スパースオートエンコーダーを用いた大規模言語モデルによる文書内トピック解釈の分析

加藤由芽 [1]　小林一郎 [1]
[1] お茶の水女子大学
{kato.yume,koba}@is.ocha.ac.jp

概要

本研究は，大規模言語モデル（LLM）が文書のトピック情報を内部でどのように表現しているかを分析することを目的とする．LLM の中間層表現を Sparse Autoencoder（SAE）により分解し，得られた潜在特徴に対して潜在的ディリクレ配分法（LDA）を適用することで，内部特徴と文書トピックとの対応関係を検証した．実験の結果，得られたトピック分布は文書ラベルと一定の対応を示した一方，同一ラベル内でも複数の異なるトピック分布が存在し，またトピック間の距離構造は文書ラベル間の類似性を必ずしも反映しないことが明らかとなった．これは, LLM における文書トピック表現が，人間が定義する意味的カテゴリより細かく, 微細なニュアンスを捉えたものになっている可能性を示唆する．

1　はじめに

大規模言語モデル（LLM）は，自然言語処理の多くのタスクにおいて高い性能を示している一方で，ハルシネーションやバイアスといった問題が指摘されている [1]．これらの問題に対処し，LLM を安全かつ信頼可能に利用するためには，モデル内部の表現や出力根拠を人間が解釈可能な形で理解することが重要である．

解釈性研究の文脈においては, 単一のニューロンが複数の無関係な概念に反応する「多義性」が大きな課題として知られている [2]. この問題に対処する手法として, Sparse Autoencoder（SAE）[3, 4] が提案されており, LLM の内部表現を高次元かつ疎な潜在空間に分解することで, 多義的なニューロン表現を単一概念に対応する特徴へと分離することが可能であることが示されている [5]. これにより, 概念レベルでの解釈性は一定程度向上している．

一方で，SAE によって抽出された多数の潜在特徴が，文書全体の意味構造，とりわけ文書レベルのトピック情報をどのように表現しているかについては，十分に明らかにされていない．先行研究では，意味的に近い特徴同士が潜在空間上で局所的な近傍構造を形成することが示唆されている [6] が，これらの特徴群と文書トピックとの対応関係は体系的に検証されていない．

本研究では，SAE によって抽出された特徴の共起構造には，文書トピックを反映した潜在的な構造が存在するという仮説を立てる．この仮説のもと，SAE 特徴の共起情報に対して潜在的ディリクレ配分法（LDA）[7] を適用し，得られたトピックと文書ラベルとの対応関係を分析することで，LLM 内部における文書トピック表現の性質を明らかにすることを目的とする．

2　関連研究

LLM の内部表現を理解するための研究として，単一のニューロンの意味的解釈を試みるものがあるが，ニューロンが複数の無関係な概念に反応する「多義性」が課題となっている [2]．これに対して，Bricken ら [8] や Cunningham ら [9] は，Sparse Autoencoder（SAE）を用いた内部表現の分解によって，多義的な反応をより一貫した特徴へと分離する手法を提案している．SAE は高次元かつ疎な潜在空間を介して内部表現を再構成することで，相対的に意味的一貫性の高い特徴を抽出する．この方法は主要な LLM 開発企業でも研究が進み，事前学習済み SAE の公開が行われている [10, 11].

Templeton ら [12] は Claude 3 Sonnet に SAE を適用し，明確な意味を持つ特徴の抽出に加えて，意味的に近い特徴同士が潜在空間上で局所的に近接して配置されることや，下流タスクへの影響を示した．この結果は，SAE によって得られる特徴空間が単なる個別概念の集合にとどまらず，より高次の意味構造

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

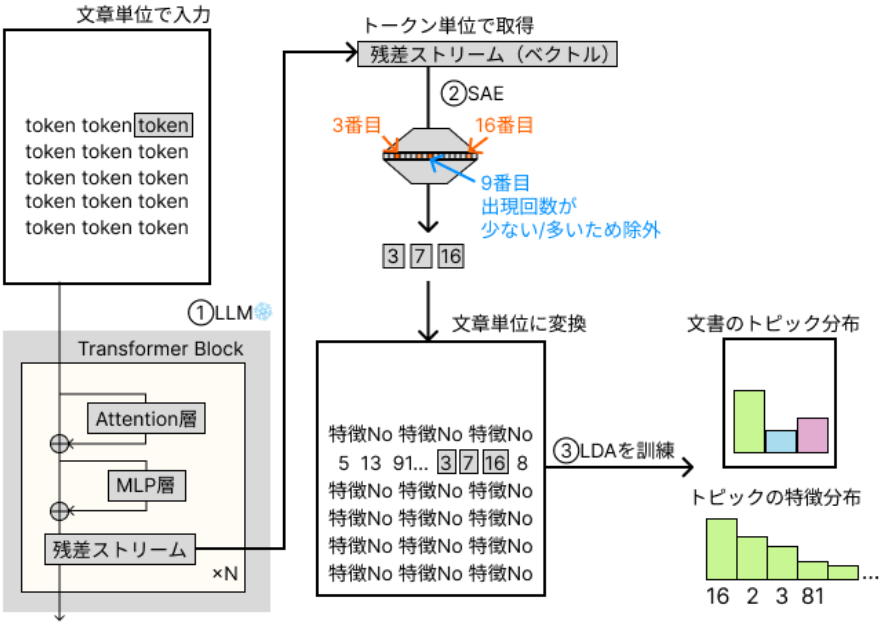

図 1　本研究における解析手法の全体像.

を内包している可能性を示唆している.

特徴全体の構造に着目した研究として，Li ら [6] が挙げられる．彼らは SAE 特徴の共起関係に基づいてクラスタリングを行い，その結果を Uniform Manifold Approximation and Projection（UMAP）[13] により可視化した．これにより，数学・コード・対話といった異なるドメインごとに意味的一貫性を持つ局所構造が形成されることが示唆された.

しかし，これらの研究は主に特徴空間の幾何学的構造や視覚的なまとまりを報告しており，抽出されたクラスタや構造が文書レベルの意味情報，特にトピック構造とどのように対応するかは十分に検証されていない．本研究では，SAE 特徴の共起に基づくトピック分析を通じて，これらの構造と文書ラベルとの対応関係を明らかにすることを目指す.

3　提案手法

3.1　研究概要

本研究では，LLM の内部表現を SAE によって分解し，得られた SAE 特徴の共起構造に基づいて文書トピック表現を分析する．具体的には，各文書において活性化した SAE 特徴を共起情報として表現し，その共起構造に対して潜在的ディリクレ配分法（LDA）を適用することで，SAE 特徴の確率分布として表されるトピックを推定する.

得られたトピックについて，文書レベルおよびトピックレベルの双方から分析を行い，SAE 特徴空間において文書トピックがどのように表現されているかを検証する．本研究の解析フローを図 1 に示す.

3.2　Sparse Autoencoder（SAE）

SAE[3, 4] は，LLM の中間表現を高次元かつ疎な潜在特徴へ分解・再構成する 2 層のオートエンコーダである．本研究では，LLM の残差ストリームを入力とする SAE を用い，エンコーダ出力の活性値が 1 を超えた場合に当該特徴が活性化したと定義する．これにより，各文書は「活性化した SAE 特徴の集合」として表現される.

3.3　潜在的ディリクレ配分法（LDA）

LDA[7] は，文書集合に潜在するトピック構造を確率的に推定するモデルであり，各文書をトピック分布，各トピックを要素分布として表現する．本研究では，単語の代わりに SAE 特徴を用い，各文書における特徴の共起情報に基づいて LDA を適用する．これにより，トピックは SAE 特徴の確率分布として推定される.

3.4　文書表現とトピック表現

LDA の結果として，各文書にはトピック分布が，各トピックには SAE 特徴の確率分布が対応付けられる．本研究では，これらの分布を用いて，文書ラベルとトピックの対応関係，およびトピック間の類

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

似構造を分析する．

4 実験

4.1 実験設定

モデル 先行研究 [6] に従い，Gemma-2-2b [14] の残差ストリームに対して事前学習された Sparse Autoencoder（SAE）`gemma-scope-pt-res` を用いる [11]．本 SAE は特徴数 16,384 を持ち，検証データにおける平均 L_0 スコアは 42 である．

データセット News Article Category Dataset[15] より，EDUCATION, SCIENCE, SPORTS の 3 カテゴリを使用する．各カテゴリから 600 サンプルずつ抽出し，合計 1,800 文書を用いた．

前処理 HTML タグ・絵文字・URL の除去を行なった．Kaggle 上の公開コード[1)]を参考にした．

ブロック・閾値 各データサンプルから先頭 256 トークンを 1 ブロックとして切り出し，そのブロック内で発火した SAE 特徴を LDA に入力する 1 つの文書として扱う．ここで特徴の「発火」とは，SAE の活性値が 1 より大きくなることを指す．発火文書割合が 35%を超える特徴，または発火文書数が 100 未満の特徴は除外した．

LDA のトピック数 トピック数を 2 から 20 まで 2 刻みで変化させて LDA を学習し，perplexity に基づくエルボ法により最適なトピック数を評価した．その結果，本研究ではトピック数を 8 に設定した．

分析に用いた層 先行研究 [16] では複数層を用いた分析が行われているが，層間での定性的差異は限定的であった．本研究では分析コストを考慮し，第 19 層の残差ストリームを分析対象とした．

5 結果と分析

5.1 文書レベルでのトピック分布

各文書に対して推定されたトピック分布を用い，文書ラベルとの対応関係を分析した．まず，文書ラベルごとにトピック分布を平均化した結果，多くのトピックにおいて特定の文書ラベルに対する確率が相対的に高くなる傾向が確認された（図 2）．一方で，トピック 0 およびトピック 3 は全体での出現頻度が低く，文書ラベルとの対応関係は相対的に弱かった．

1） https://www.kaggle.com/code/whdhdyt/news-article-category-classification

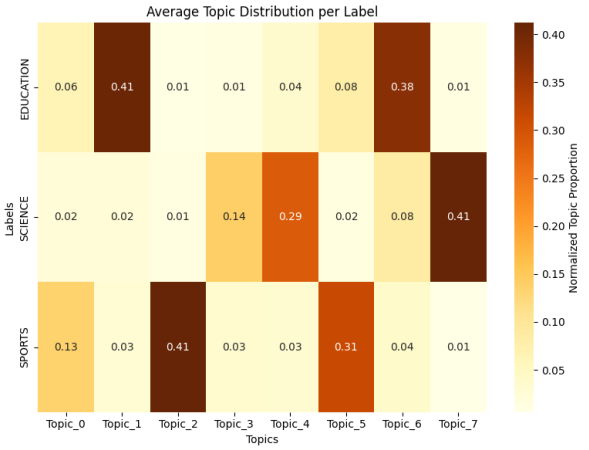

図 2 各文書のトピック分布の文書ラベルごとの平均．

Topic	Topic Interpretation
0	sports, legal issues, and international events
1	education and economic policy themes
2	sports and competition related contexts
3	semantic themes in discussions and achievements
4	semantic exploration of social and theoretical contexts
5	social issues and emotional narratives
6	educational and emotional engagement factors
7	astronomy and environmental science concepts

表 1 それぞれのトピックにおける，代表特徴の解釈を用いた，トピックの解釈

次に，各文書のトピック分布を UMAP により 2 次元空間に可視化した結果を図 4 に示す．UMAP 空間上では，文書ラベルごとに分布の分離が概ね確認できる一方で，異なるラベル間に連続的な遷移も観察された．

5.2 トピックレベルでの特徴分布および解釈

各トピックに対応する SAE 特徴分布間の類似度を，Jensen–Shannon Divergence（JSD）により算出し，その距離行列を多次元尺度構成法（MDS）を用いて 2 次元に投影した（図 3）．本可視化は LDAvis [17] で提案された手法に基づく．その結果，トピック 0 とトピック 3，およびトピック 5 とトピック 7 の間で比較的高い類似度が確認された一方，その他のトピック間では一定の距離が保たれていた．

さらに，各トピックについて確率の高い SAE 特徴を代表特徴として選出し，それらの発火トークン列をもとに LLM を用いてトピックの解釈を行っ

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

図 3 JSD に基づくトピック間距離の MDS 可視化（LDAvis）．各円はトピックを表し，円間距離は特徴分布の類似度を反映する．円の面積はコーパス内でのトピック出現頻度に比例する．

た（表 1）．その結果，トピック 0 およびトピック 2 では SPORTS，トピック 1 およびトピック 6 では EDUCATION，トピック 7 では SCIENCE に関連する概念が抽出された．一方で，トピック 3，4，5 については，特定の文書ラベルに限定されない，ニュース記事全般に共通する広範な概念が得られた．

6 考察

本研究では，SAE 特徴の共起構造に基づいて LDA を適用することで，LLM 内部における文書トピック表現の性質を分析した．その結果，文書ラベルとトピックとの間には一定の対応関係が確認され，トピック解釈においても文書カテゴリと関連する概念が抽出された．これらの結果は，LLM 内部の SAE 特徴空間において，文書トピックに関する情報がある程度明示的に表現されている可能性を示唆している．

特に，UMAP による可視化において，各文書ラベルごとにおおよそ 2 つの高密度領域が観察された点は示唆的である。これらの領域は明確に分離したクラスタというよりも，連続的に遷移する構造を持っており，同一の文書ラベルに属する文書であっても，内部表現上では複数の異なるトピック構造として表現されている可能性がある．

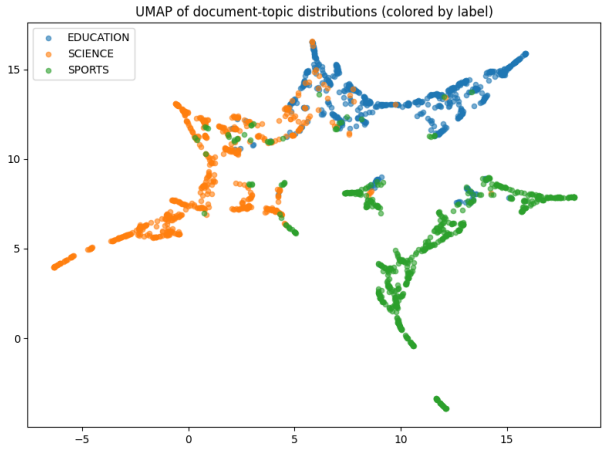

図 4 各文書のトピック分布の UMAP による 2 次元写像．一点一点が一つの文書である

さらに，トピック間の特徴分布に基づく類似度分析では，特徴分布が近接するトピックが必ずしも同一の文書ラベルに対応せず，逆に同一ラベルに対応するトピック同士が近接しない場合も確認された．このことは，LLM におけるトピック表現が，文書カテゴリそのものだけでなく，文体的特徴や意味的ニュアンスといった要素を含んで形成されている可能性を示唆している．

7 結論

本研究では，SAE 特徴の共起構造に基づいて LDA を適用することで，LLM 内部における文書トピック表現の構造を分析した．その結果，得られたトピックの多くが文書ラベルと一定の対応関係を示し，トピック解釈においても文書カテゴリと関連する概念が確認された．このことから，LLM 内部の SAE 特徴空間には，文書トピックに関する情報がある程度明示的に表現されている可能性がある．一方で，トピック分布の可視化およびトピック間類似度の分析からは，同一の文書ラベルに対応するトピックが必ずしも近接して配置されないことや，各ラベル内で多峰的な分布構造が形成されることが確認された．これらの結果は，LLM におけるトピック表現が，人間が定義する文書カテゴリとは異なる粒度や観点に基づいて形成されている可能性を示している．

今後の課題としては，トピック数や特徴選択条件の変更による影響の検証，トピックに寄与する特徴の定量的評価手法の検討，およびトピック解釈の再現性向上が挙げられる．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は科研費（23K28143）の支援を受けた．ここに深謝する．

参考文献

[1] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence, 2025.

[2] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

[3] Marc' aurelio Ranzato, Y-lan Boureau, and Yann Cun. Sparse feature learning for deep belief networks. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, **Advances in Neural Information Processing Systems**, Vol. 20. Curran Associates, Inc., 2007.

[4] Andrew Ng. Sparse autoencoder. Cs294a lecture notes, Stanford University, 2011.

[5] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. **Transformer Circuits Thread**, 2024.

[6] Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. **Entropy**, Vol. 27, No. 4, p. 344, March 2025.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. **J. Mach. Learn. Res.**, Vol. 3, pp. 993–1022, 2003.

[8] Trenton Bricken, et al. Towards monosemanticity: Decomposing language models with dictionary learning. **Transformer Circuits Thread**, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[9] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

[10] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024.

[11] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024.

[12] Adly Templeton, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. **Transformer Circuits Thread**, 2024.

[13] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[14] Gemma Team, et al. Gemma 2: Improving open language models at a practical size, 2024.

[15] Rishabh Misra. News category dataset. **arXiv preprint arXiv:2209.11429**, 2022.

[16] Yume Kato Ichiro Kobayashi. An analysis of document topic features in a large language model using a sparse autoencoder. In **The 26th International Symposium on Advanced Intelligent Systems(ISIS2025)**, November 2025.

[17] Carson Sievert and Kenneth Shirley. LDAvis: A method for visualizing and interpreting topics. In Jason Chuang, Spence Green, Marti Hearst, Jeffrey Heer, and Philipp Koehn, editors, **Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces**, pp. 63–70, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

A　LLM 解釈における諸設定

本研究では，SAE 特徴およびトピックの解釈において，大規模言語モデル（LLM）を補助的な分析手段として用いた．本付録では，その際に用いたモデルおよびプロンプト設定の詳細を記す．

A.1　用いたモデル

特徴およびトピックの解釈には，OpenAI が提供する `gpt-4o-mini` を用いた．すべての出力は API を介して取得しており，モデルの温度やその他の生成パラメータはデフォルト設定を使用した．

A.2　プロンプト設計

LLM による解釈の一貫性および過剰な推測を防ぐため，以下のシステムプロンプトを全ての問い合わせに共通して使用した．

```
You analyze internal features of a language model.
Be conservative. If evidence is weak or inconsistent,
     answer "unclear".
Do not guess or hallucinate.
Respond in valid JSON only.
```

A.2.1　SAE 特徴の解釈

各 SAE 特徴について，その特徴が最も強く活性化したトークンを含む短いテキスト断片を入力とし，特徴の意味的性質を解釈させた．強く活性化したトークンは ** ** により明示した．

使用したプロンプトは以下の通りである．

```
Below are short text snippets where a single internal
     feature z
of a language model activates strongly.
The strongly activated token is marked with ** **.

Examples:
{examples}

Task:
1. Briefly describe what this feature responds to (max 3
     words).
2. Choose one category:
   - semantic (topic or meaning)
   - syntactic (structure, quotation, formatting)
   - unclear (cannot determine)

Output format (JSON only):
{
  "interpretation": "...",
  "category": "semantic | syntactic | unclear"
}
```

この出力により，各特徴が主に意味的情報を捉えているか，あるいは構文的・形式的な性質を捉えているかを判別した．

A.2.2　トピックの解釈

トピック（LDA クラスタ）の解釈には，同一クラスタに属する複数の代表 SAE 特徴の解釈結果を入力とし，クラスタ全体としての共通概念を要約させた．

使用したプロンプトは以下の通りである．

```
Below are interpretations of features belonging to the
     same cluster
in a language model.

Cluster ID: {cluster_id}

Feature interpretations:
{items}

Task:
Provide a short cluster-level summary as a noun phrase
     (max 10 words).
Be conservative. If no clear common pattern exists,
     output "unclear pattern".

Output format (JSON only):
{
  "summary": "..."
}
```

この手順により，個々の特徴解釈を統合したトピックレベルでの解釈を得た．なお，不明確な場合には `"unclear pattern"` を出力させることで，過度な一般化を避ける設計とした．

B　トピック数 3 における補足実験

本節では，トピック数を 3 に設定した場合の結果を補足的に示す．LDA のトピック数を 3 とした上で，各文書について得られたトピック分布を三角図として可視化した．各点は 1 文書を表し，色は文書ラベルに対応する．

図 5 に示すように，文書ラベルごとに分布は概ね異なる領域に分かれており，トピック分布が文書ラベルの情報を一定程度反映していることが確認できる．一方で，各ラベルの分布は三角図の頂点付近には集中せず，むしろ辺付近に多く分布している．

これらの結果から，本手法の枠組みにおいては，文書が単一のトピックに強く支配される構造は明確には観測されなかったと言える．一方で，文書表現が複数のトピックの相対的な寄与として表現される構造が，安定して観測された．

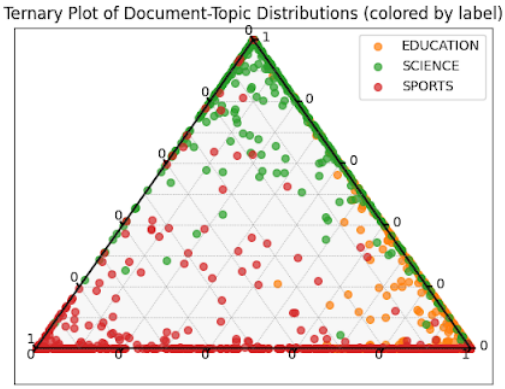

図 5　トピック数 3 における各文書のトピック分布の三角図．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).