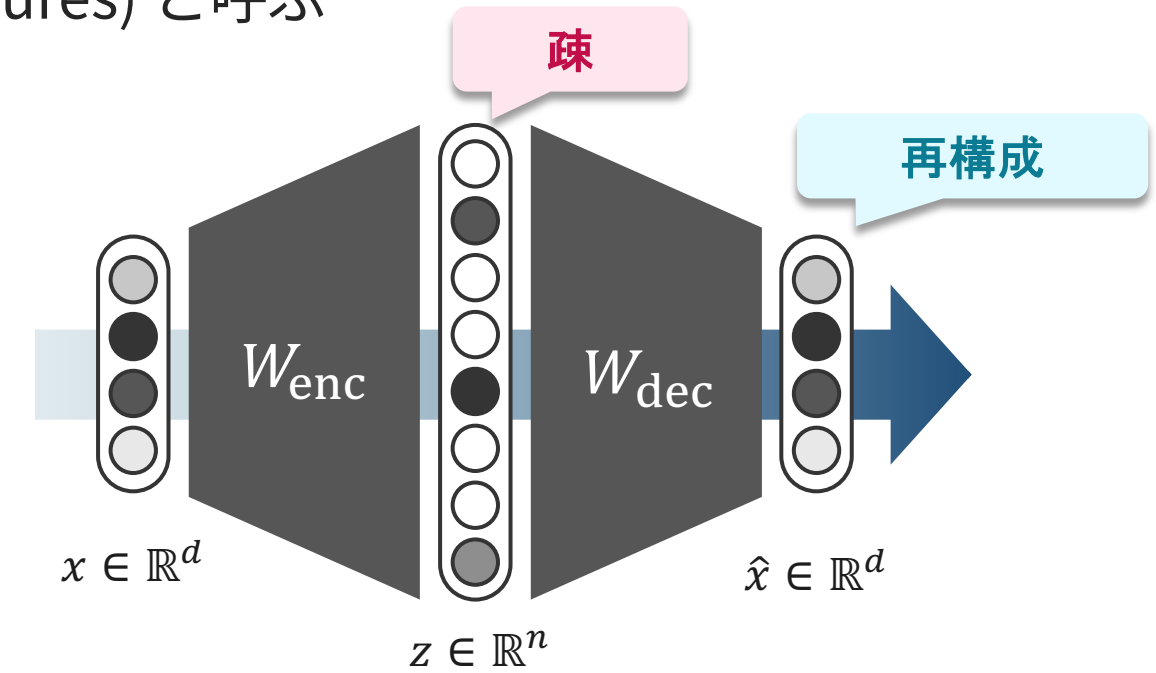
定式化

$$z = \mathrm{ReLU}\left(W_{\mathrm{enc}}\left(x - b_{\mathrm{pre}}\right)\right)$$
$$\hat{x} = W_{\mathrm{dec}} z + b_{\mathrm{pre}}$$

ロス関数

$$\mathcal{L} = \|x - \hat{x}\|_2^2 + \lambda \|z\|_1$$

再構成　疎



特徴量の定量化

Mohamed bin Zayed University of Artificial Intelligence

- ある特徴量を発火させるトークンたちから、その傾向を分類/数値化
- 2つの指標:
  - **Language**: English, Japanese, or Mixed
  - **Monosemancitiy** (意味のまとまり): 0~1 (0=意味がバラバラ、1=意味が単一)

Activating Tokens	Language
Dogs are scary/Cats are cute/Cat cafe	**English**
犬は怖い/猫は可愛い/猫カフェ	**Japanese**
Dogs are scary/猫は可愛い/Cat cafe	**Mixed**

Activating Tokens	Monosemanticity
犬は怖い/I like apples/I love that guitar	**0.00**
犬は怖い/Cat cafe/I like apples	**0.50**
Dogs are cute/猫は可愛い/Cat cafe	**1.00**

特徴量の定量化

Mohamed bin Zayed University of Artificial Intelligence

- ある特徴量を発火させるトークンたちから、その傾向を分類/数値化
- 2つの指標:
  - **Language**: English, Japanese, or Mixed
  - **Monosemancitiy** (意味のまとまり): 0~1 (0=意味がバラバラ、1=意味が単一)

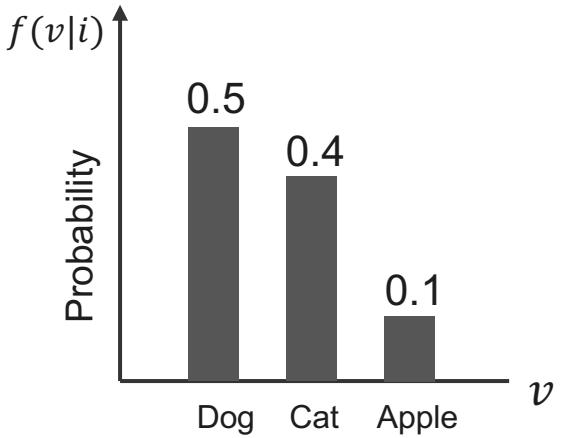
Activating Tokens	Language
Dogs are scary/Cats are cute/Cat cafe	**English**
犬は怖い/猫は可愛い/猫カフェ	**Japanese**
Dogs are scary/猫は可愛い/Cat cafe	**Mixed**

Activating Tokens	Monosemanticity
犬は怖い/I like apples/I love that guitar	**0.00**
犬は怖い/Cat cafe/I like apples	**0.50**
Dogs are cute/猫は可愛い/Cat cafe	**1.00**

Monosemanticity

Mohamed bin Zayed University of Artificial Intelligence

1. Token entropy を計算: $H_{\text{token}}(i) = -\sum_{v \in V} f(v|i) \log f(v|i)$
2. Semantic entropy を計算:
   1. コサイン類似度に基づいてトークンをクラスタリング
   2. クラスターレベルの entropy を計算: $H_{\text{semantic}}(i) = -\sum_{c \in C_i} p(c|i) \log p(c|i)$
3. **Monosemanticity** を計算: $R_{\text{mono}}(i) = 1 - \frac{H_{\text{semantic}}(i)}{H_{\text{token}}(i)}$

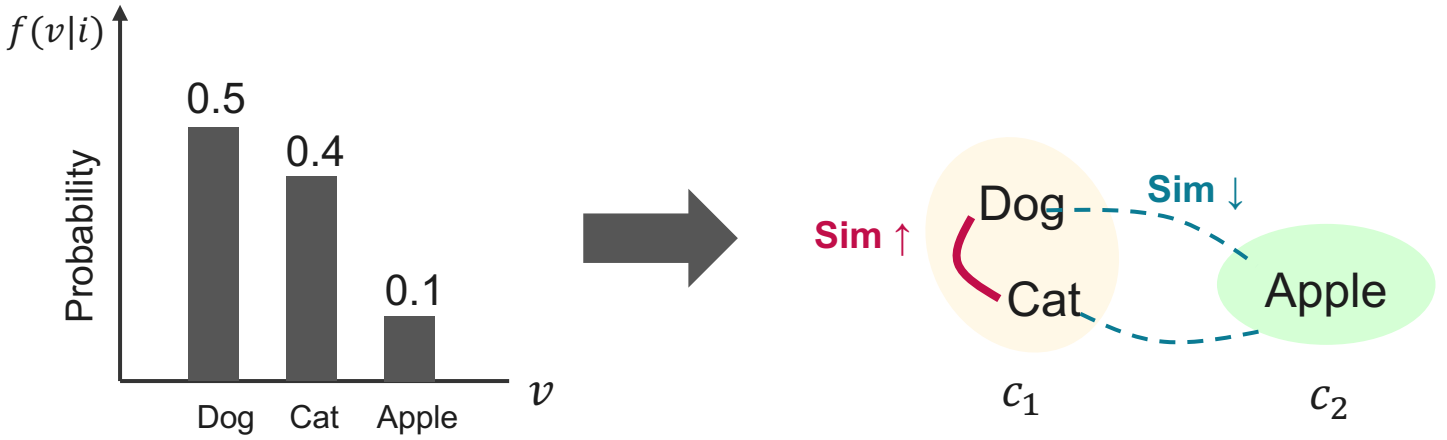


Activation patterns of i-th Feature

Monosemanticity

Mohamed bin Zayed University of Artificial Intelligence

1. Token entropy を計算: $H_{\text{token}}(i) = -\sum_{v \in V} f(v|i) \log f(v|i)$
2. Semantic entropy を計算:
   1. コサイン類似度に基づいてトークンをクラスタリング
   2. クラスターレベルの entropy を計算: $H_{\text{semantic}}(i) = -\sum_{c \in C_i} p(c|i) \log p(c|i)$
3. **Monosemanticity** を計算: $R_{\text{mono}}(i) = 1 - \frac{H_{\text{semantic}}(i)}{H_{\text{token}}(i)}$



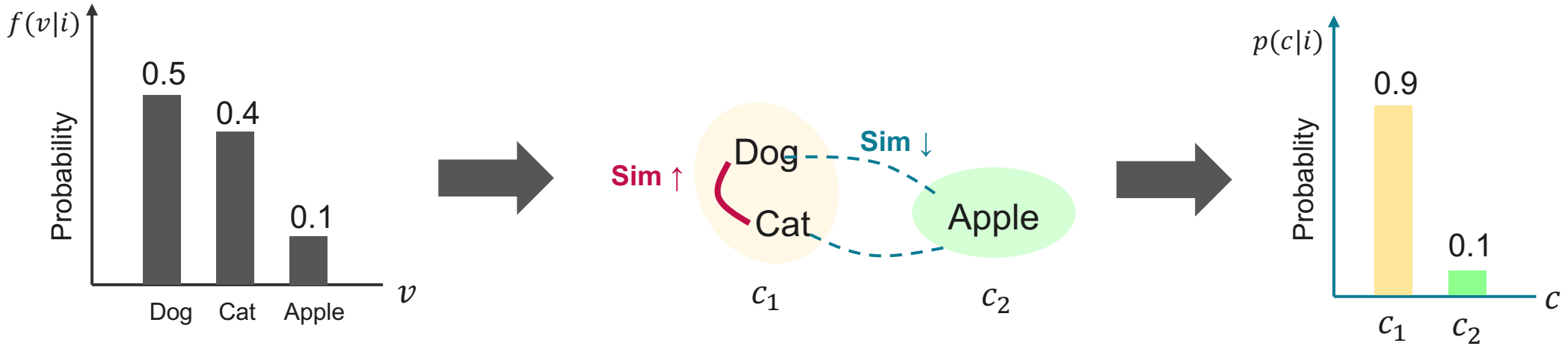
Activation patterns of i-th Feature

Clustering in the embedding space

Monosemanticity

Mohamed bin Zayed University of Artificial Intelligence

1. Token entropy を計算: $H_{\text{token}}(i) = -\sum_{v \in V} f(v|i) \log f(v|i)$
2. Semantic entropy を計算:
   1. コサイン類似度に基づいてトークンをクラスタリング
   2. クラスターレベルの entropy を計算: $H_{\text{semantic}}(i) = -\sum_{c \in C_i} p(c|i) \log p(c|i)$
3. **Monosemanticity** を計算: $R_{\text{mono}}(i) = 1 - \frac{H_{\text{semantic}}(i)}{H_{\text{token}}(i)}$



Activation patterns of i-th Feature

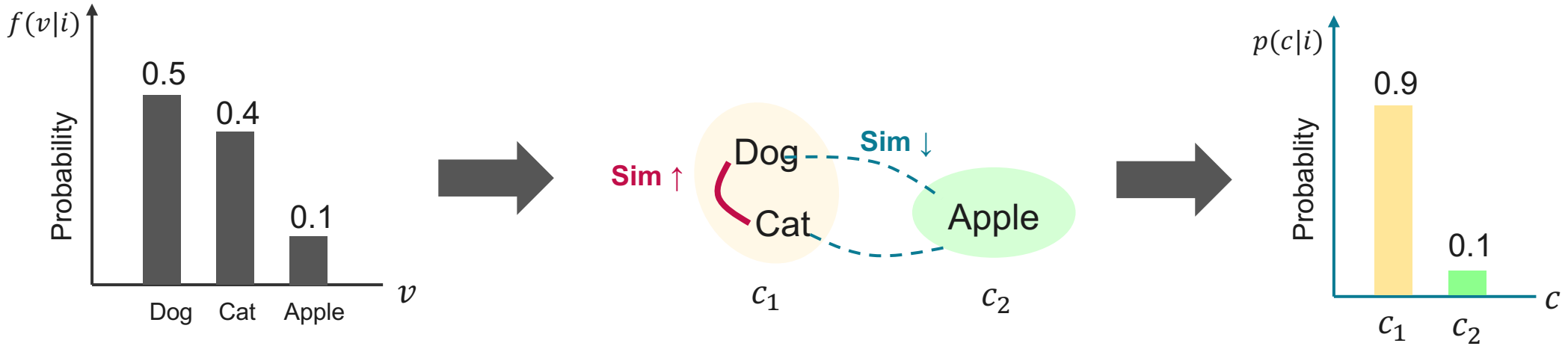
Clustering in the embedding space

Monosemanticity

Mohamed bin Zayed University of Artificial Intelligence

1. Token entropy を計算: $H_{\text{token}}(i) = -\sum_{v \in V} f(v|i) \log f(v|i)$
2. Semantic entropy を計算:
   1. コサイン類似度に基づいてトークンをクラスタリング
   2. クラスターレベルの entropy を計算: $H_{\text{semantic}}(i) = -\sum_{c \in C_i} p(c|i) \log p(c|i)$

3. **Monosemanticity** を計算: $R_{\text{mono}}(i) = 1 - \frac{H_{\text{semantic}}(i)}{H_{\text{token}}(i)}$



Activation patterns of i-th Feature

Clustering in the embedding space

実験設定

Mohamed bin Zayed University of Artificial Intelligence

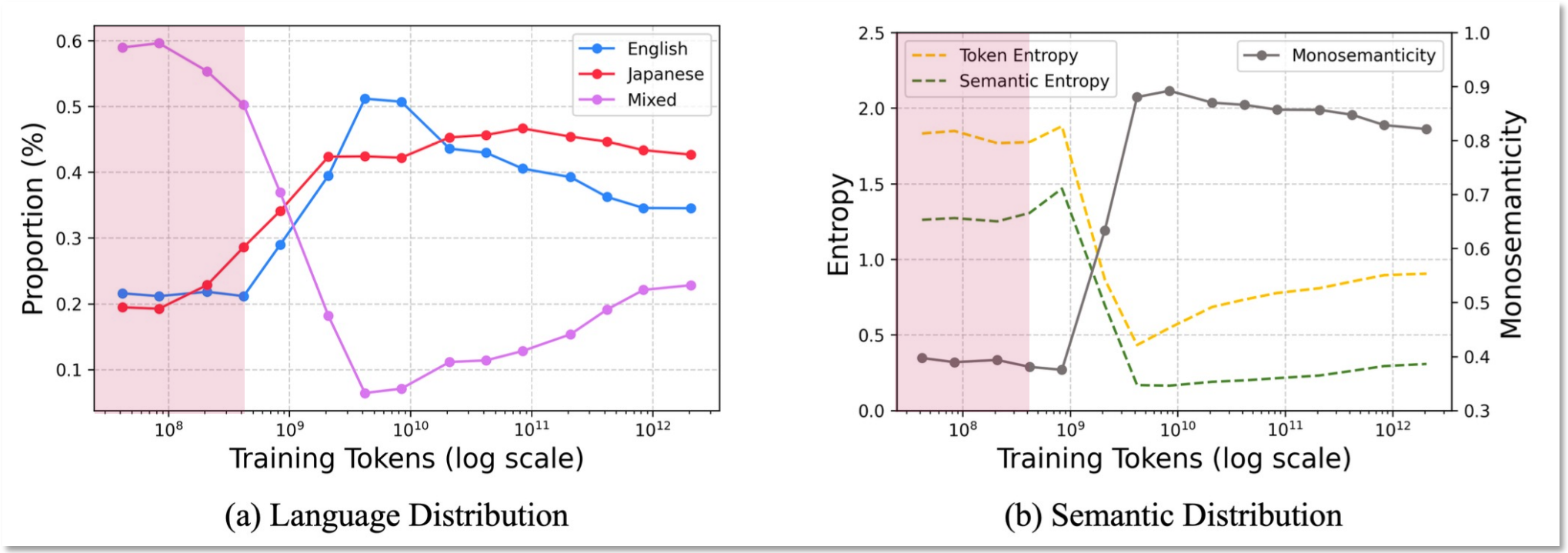
- モデル: LLM-jp-3 (日英バイリンガルモデル)
  - (150M-3.7B) × (16 checkpoints) × (All layers) => 1572 setups
- データ: En-Wikipedia / Ja-Wikipedia (1:1)
- TopK-SAE
  - sparsity K=32 / 中間層 n=32,768
  - 学習率はグリッドサーチにより決定

結果: 学習ステージ方向 (3.7B, L=14)

Mohamed bin Zayed University of Artificial Intelligence

学習初期

- Mixed language の割合が多い & monosemanticity が低い
  → ほとんどの特徴量が**ランダムな**トークンで発火



結果: 学習ステージ方向 (3.7B, L=14)

Mohamed bin Zayed University of Artificial Intelligence

学習初期

- Mixed language の割合が多い & monosemanticity が低い
  → ほとんどの特徴量が**ランダムな**トークンで発火

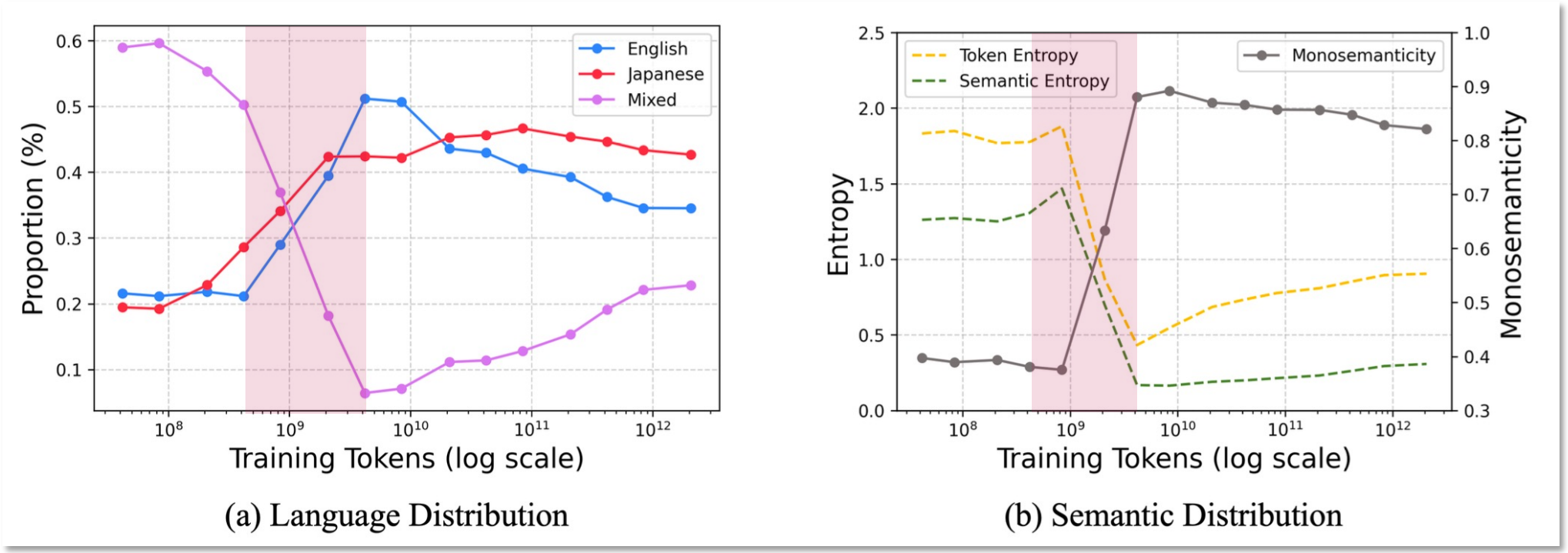
Activating tokens	Language	Monosemanticity
・Born 20 June 1967) is ・This American Life episodes ・西部、ジュネーブ州の	Mixed	0.19
・in Houston County, Albama ・Trichromia repanda is a ・大会は1938年の2月	Mixed	0.24

結果: 学習ステージ方向 (3.7B, L=14)

Mohamed bin Zayed University of Artificial Intelligence

学習中期

- 英語特徴量 / 日本語特徴量の割合が増加 & Monosemanticity も増加
  → 特徴量が**各言語内**で意味をとらえるように



結果: 学習ステージ方向 (3.7B, L=14)

Mohamed bin Zayed University of Artificial Intelligence

学習中期

- 英語特徴量 / 日本語特徴量の割合が増加 & Monosemanticity も増加
  → 特徴量が**各言語内**で意味をとらえるように

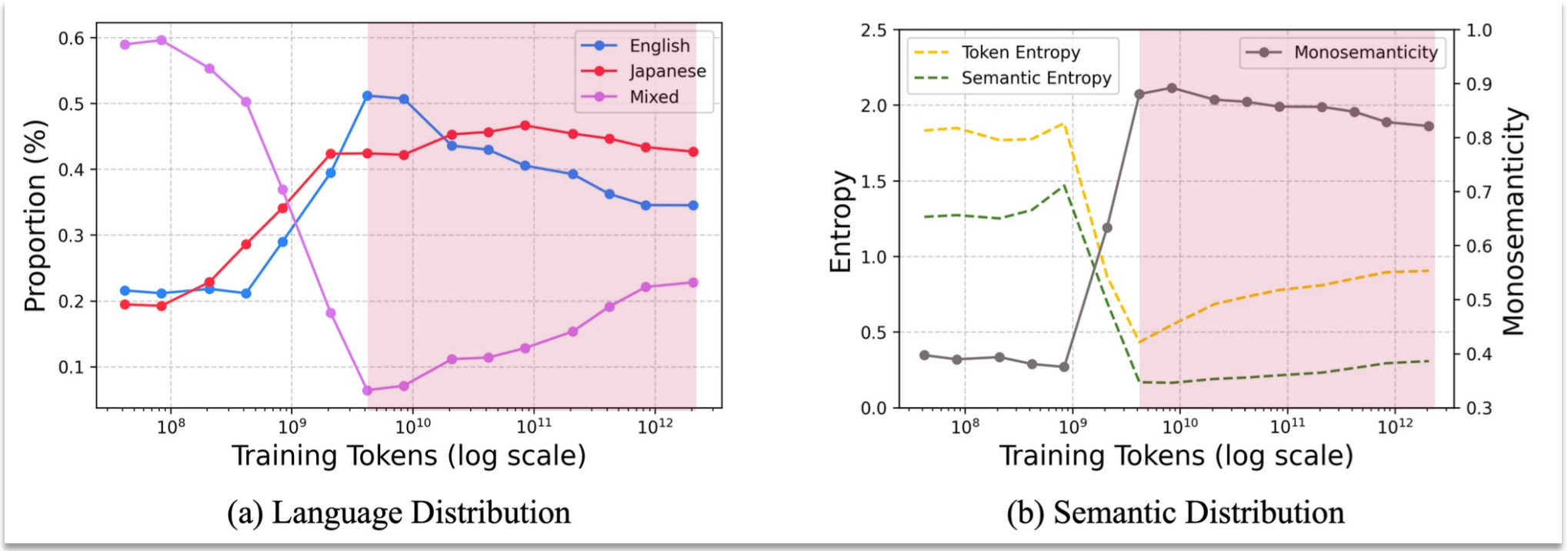
・which give rise to ・secretly gave assistance to ・which had given some	English	1.00
・は、ドイツの哲学者 ・、日本の明治期の ・は、イギリスの法学者	Japanese	1.00

結果: 学習ステージ方向 (3.7B, L=14)

Mohamed bin Zayed University of Artificial Intelligence

学習後期

- Mixed Language の割合が増加 & Monosemanticity は比較的高いまま
  → 一部の Feature が**言語を超えた意味**を捉え始める



結果: 学習ステージ方向 (3.7B, L=14)

Mohamed bin Zayed University of Artificial Intelligence

学習後期

- Mixed Language の割合が増加 & Monosemanticity は比較的高いまま
  → 一部の Feature が**言語を超えた意味**を捉え始める

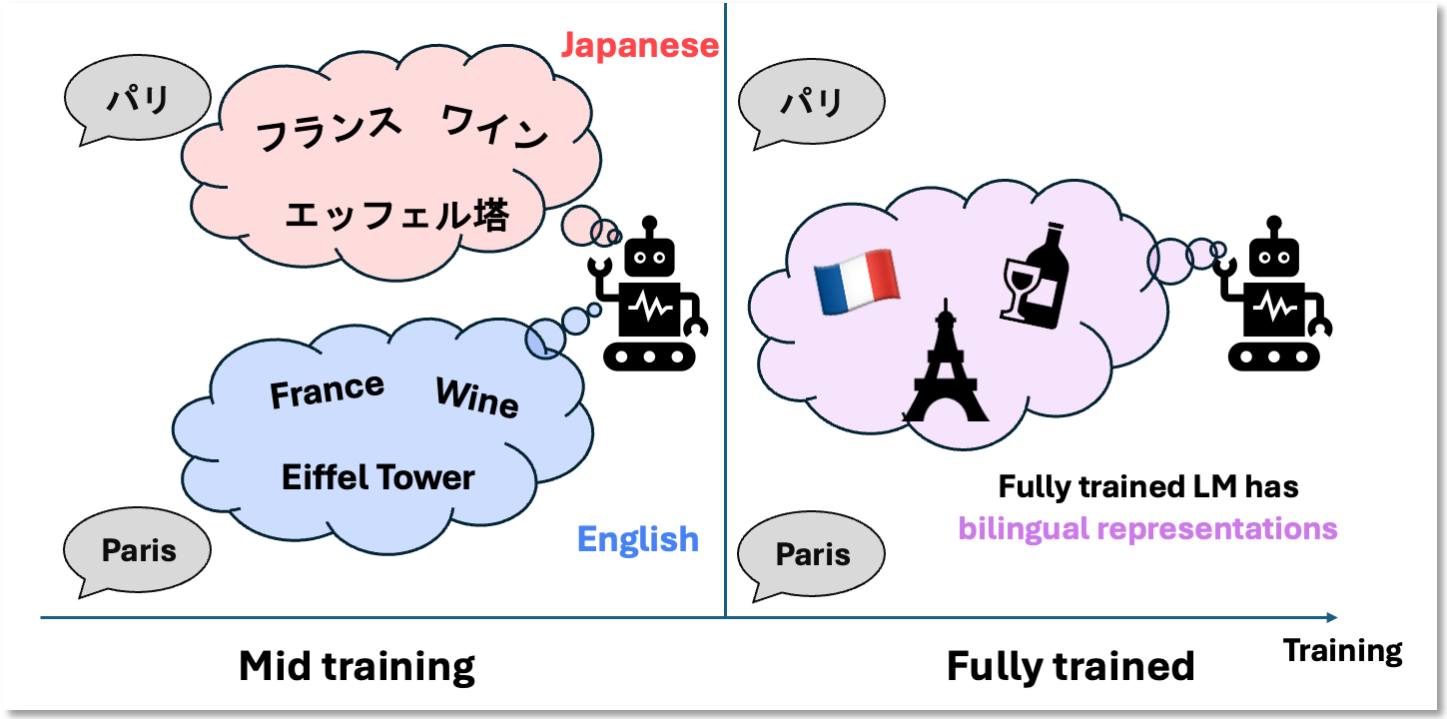
・It was last assigned to the ・The channel assigns series ・に割り当てられており、	Mixed	0.85
・different ritual and social ・as a ceremonial or heraldic ・のような儀式用の穀物	Mixed	0.62

結果: 学習ステージ方向 (3.7B, L=14)

Mohamed bin Zayed University of Artificial Intelligence

まとめ

- 学習初期~中期: **言語内の意味**を捉えた内部表現を獲得
- 学習中期~後期: **言語を超えた意味**を捉えた内部表現を獲得



結果: 層 / モデルサイズ 方向

Mohamed bin Zayed University of Artificial Intelligence

- **層方向** / **モデルサイズ方向**でも同様の分析を実施

層方向 (Fully-trained, 3.7B)

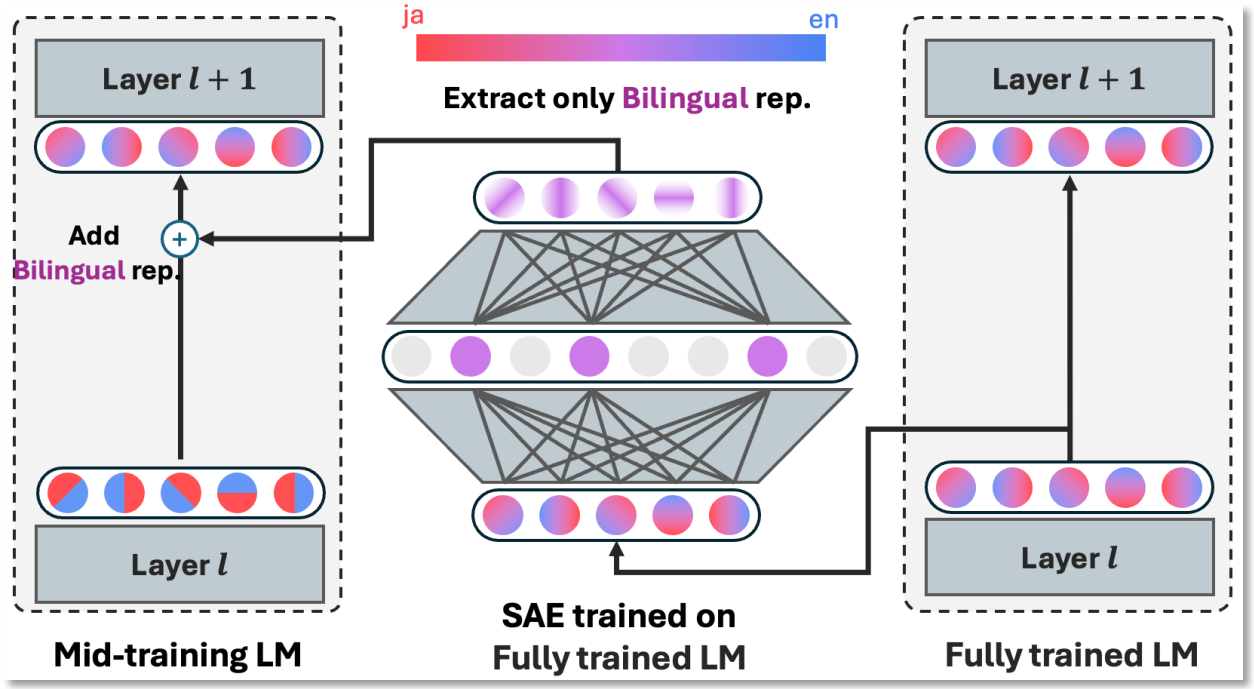
Layer (region)	Mixed lang. proportion	Monosemanticity	Findings
Lower layers	High	Low	Contain many polysemantic features
Middle layers	Medium	High	Contain many bilingual features
Upper layers	Low	Medium	Contain many contextual features (see our paper for details)

モデルサイズ方向 (Fully-trained, middle-layers)

Model sizes	Mixed lang. proportion	Monosemanticity	Findings
Small	Low	Medium	Contain almost no bilingual features
Large	Medium	High	Contain many bilingual features

介入実験: 仮説と手法

- 仮説: 「バイリンガルな表現がモデルの性能に大きな影響を与える」
- SAE を使用して特定の種類の表現を学習済みモデルから抽出し (下図右)、学習途中のモデルに加えて性能変化を測定



介入実験: 結果

Mohamed bin Zayed University of Artificial Intelligence

- Perplexity と LLM-jp-eval の翻訳タスクで評価
- バイリンガルな表現を注入した時の性能向上度合いが一番大きかった
- 日本語の表現を注入すると日本語の性能が上がる（英語も同様）

	Perplexity (↓)			COMET-22 (↑)	
Add	En	Ja	all	En → Ja	Ja → En
-	18.70	25.78	22.43	61.1	56.4
En	18.53	25.64	22.28	61.4	56.9
Ja	18.65	25.33	22.17	61.3	**57.2**
Bi	**18.36**	**25.20**	**21.96**	**62.5**	**57.2**

まとめ（再掲）

Mohamed bin Zayed University of Artificial Intelligence

- 学習ステージ・層・モデルサイズの変化に伴い LM 内部表現がどう変化するか

Findings

(i) 言語を別々に習得 → 言語間の対応を学習

(ii) 中間層付近が言語間の対応をより獲得

(iii) 大きいモデルほどバイリンガルで文脈の意味をとらえた表現を獲得

(iv) 言語間の対応に関する表現は性能に大きな影響