# 文脈内学習におけるデモの親和性と多様性の提案

加藤 万理子 [1]　趙 羽風 [1]　坂井 吉弘 [1]　井之上 直也 [1,2]
[1] 北陸先端科学技術大学大学院　[2] 理化学研究所
mariko.k@jaist.ac.jp

## 概要

文脈内学習 (**I**n-**C**ontext **L**earning; ICL) において, デモンストレーション (デモ) の選択はタスク性能に大きな影響を与える. 既存研究ではデモの選択手順については研究されているが, 選択基準であるデモの性質は十分に調べられていない. 本研究では, デモの「親和性」と「多様性」という 2 つの性質を新たに提案し, その内の親和性が性質が複数のモデルおよびデータセットにおいてデモ選択に望ましい性質であることを示した. さらに, 既存手法で選ばれたデモが, 2 つの性質のタスク性能を向上させる方向へ集約していることを示し, デモ選択とタスク性能のメカニズム解明への示唆を得た.

## 1　はじめに

ICL とは, 大規模言語モデルに少数の入出力例 (デモ) と解かせる入力 (クエリ) をプロンプトとして与え, パラメータの更新を行わずにクエリに対する予測を行わせる利用法や能力を指す. 近年, タスク性能に影響するデモの性質を用いてデモ選択を行うことで, タスク性能が向上できることがわかっている [1]. 例えば, BM25 など表層の類似度を基にデモを選択する手法では [2], 『クエリとの類似性』という”デモの性質”に基づいてデモを選択しタスク性能が向上しているが, なぜ性能改善に繋がるのか十分に解明されていない. ここで我々は, 既存のデモ選択手法が, 限られた特定のデモの性質を最大化するようなデモを選択することにより, タスク性能を改善していると考えている. この仮説を検証するために, 次のような理想的なデモの性質を検討する:

**要件 1**　提案するデモの性質とタスク性能に相関があること.

**要件 2**　さまざまな既存手法が, 提案するデモの性質を改善していること.

もしこのような性質を持つデモを確認できれば, 既存のデモ選択手法がこの性質を実質的に改善することによってタスク性能を向上させていると考えることができ, この性質を分析することでデモ選択とタスク性能のメカニズムの解明に繋がる.

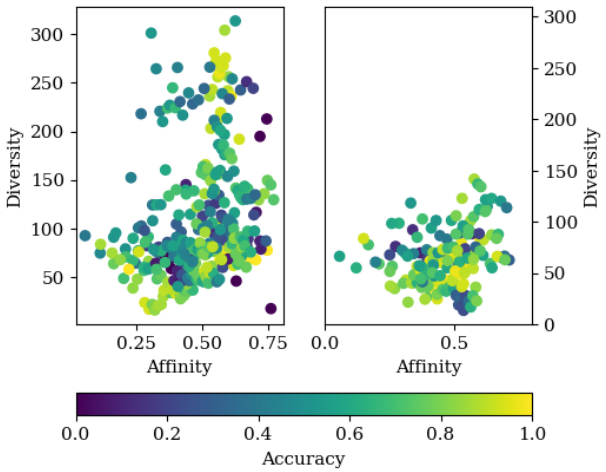


**図 1**　Falcon 7B: デモ数 16 個での親和性および多様性 (左: ランダムサンプリング, 右: BGE M3 に基づくデモ選択). 色が Accuracy を表しており, ランダムにデモを選択した場合 (左) と比較して, 類似度に基づき選択されたデモは親和性度および多様性度が収束する

本研究では, 既存のデモ選択基準 [1] に基づき, 理想的な良いデモの性質として, 内部表現上の「デモとクエリの親和性」と「デモ同士の多様性」を導入する. 具体的には, 言語モデルの推論において重要な役割を果たす特定のインダクションヘッドに着目し (2 章), このインダクションヘッドの部分空間上で内部表現の親和性度および多様性度を定義する (3.1 節). そして, これらのデモの性質が要件 1,2 を満たすかを確認する.

実験の結果, 要件 1 について, Llama 3 8B において親和性度はさまざまなデータセットおよびデモ数を通してタスク性能と相関を示した一方, 多様性度についてはモデルやデータセットによって異なる傾向がみられた (図 2). これにより, 親和性は要件 1 を満たし, デモ選択において望ましい性質であることが確認された. 次に要件 2 について, 既存のデモ選択

手法を用いた実験を行った結果, Falcon 7B では親和性度および多様性度が Accuracy が高くなりやすい範囲に収束する傾向が見られ (図 1), 既存手法が親和性度および多様性度を収束させるようなデモ選択を行っていることが確認された. 一方, Llama 3 8B では既存手法で親和性度および多様性度を収束する傾向は見られなかった.

本研究の貢献は次の通りである:

- 既存手法に対する良いデモの性質として, 「デモとクエリの親和性」と「デモ同士の多様性」を提案し, その中で親和性が特にタスク性能に寄与するデモ選択における望ましい性質であることを示した.
- Falcon 7B においては, 既存のデモ選択基準が, タスク性能を向上させるような性質を持つデモを選んでいる可能性があることを確認した.

# 2 インダクションヘッドにおける内部表現の抽出

## 2.1 ICL の定式化

分類問題における ICL では, 入力文と出力ラベルのペアからなる自然言語で書かれたデータセットを用いる. データセットからデモ用の $k$ 個の入力文と出力ラベルを取り出し, クエリとなる入力文を用いてトークン列 $s$ を作成する ($k = 2$ の場合の例: 「"Good movies": Positive, "That's too cruel.": Negative, "I like it.":」). ここで, 「:」などの入力文と出力ラベルの間に用いる区切り文字を予兆トークン (Forerunner token) と呼ぶこととし, 本稿では, クエリの出力ラベルと一致しているデモのサンプルを, 「正しいサンプル」と呼ぶこととする.

## 2.2 インダクション回路

インダクション回路は大規模言語モデルなど複雑なモデルを分解し解析するために提案された手法であり, 大規模言語モデルにおける ICL を可能にする重要な回路である [3]. インダクション回路は異なる層の 1 組の注意ヘッドで構成され, 前のトークン情報を各トークンにコピーを行う注意ヘッドと, コンテキストに基づいてトークンへの注意を張る注意ヘッドがある. 特に後者をインダクションヘッドと呼び, [A][B]...[A'] を入力としたときに [B] を予測する確率を高めることができる.

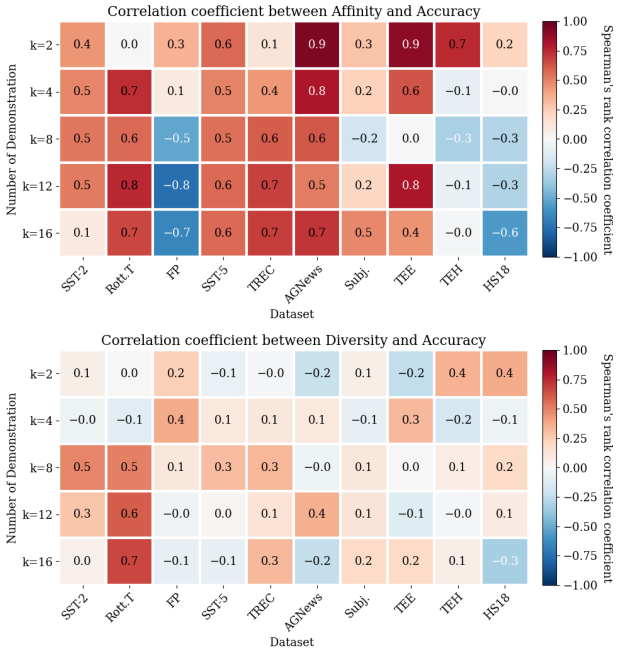


**図 2** LLama 3 8B: ランダムサンプリングしたデモに対する親和性および多様性とそれぞれの Accuracy とのスピアマン相関係数 (上: 親和性, 下: 多様性).

## 2.3 内部表現の抽出

前節で導入した ICL で重要な役割を担うインダクションヘッドのうち, 正しいサンプルのラベルトークンへの注意スコアが最も高いインダクションヘッド (「最も正しいインダクションヘッド」) を検出する. 先行研究に従い [3], 次の手順で最も正しいインダクションヘッドを検出する: (1) すべての正しいサンプルのラベルトークンからの注意スコアの合計が基準値 ($5 \times k /$ 出力ラベルの種類数 $/ s$ の長さ) を超え, 最も注意スコアの合計が高いヘッドを「最も正しいインダクションヘッド」として検出する. (3) 「最も正しいインダクションヘッド」において, 前の層の隠れ状態並びにアテンションのクエリおよびキーの重み行列の積からデモのラベルトークン表現 $\left\{d_{\text{label}}^{(i)}\right\}_{i=1}^{k}$ とクエリの予兆トークン表現 $d_{\text{query}}$ として抜き出す.

# 3 デモ選択基準として多様性と親和性は望ましい性質か

## 3.1 親和性および多様性の定義

実際に要件を満たすようなデモの性質を計算するためには, タスク性能により影響を与える表現を用いる必要がある. 本研究では, より ICL におけるタスク性能に直結する表現として, 別のモデルを用いて

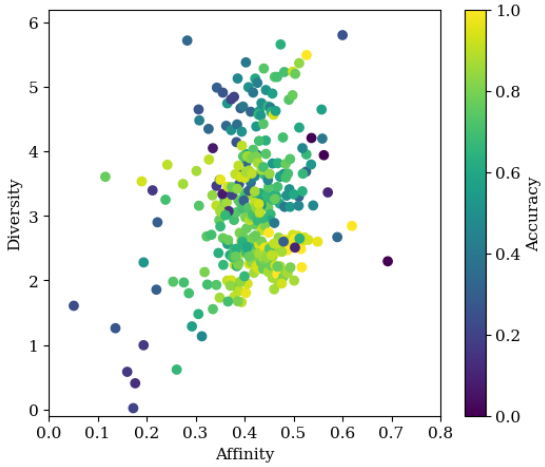


**図 3** Llama 3 8B: 全てのデータセットデモ数 16 個での親和性および多様性. 色が Accuracy を表しており, Accuracy が高くなりやすい度合いが存在する.

求める埋め込み表現ではなく, ICL を行う言語モデル内のインダクションヘッドの内部表現に着目し, 新たなデモ性質を提案する.

#### 3.1.1 多様性

ICL の推論に重要な役割を担う Induction Head の内部状態において, 多様性をデモの全てのラベルトークン表現間の共分散として, 式 1 のように定義する. ここで, D は $d^{(i)}_{\text{label}}$ の共分散行列とする.

$$\text{Div}\left(\left\{d^{(i)}_{\text{label}}\right\}_{i=1}^{k}\right) = \frac{1}{k}\text{tr}\left(\underset{i\in[1,k]}{\text{D}}\left[d^{(i)}_{\text{label}}\right]\right) \tag{1}$$

#### 3.1.2 親和性

多様性と同じく Induction Head の内部状態において, 親和性をデモの全てのラベルトークン表現とクエリの予兆トークン表現間のコサイン類似度として, 式 2 のように定義する.

$$\text{Aff}\left(d_{\text{query}}, \left\{d^{(i)}_{\text{label}}\right\}_{i=1}^{k}\right) = \frac{1}{k}\sum_{i=1}^{k}\cos\left(d_{\text{query}}, d^{(i)}_{\text{label}}\right) \tag{2}$$

### 3.2 実験設定

**モデル.** Llama 3 8B, Falcon 7B を用いて実験を行った. モデルのパラメータは `HuggingFace` から取得した.

**データセット.** 全ての実験で 10 個の文分類タスクのデータセットを用いた. データセットの詳細は付録 B を参照されたい. デモ数はそれぞれ $k = 2, 4, 8, 12, 16$ とし, データセットからランダムに $k$ 個選択を行い入力するトークン列を構成した.

### 3.3 実験結果

Llama 3 8B における親和性および多様性について, Accuracy とのスピアマン順位相関係数をそれぞれ図 2 に示す. 親和性度においては多くのデータセットおよび $k$ において相関が確認できた (図 2 上). 一方, 多様性度はデータセットにより異なる傾向が見られ, 多くのデータセットにおいて相関がないことが確認された. また, Falcon 7B においても親和性および多様性において同様の傾向が見られた (図 5). よって, 親和性は要件 1 を満たしデモ選択において望ましい性質であるが, 多様性は満たさないことが示された.

$k = 16$ において, Llama 3 8B における親和性と多様性をそれぞれプロットした図を示す (図 3). 親和性および多様性は右下の範囲で Accuracy が高くなる傾向が見られ, タスク性能を向上させる方向が存在する. また, Falcon 7B においても, 親和性および多様性は右下の範囲で Accuracy が高くなる傾向が見られ (図 7), このことから, 2 つのモデルでタスク性能を向上させる方向の存在を確認できた.

## 4 既存手法は親和性と多様性を集約させるか

本章では既存手法で選択したデモに対して, 前章で示したタスク性能を向上させる度合いに収束するか検証する. デモ選択以外の実験設定は, 3 章に従う.

### 4.1 既存手法

本研究では, デモ検索機の学習が不要なスコアリング手法 [4] を採用し, クエリを検索キーとし, デモの選択を行う.

- TopK-BM25: BM25 は TF-IDF を拡張した古典的な検索手法 [5] であり, 上位 $k$ 個のサンプルがデモとして選択される.
- TopK-BGE M3: BGE M3[6] を用いて生成した文埋め込みに対して k-NN を使用し上位 $k$ 個のデモを選択する.

### 4.2 実験結果

Falcon 7B におけるランダムサンプリングおよび BGE M3 で選択したデモに対し, $k = 16$ における親和性および多様性をプロットしたものをそれぞれ図 1 に示す. ランダムサンプリングしたデモと比較して既存手法で選択したデモは, 前章で確認されて

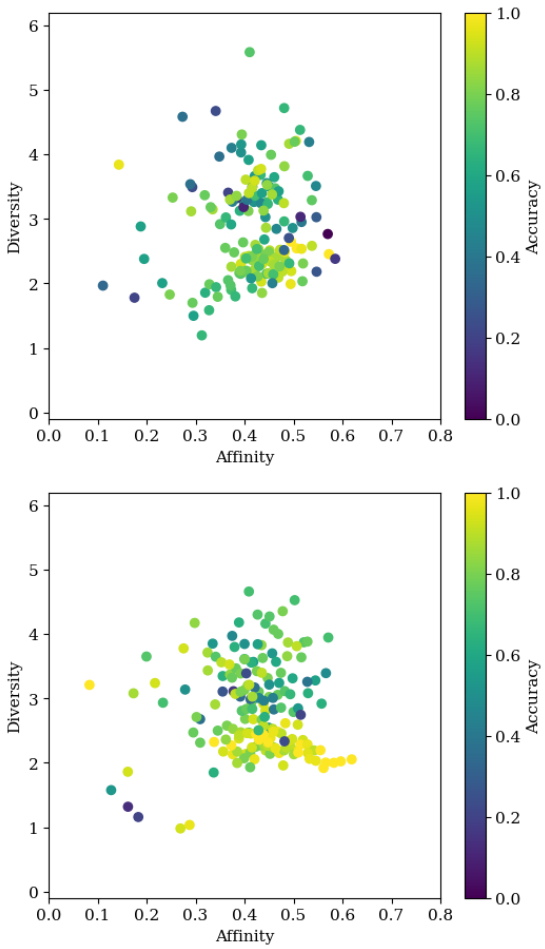


**図 4**　LLama 3 8B: 全てのデータセットにおけるデモ数 16 個での親和性および多様性 (上: BM25, 下: BGE M3). 既存手法によるタスク性能を向上する度合いへの収束は確認できない.

いたタスク性能を向上させる方向 (右下) に集まるような傾向が確認できる. また, この傾向は BM25 においても確認できる (図 6). 一方, Llama 3 8B においてはこの傾向は確認できない (図 4).

## 5　まとめ・今後の展望

デモ選択基準における良いデモの性質として, インダクションヘッドに基づく多様性と親和性を提案し, 特に親和性については, 複数のモデルとデータセットで望ましい性質であることを示した. さらに, Falcon 7B において, 既存のデモ選択手法（BM25 や BGEM3 を用いた手法）で選択されたデモが, これらの性質をタスク性能を向上させる方向へ集約させることを示した. 本稿では学習の必要がないデモ検索機を検証対象としたが, 学習を行うデモ検索機については今後の課題である. また, モデルやデータセットにより一部異なる結果が得られたことから, その原因等の詳しい調査も, 今後の課題である.

## 謝辞

本研究は中島記念国際交流財団の助成を受けたものです.

## 参考文献

[1] Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey, 2024.

[2] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3?, 2021.

[3] Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. Revisiting in-context learning inference circuit in large language models, 2024.

[4] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning, 2023.

[5] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. **Found. Trends Inf. Retr.**, Vol. 3, No. 4, p. 333–389, April 2009.

[6] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[7] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[8] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In **Proceedings of the ACL**, 2005.

[9] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. **Journal of the Association for Information Science and Technology**, Vol. 65, , 2014.

[10] Xin Li and Dan Roth. Learning question classifiers. In **COLING 2002: The 19th International Conference on Computational Linguistics**, 2002.

[11] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. Toward semantics-based answer pinpointing. In **Proceedings of the First International Conference on Human Language Technology Research**, 2001.

[12] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. **Advances in neural information processing systems**, Vol. 28, , 2015.

[13] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. **arXiv preprint arXiv:1803.05449**, 2018.

[14] Saif Mohammad, Felipe Bravo-Marquez, Mohammad

Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In **Proceedings of the 12th international workshop on semantic evaluation**, pp. 1–17, 2018.

[15] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In **Proceedings of the 13th International Workshop on Semantic Evaluation**, pp. 54–63, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics.

[16] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate Speech Dataset from a White Supremacy Forum. In **Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)**, pp. 11–20, Brussels, Belgium, October 2018. Association for Computational Linguistics.

# A　Falcon 7B の実験結果

Falcon 7B での実験結果を次に示す.

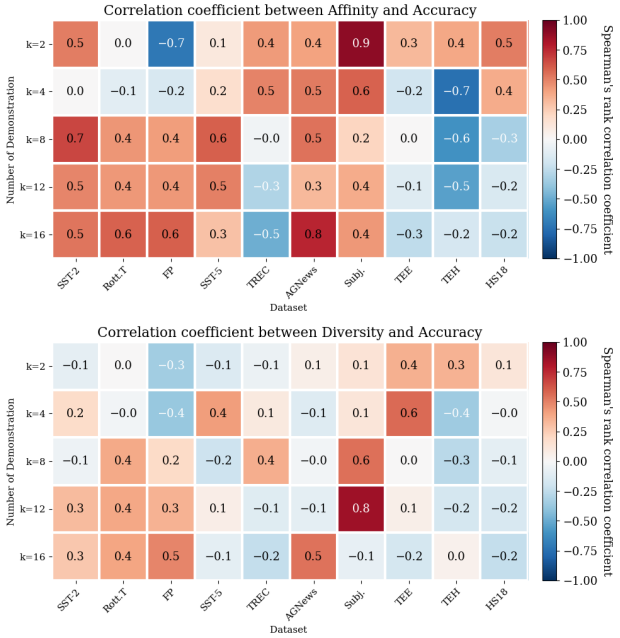


**図 5**　Falcon 7B: ランダムサンプリングされたデモに対する親和性および多様性とそれぞれの Accuracy とのスピアマン相関係数 (上: 親和性, 下: 多様性).

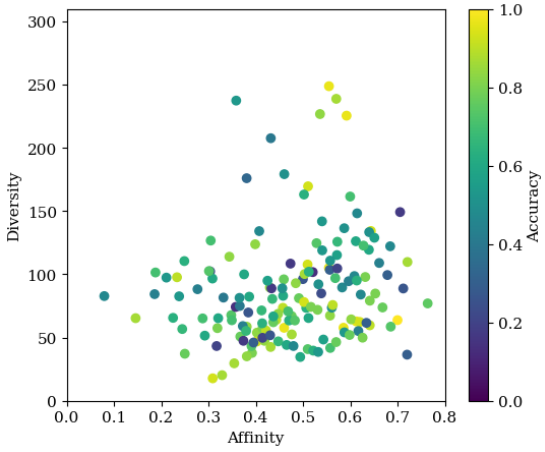


**図 6**　Falcon 7B: 全てのデータセットに対し BM25 で選択したデモに対する親和性および多様性 (デモ数 16 個). 色が Accuracy を表している.

# B　データセット

全ての実験で GLUE-SST2(SST-2)[7], Rotten tomatoes(Rott.T)[8], Finacial Phrasebank(FP)[9], Stanford Sentiment Treebank(SST5)[7], TREC(TREC) [10, 11], AGNews(AGNews)[12], Subjective(Subjective)[13], Tweet Eval Emotion(TEE)[14], Tweet Eval Hate(TEH)[15], Hate Speech 18(HS18)[16] を用いた.

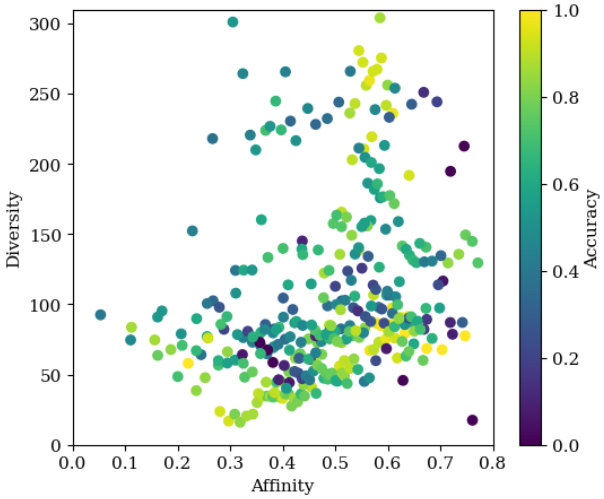


**図 7**　Falcon 7B: 全てのデータセットに対しランダムサンプリングで選択したデモに対する親和性および多様性 (デモ数 16 個). 色が Accuracy を表している.