

スパースオートエンコーダを用いた 大規模言語モデルのチェックポイント横断分析

稲葉 達郎, 乾 健太郎, 宮尾 祐介, 大関洋平, Benjamin Heinzerling*, 高木 優*

第17回LLM勉強会, 2025/3/25

コード・デモ([url](#)) →

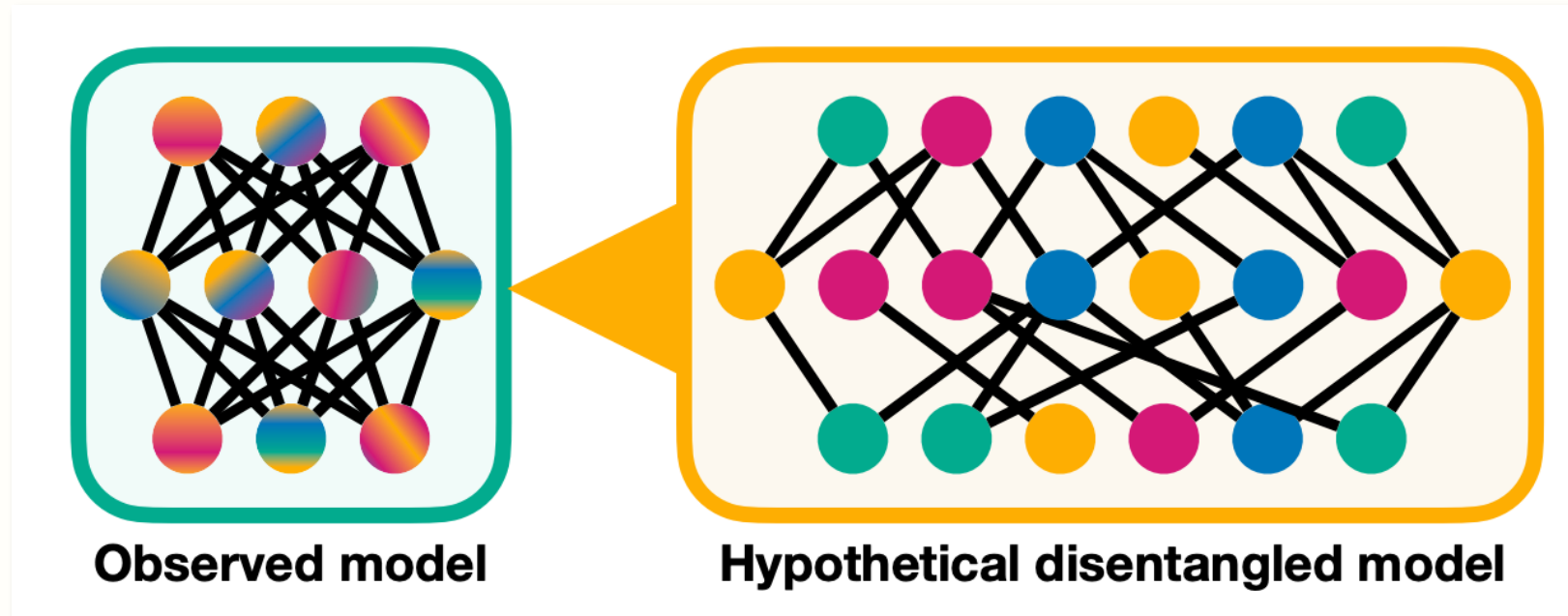


研究概要

- 大規模言語モデルの内部表現が含む情報が学習経過に伴いどう変化するか
- 特徴量抽出機のスパースオートエンコーダ (SAE)を使用して分析
- 結果
 - 言語を個別に学習後、言語間の対応関係を習得していそう
 - トークンレベルの知識を学習後、抽象度の高い概念レベルの知識を習得していそう

言語モデル解釈の難しさ [Bereska+, 24]

- 言語モデルの内部表現は **Polysemantic** (多義的) で解釈するのが難しい
- 内部表現を **Monosemantic** (一義的) な表現の足し合わせで表現したい



Polysemantic な表現のもつれを解いて Monosemantic に分解したい

スパースオートエンコーダ (SAE) [Olshausen+, 97; Huben+, 23]

- Polysemantic な表現を Monosemantic な表現の足し合わせに分解
- 中間層が**疎**になるように制約をかけた**オートエンコーダ**
 - **特徴量抽出機**の一種

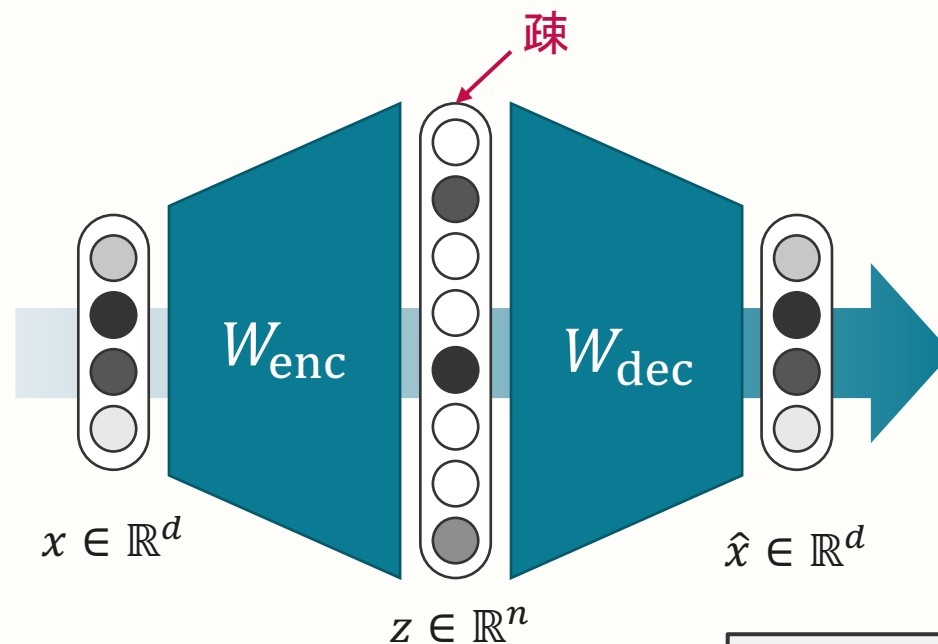
定式化

$$z = \text{ReLU} \left(W_{\text{enc}}(x - b_{\text{pre}}) \right)$$

$$\hat{x} = W_{\text{dec}}z + b_{\text{pre}}$$

損失

$$\mathcal{L} = \underbrace{\|x - \hat{x}\|_2^2}_{\text{再構成損失}} + \underbrace{\lambda \|z\|_1}_{\text{疎にする制約}}$$



$$\begin{aligned} x, b_{\text{pre}} &\in \mathbb{R}^d, z \in \mathbb{R}^n, d < n \\ W_{\text{enc}} &\in \mathbb{R}^{n \times d}, W_{\text{dec}} \in \mathbb{R}^{d \times n} \end{aligned}$$

※ b_{pre} と ReLU は省略して図示

実験設定

LLM-jp-3-1.8B の12層目の表現を**チェックポイント横断**で分析

- SAE のバリエーションの一つ TopK-SAE を使用 ($\frac{n}{d} = 16, K = 32$)
- SAE で抽出した特徴量の分布を見ることでモデルが内部で扱う情報がどう進化するか

特徴量の分布

- 言語傾向: 日本語 or 英語 or 日英混合
- 意味粒度: トークンレベル(**猫**と**猫**) or 概念レベル(**猫**と**犬**) or 無関係

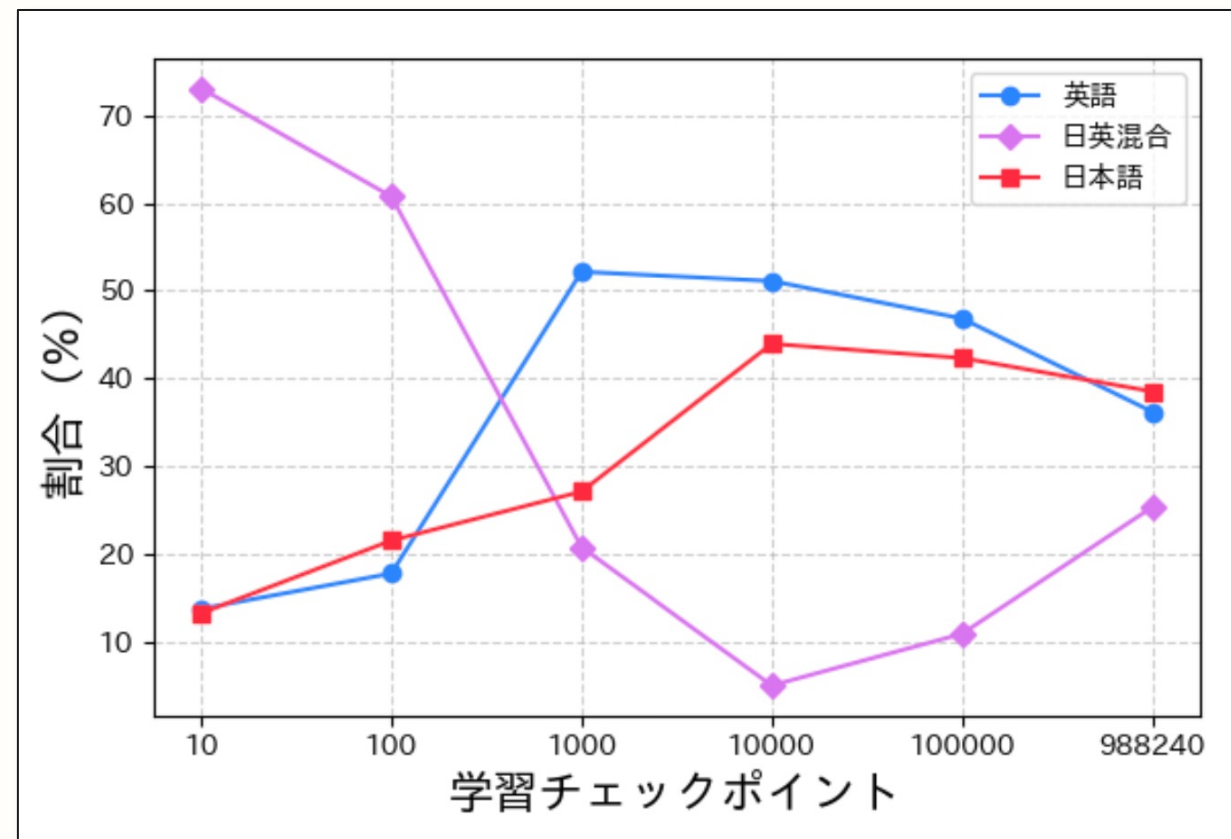
Feature(特徴量)の可視化

- (a) ckpt=100, (b) ckpt=10000, (c) ckpt=988240 での活性化パターン例

	Feature番号	Featureが活性化する文章例	言語傾向 (4.3章)	意味粒度 (4.4章)
(a)	F _{ckpt=100} #00002	・) は、「日本の貴婦人 ・ または HMG-CoA レダクターゼ ・ called radiological pollution, is	日英混合	無関連
	F _{ckpt=100} #00004	・ から20世紀前半にかけて ・ は日本の防衛官僚。 ・ investigations are performed by geotechnical	日英混合	無関連
(b)	F _{ckpt=10000} #00004	・ dorsalis), also known as the scrub ・ regnans, known variously as ・ nerve) also known as the fourth	英語	トークンレベル: 「known」
	F _{ckpt=10000} #00009	・ 石油生産設備から ・ 冷暖房設備、冷凍冷蔵設備、動力設備又は ・ のプラント設備を	日本語	トークンレベル: 「設備」
(c)	F _{ckpt=988240} #00009	・ ここで言う「都市」には ・ , where fluency is defined as linguistic ・ "Arbitrary" here means that the	日英混合	概念レベル (同義): 「定義」
	F _{ckpt=988240} #00016	・ 特有の臭気のある白色個体で、 ・ 物で、白色の粉末である ・ It is a colorless liquid with a smell reminiscent	日英混合	概念レベル (意味的共通性): 「物質の特性」

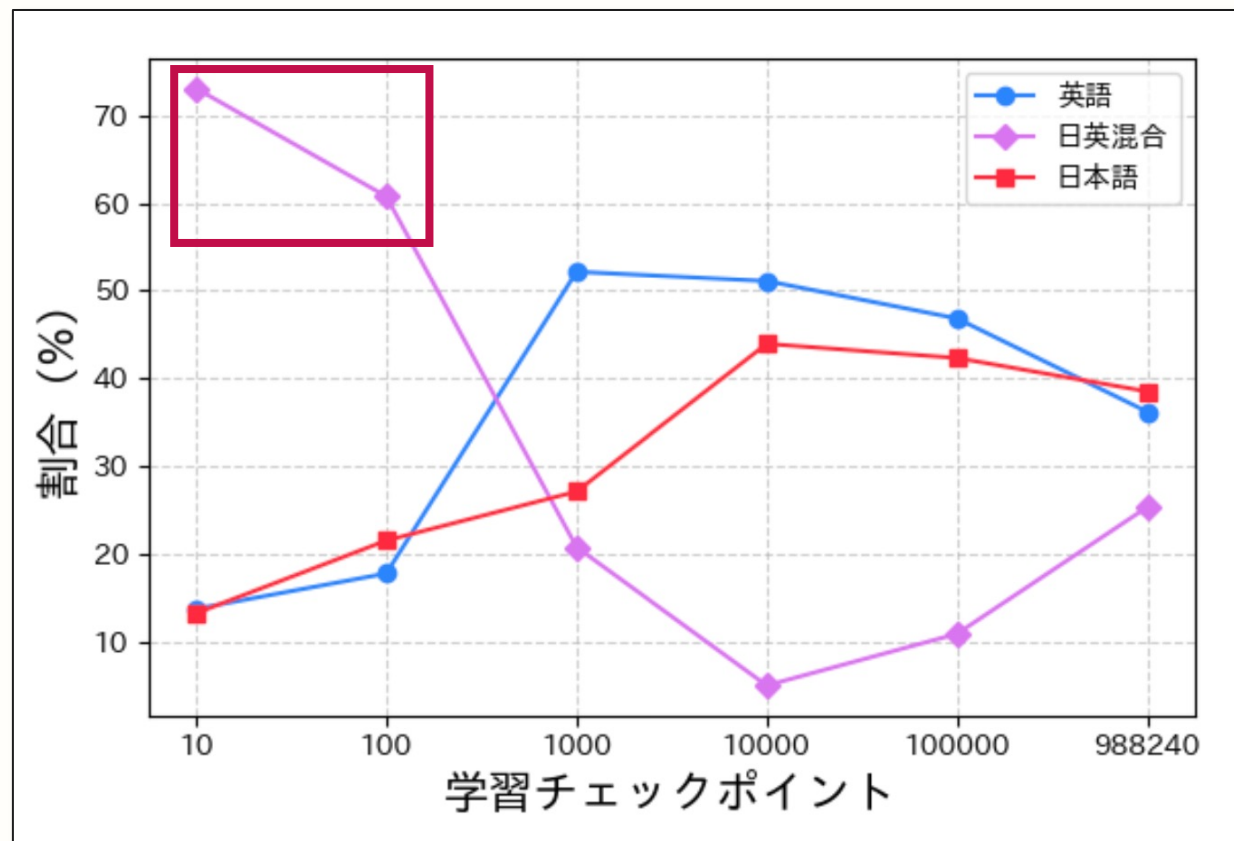
活性化パターンの言語傾向推移

- 学習初期は日英混合 Feature が大半
 - 無作為なトークンに活性化
- 英語Feature と日本語Feature は学習中期に増加
- 日英混合 Feature は学習経過で一旦減り学習後期に再び増加
 - 日英間で同一の意味を持つトークンや文章に対して活性化



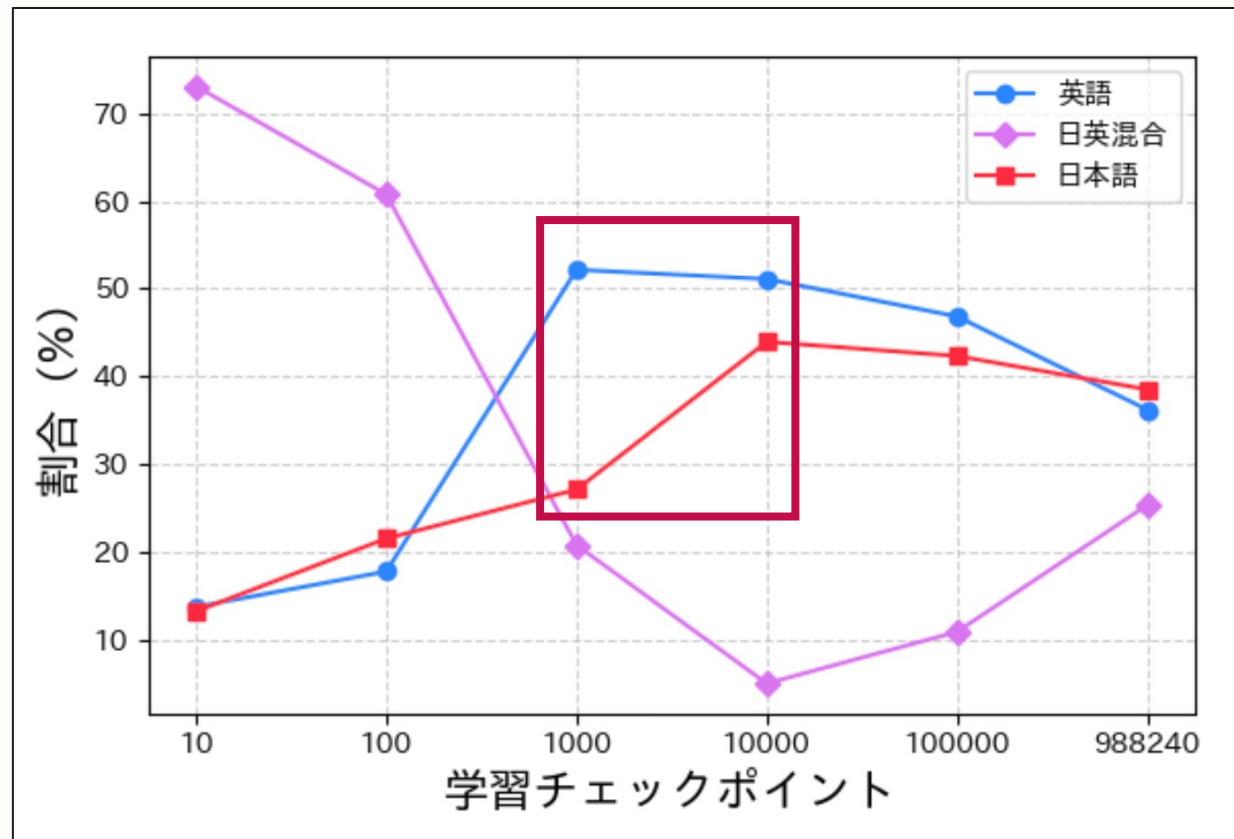
活性化パターンの言語傾向推移

- 学習初期は日英混合 Feature が大半
 - 無作為なトークンに活性化
- 英語Feature と日本語Feature は学習中期に増加
- 日英混合 Feature は学習経過で一旦減り学習後期に再び増加
 - 日英間で同一の意味を持つトークンや文章に対して活性化



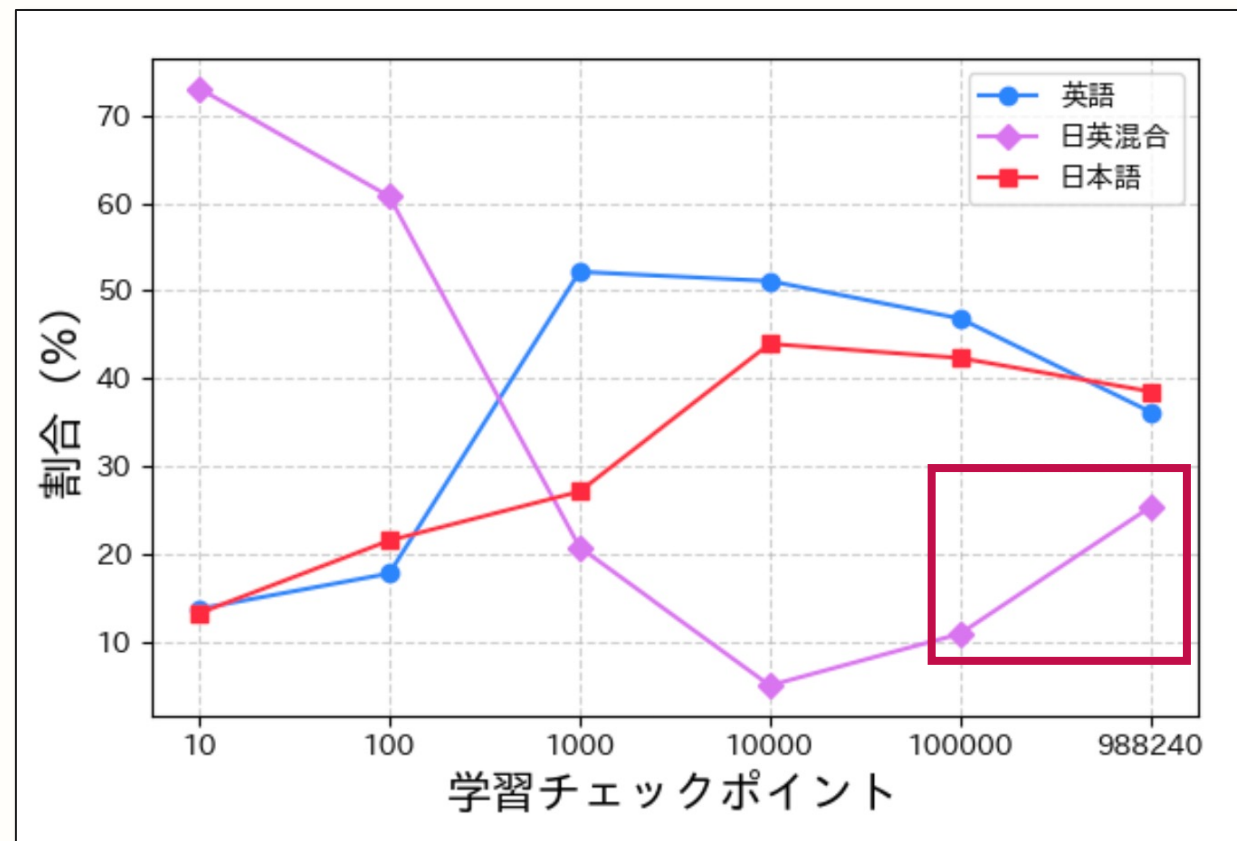
活性化パターンの言語傾向推移

- 学習初期は日英混合 Feature が大半
 - 無作為なトークンに活性化
- **英語Feature と日本語Feature は学習中期に増加**
- 日英混合 Feature は学習経過で一旦減り学習後期に再び増加
 - 日英間で同一の意味を持つトークンや文章に対して活性化



活性化パターンの言語傾向推移

- 学習初期は日英混合 Feature が大半
 - 無作為なトークンに活性化
- 英語Feature と日本語Feature は学習中期に増加
- 日英混合 Feature は学習経過で一旦減り学習後期に再び増加**
 - 日英間で同一の意味を持つトークンや文章に対して活性化**



活性化パターンの言語傾向推移

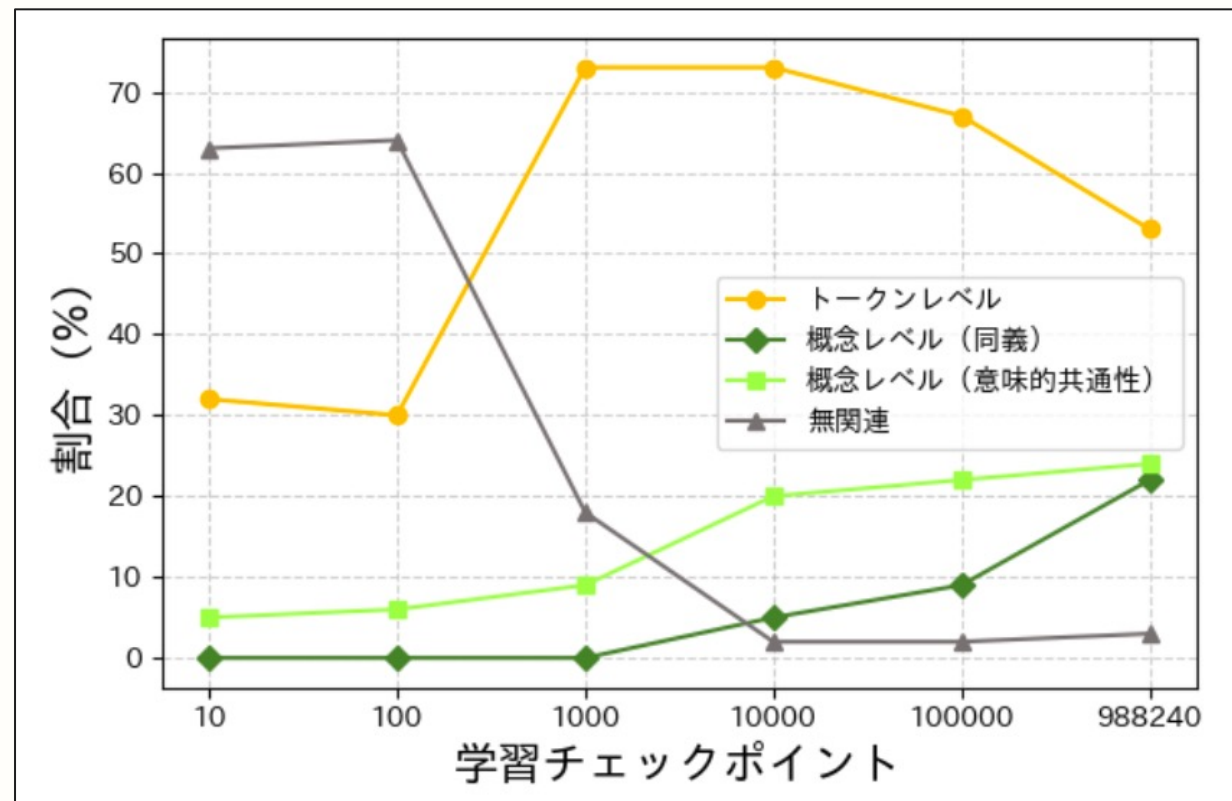
- 学習初期は日英混合 Feature が大半
 - 無作為なトークンに活性化
- 英語Feature と日本語Feature は学習中期に増加
- 日英混合 Feature は学習経過で一旦減り学習後期に再び増加
 - 日英間で同一の意味を持つトークンや文章に対して活性化



1. **学習初期から中期**にかけて**言語別**にトークンや文章の意味を習得
2. **学習中期から後期**にかけてトークンや文章の**言語間での対応関係**を習得

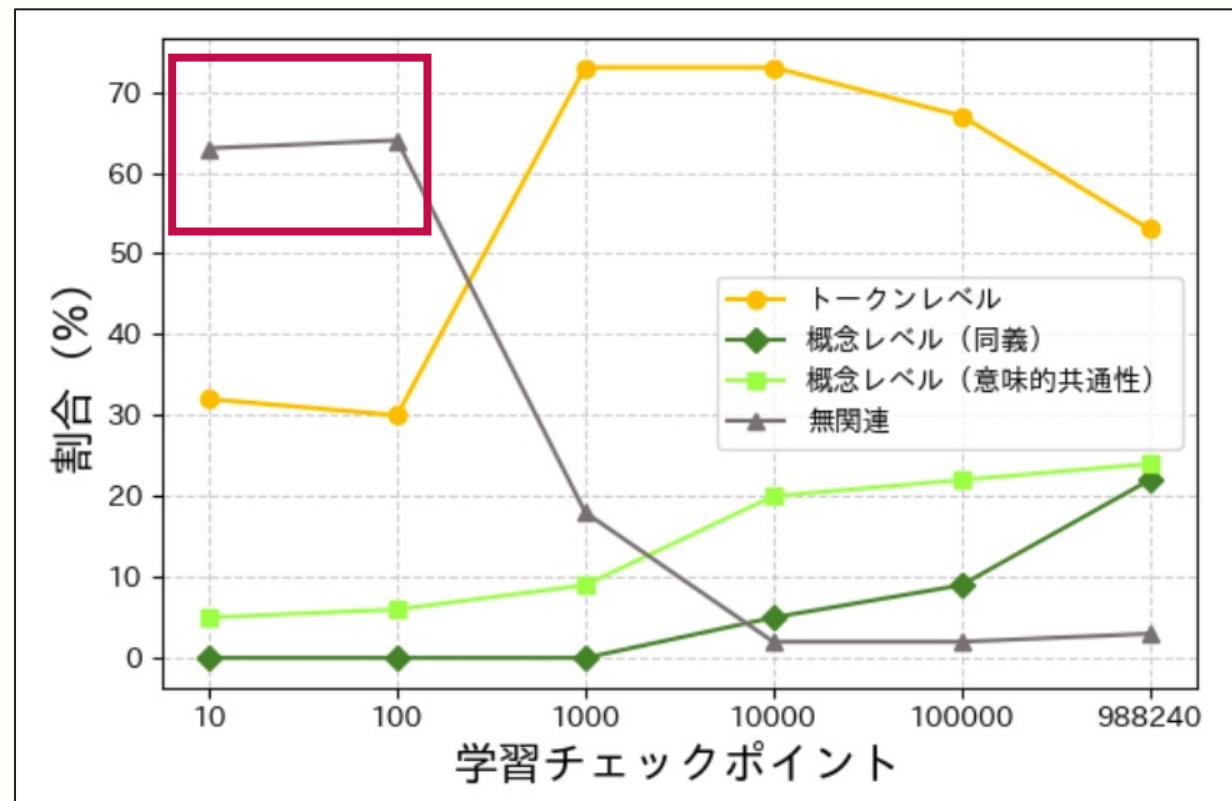
活性化パターンの意味粒度推移

- 学習初期は無関連 Feature が大半
- 学習中期にトークンレベル Feature が増加
- 学習後期にかけ概念レベル Feature 増加



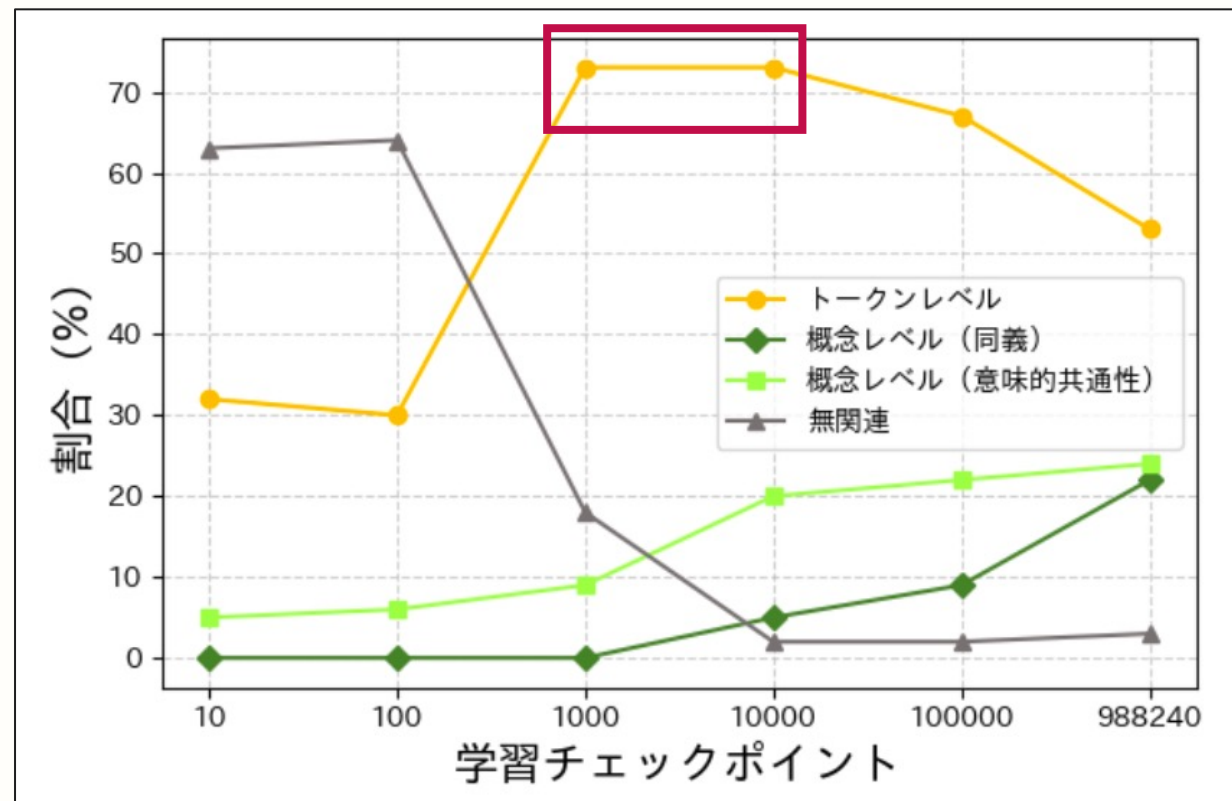
活性化パターンの意味粒度推移

- 学習初期は無関連 Feature が大半
- 学習中期にトークンレベル Feature が増加
- 学習後期にかけ概念レベル Feature 増加



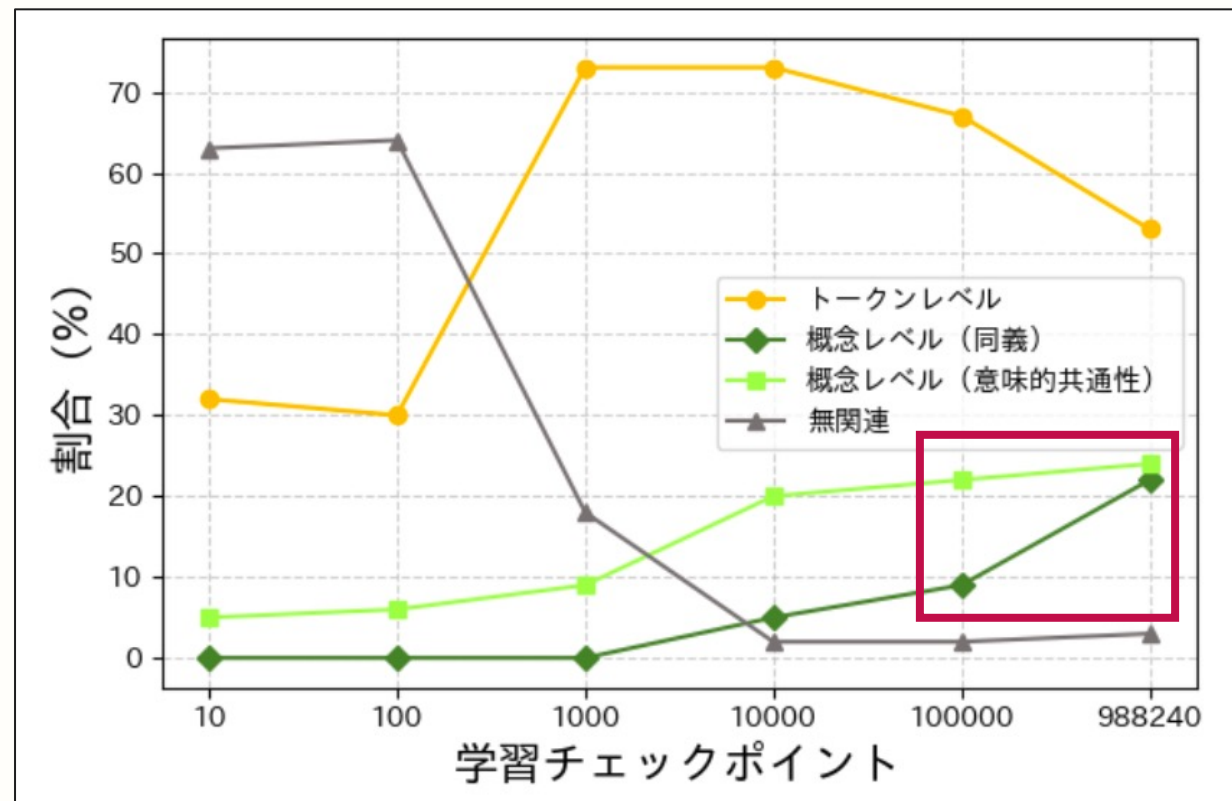
活性化パターンの意味粒度推移

- 学習初期は無関連 Feature が大半
- **学習中期にトークンレベル Feature が増加**
- 学習後期にかけ概念レベル Feature 増加




活性化パターンの意味粒度推移

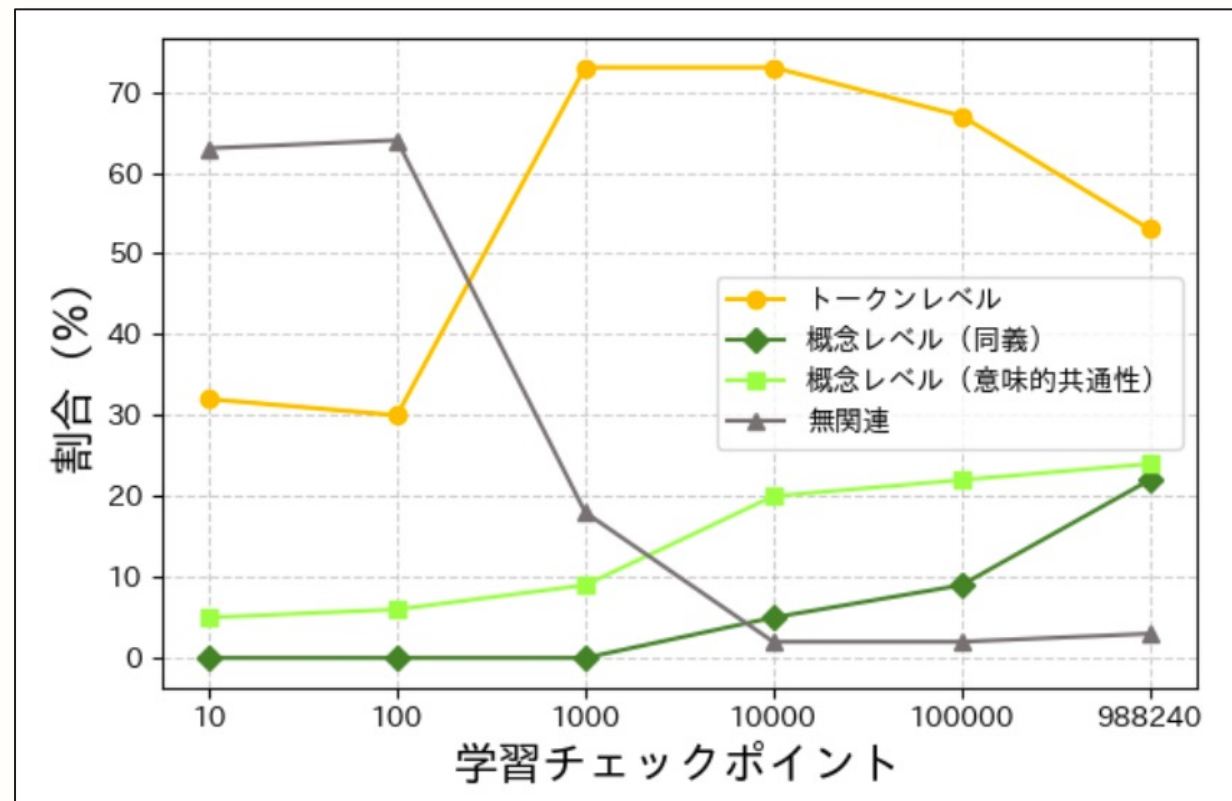
- 学習初期は無関連 Feature が大半
- 学習中期にトークンレベル Feature が増加
- **学習後期にかけ概念レベル Feature 増加**



活性化パターンの意味粒度推移

- 学習初期は無関連 Feature が大半
- 学習中期にトークンレベル Feature が増加
- 学習後期にかけ概念レベル Feature 増加

- 
1. 学習初期から中期にかけて**トークンレベル**の知識を習得
 2. 学習中期から後期にかけて**概念レベル**の知識を習得



まとめ

- スパースオートエンコーダを用いてチェックポイント横断で LLM を分析した
- その結果以下の可能性が示唆された
 - LLM は個別言語でのトークンや文章の意味を習得し，その後言語間の対応関係を理解すること
 - トークンレベルの知識を習得した後に概念レベルの知識体系を構築している