LLM への軽量ベクトル介入による金融センチメント制御

平間 太規 [1]　伊藤 友貴 [2]　坂地 泰紀 [1]　野田 五十樹 [1]
[1] 北海道大学
[2] 国立研究開発法人情報通信研究機構
hirama.taiki.e1@elms.hokudai.ac.jp, tomoki.ito@nict.go.jp
{sakaji,i.noda}@ist.hokudai.ac.jp

概要

大規模言語モデル (LLM) は有用である一方，金融センチメント分析に使う場合，モデルの再学習（ドメイン適応）にかかる計算コストが運用時のボトルネックとなる．そこで，本研究では，重みを固定した LLM に対し，推論時に特定層の内部表現へ学習済みベクトルを加算し，3 値（POSITIVE/NEUTRAL/NEGATIVE）出力を制御する軽量手法 Polar2 を提案する．Polar2 では，末尾トークン表現から中立度軸と極性軸を学習し，介入強度で出力を調整する．更に，本手法の有効性を日本語及び英語のデータセットを用いて実証した．検証の結果，本手法により，比較手法に比べ低コストかつ高性能にドメイン適応できることを確認できた．

1　はじめに

近年，大規模言語モデル（LLM）は多様な自然言語処理タスクで高い性能を示している一方，実運用では汎用事前学習だけでは金融・経済の専門的文脈を十分に扱えない場合が多く，低コストにドメイン適応する方法が重要となっている．金融ドメインのセンチメントは，市場環境や政策，物価・為替などの状況変化により評価基準が変動し得る．例えば円安は，輸出企業の利益増が重視される局面では肯定的に捉えられうる一方，輸入価格高騰やインフレ圧力が問題化する局面では否定的に捉えられうる．このような「センチメントシフト」が起きる環境では，シフト毎にモデルを再学習する運用は計算・運用コストが大きい．このような問題を解決する手法として知識編集を用いたアプローチが近年，いくつか提案されている．知識編集では，モデル全体を学習するのではなく，ドメイン適応に有効なパラメータのみ，効率的に学習，あるいは，介入することで，効率的に LLM のドメイン適応を行う．その中でも，ポジティブコメントとネガティブコメントの対のような，対照例から得た差分を残差ストリームへ加算するステアリングベクトルを用いた手法である Contrastive Activation Addition（CAA）[1] は有望な手法の一つである．ただし,CAA は multiple-choice の 2 択（A/B）により「対象行動 vs その反対」の 二項対比 からステアリング方向を抽出する設計であり 3 値以上の分類問題にはそのまま適用できない．

そこで，本研究では，CAA で使われている考え方を 3 値以上の分類問題にも使えるように拡張した，金融センチメントのドメイン適応手法 Polar2 を提案すると共に，その有効性を実験的に評価する．Polar2 では，モデル重みを固定したまま，推論時に特定層の内部表現へ学習済みベクトルを微小加算することで，金融センチメント出力を制御する．このとき，少数パラメータ（中立度軸と極性軸の 2 本のベクトル）を外部に学習し，推論時に内部表現へ加算介入する．従来の CAA では極性軸のみ（1 本のベクトル）で学習をするが，本手法では，中立度軸と極性軸の 2 本のベクトルで学習する．これにより，CAA ではできなかった「多クラスセンチメント分類タスクに関する適応」が可能となる．

本研究の貢献は以下の通りである．

（1）少ない計算コストで，金融センチメントの多クラス分類タスクに関するドメイン適応を行える，ドメイン適応手法 Polar2 を提案した．

（2）実テキストデータを用いた検証により．提案手法によってドメイン適応が実際にできること，そして，提案手法が既存手法に比べ，低コストかつ高性能にドメイン適応できることを実証した．

2　関連研究

専門ドメインへの軽量適応に関する既存手法として，Low-Rank Adaptation (LoRA) [2] や知識編集（Knowledge Editing）を用いた手法が挙げられる．

LoRA では低ランク更新を導入した学習により，計算コストを抑えつつモデルのドメイン適応を行う．

LoRA は全パラメータ更新を行うフルファインチューニングに比べ，低コストでモデルのドメイン適応ができる一方，依然として計算コストが高い．

この問題意識と関係する枠組みとして，知識編集がある．知識編集はモデル全体の再学習に依存せず，特定の知識や振る舞いを意図通りに更新することを目指す．代表的手法として，勾配情報から更新を推定する MEND[3]，Transformer の MLP を介した局所更新で事実連想を書き換える ROME[4] や MEMIT[5]，また，これらの外部メモリ参照版である SERAC[6] 等が提案されている．このようなパラメータ更新による永続的編集に対し，推論時に内部表現へ介入して挙動を制御する方法も近年，いくつか提案されている [7, 8, 9]．特に，対照例から得た差分を残差ストリームへ加算するステアリングベクトルを用いたアプローチ Contrastive Activation Addition（CAA）[1] は有望なアプローチの一つである．ただし,CAA は 1 節で触れたように，3 値以上の分類問題にはそのまま適用できない点に課題がある．それに対し，本研究で提案する Polar2 は 3 値の分類問題にも利用可能である．

3　金融センチメント分析

本研究では，金融センチメント分析とは、景気コメント等の金融テキストに対し，ポジネガ分類を行うタスク [10, 11, 12, 13] であるとする．但し，分類を短いテキスト生成に基づいて行うことにする．即ち，金融テキストに対するポジネガを問うプロンプト（例: A.4) を入力に，LLM に `POSITIVE`, `NEUTRAL`, `NEGATIVE` のいずれかの文字列を生成させ，それを予測ラベルとする．このとき，生成結果から予測ラベルが抽出不能な場合は `UNKNOWN` として扱い，評価時は `NEUTRAL` にマップする．

4　提案手法: Polar 2

本節では，提案手法 Polar 2 を説明する．Polar 2 では，LLM の重みを固定したまま，推論時に LLM の特定層 1 層（以下，介入層 L とする）の内部表現へ少数パラメータのベクトルを加算（介入層 L の末尾トークンの隠れ状態へ介入）して生成出力を制御する（図 1）．まず，介入層 L の末尾トークン隠れ状態から中立度軸 u_r と極性軸 u_θ を学習する (4.1 節)．次に，推論時にそれらを加算して隠れ状態を $h \to h'$ と修正することで出力分布を調整する (4.2 節)．

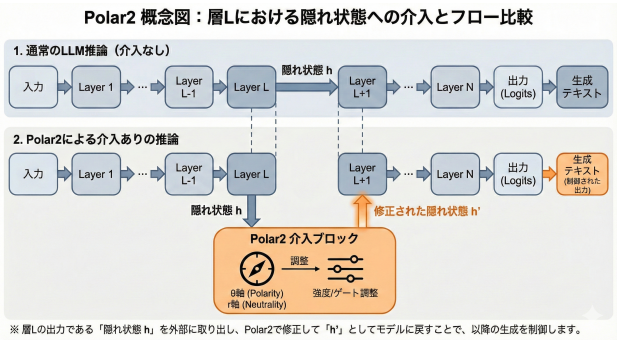

図 1　Polar2 概念図：通常推論（上）と，介入層 L の末尾トークン隠れ状態 h に介入して生成を制御する推論（下）．

4.1　2 軸（中立度軸・極性軸）の学習

図 1 のように，モデルの層数を N，介入層を L（$1 \le L \le N$）とする．本手法は生成による分類を前提とするため，末尾トークンに対応する層 L の隠れ状態 $h \in \mathbb{R}^d$ を特徴量として用いる．軸学習用データは JSONL 形式（`prompt`, `label`）で与える．

特徴抽出:　軸学習では，層 L における末尾トークンの隠れ状態を特徴量として用いる．具体的には，モデルの forward 計算で得られる隠れ状態列から，層 L の末尾トークン表現 $h \in \mathbb{R}^d$ を抽出し，これを特徴量ベクトルとする．

2 本の軸を用いた学習:　学習データから 2 つの線形軸を学習する：(i) 中立度軸 u_r（`NEUTRAL` vs. `POSITIVE/NEGATIVE`），(ii) 極性軸 u_θ（`POSITIVE` vs. `NEGATIVE`；`NEUTRAL` は除外）．後者では，極性（正負）判断を中立判定から独立に扱うため，`NEUTRAL` を学習対象から除外する．この 2 軸分解により，3 値分類を「中立か否か」と「正負どちらか」の 2 段として捉えられる．中立近傍では u_r に沿う調整を優先し，非中立と判断される例に対して u_θ により正負の分離を補正する．これにより，NEUTRAL の混同を抑えつつ，必要な方向にのみ介入できる．

SVM による軸推定と境界の保持:　各軸は，層 L の特徴ベクトルを入力として線形 SVM を学習し，得られた重み w を正規化して単位ベクトル u として定義する．同時に，判別境界もスケールに不変な形で保持するため，以下の u と γ を用いる（b は切片）．

$$u := \frac{w}{\|w\|}, \qquad \gamma := -\frac{b}{\|w\|} \tag{1}$$

パラメータの保存:　学習結果として，軸ベクトル u_r, u_θ と各境界 γ_r, γ_θ，および適用する層 L などの設定を保存する．学習パラメータは主に $u_r, u_\theta \in \mathbb{R}^d$

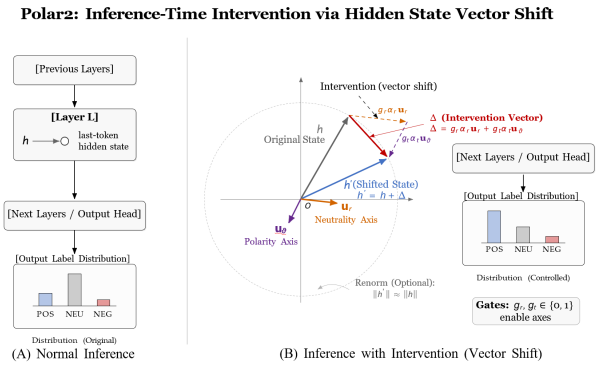

図 2　推論時介入による隠れ状態のシフト：(A) 介入なしでは層 L の末尾トークン隠れ状態 h が後段の生成へ伝播する．(B) Polar2 では $\Delta = \text{use}_r\,\alpha_r u_r + \text{use}_t\,\alpha_t u_\theta$ を加算して $h' = h + \Delta$ を得る（破線円は renorm のイメージ）．

（および少数の閾値）であり，総数は概ね $2d$ に相当する．ここで d は介入対象層の隠れ状態次元であり，本実験では $d = 4096$（$2d = 8{,}192$）である．

4.2　推論時介入（ベクトル加算）

推論時には，図 1 のように，介入層 L の内部表現に対して加算介入を行う．原則として**末尾トークン**（`Answer:` など）の隠れ状態 $h \in \mathbb{R}^d$ を操作対象とする．介入は 2 本の軸に沿った加算として定義し，以下で定義される h' を h の代わりに用いて推論する．

$$h' := h + \Delta, \quad \Delta := \text{use}_r\,\alpha_r u_r + \text{use}_t\,\alpha_t u_\theta \tag{2}$$

ここで $\alpha_r, \alpha_t \in \mathbb{R}$ はそれぞれ r 軸・θ 軸に沿った介入強度であり，式 (7) に従って決定する．$\text{use}_r, \text{use}_t \in \{0, 1\}$ はサンプル単位で各軸の介入を適用するか否か（式（5）及び式（6）にて定義）を表す．また，介入前後で $\|h\|$ を保つ正規化（`renorm`）を有効化し，表現のスケール変化が出力へ過度に影響することを抑制する．図 2 に示すように，通常推論では層 L の末尾トークン隠れ状態 h がそのまま後段へ伝播するのに対し，Polar2 では u_r, u_θ に沿う介入ベクトル Δ を加算して h を h' にシフトし，生成されるラベル語の分布を調整する．

スコア計算（適用判定と符号決定に使用）：
各サンプルについて，入力プロンプトに対し，各ラベル語（例：`Positive`, `Neutral`, `Negative`）を継続として与えた場合の teacher-forcing 対数確率を計算する．各ラベルの総対数確率 $lp_{pos}, lp_{neu}, lp_{neg}$ を

$$s_{pos} := lp_{pos} - lp_{neu}, \quad s_{neg} := lp_{neg} - lp_{neu} \tag{3}$$

として定義する．さらに

$$s_{diff} = |s_{pos} - s_{neg}|, \quad n_{dist} = \max(s_{pos}, s_{neg}) \tag{4}$$

を計算し，後述の適用判定や符号決定に用いる．

サンプル単位の適用判定（ヒューリスティック）：
本研究では，サンプルごとに「介入する/しない」を決めるため，s_{diff} と n_{dist} に基づくヒューリスティックな判定を用いる．具体的には，極性軸（θ 軸）は s_{diff}、中立度軸（r 軸）は n_{dist} が小さい場合にのみ適用し，

$$\text{use}_t := \mathbb{I}[\, s_{diff} \le \tau_t \,], \tag{5}$$

$$\text{use}_r := \mathbb{I}[\, n_{dist} \le \tau_r \,] \tag{6}$$

とする（$\mathbb{I}[\cdot] \in \{0, 1\}$ は指示関数）．use_t は **POS/NEG の判定に迷っている例**に極性軸の介入を限定する意図である．また，use_r を式 (2) のように使うことで，**中立近傍にある例**へ中立度軸の介入を限定する．

強度 α の決め方（境界依存／定数）：投影 $z = u^\top h$ と境界 γ を用いて介入強度を定める．強度を境界依存とする場合は，係数 k と定数 $a_{\max}$ を用いて

$$\alpha = \text{clip}(k(\gamma - z), [-a_{\max}, a_{\max}]) \tag{7}$$

とし，境界からの距離に比例して強度を変化させる．強度を定数とする場合は，α は定数とする．

θ 軸の符号決定: 極性軸（θ 軸）の符号は，強度が定数の場合は $|\alpha_t|$ を一定とし，ラベル尤度差 $\text{sign}(s_{pos} - s_{neg})$ により符号を与える．境界依存の場合は，$k(\gamma - z)$ の符号（すなわち z が境界のどちら側にあるか）により符号を決める．

5　実データを用いた有効性の評価

Polar2 の金融センチメント分類タスクにおける有効性を実データを用いて，以下の通りに評価した．

5.1　評価方法

まず，ベースモデルとなる LLM に対し，Polar2 により，訓練データを用いて，中立度軸 u_r と極性軸 u_θ を学習する (4.1 節)．その後，評価データの各入力に対し，Polar2 によるドメイン適応後の LLM を用いて，最大 3 トークンを生成する．生成結果から `POSITIVE/NEUTRAL/NEGATIVE` を抽出して予測ラベルとし，その結果が正しいか否かによって評価する．評価指標には Accuracy と Macro-F1 値を利用した．

5.2　データセット

本検証では，英語及び日本語の 3 値センチメント分類タスクに関するデータセットである Tweet-FinSent，及び Economy Watchers Survey (EWS) データセットを用いた．

TweetFinSent: TweetFinSent[14] は株式ティッカー

（例: $TSLA, $BABA）を含むツイートに POSITIVE/NEUTRAL/NEGATIVE のラベルを付与したデータセットである．本研究では公開データ取得後に前処理・整形を行い，訓練データには 132 件（POS=36, NEU=41, NEG=55），テストデータには 313 件（POS=111, NEU=110, NEG=92）を使用した．

EWS: 本データは，内閣府が毎月実施している景気ウォッチャー調査を収集・統合して構築された，現状判断・先行き判断のコメントとその 5 段階評価（◎, ○, □, ▲, ×）からなるデータセット [15] である．今回はこの 5 段階評価を 3 値（POSITIVE/NEUTRAL/NEGATIVE）に集約して用いた．訓練データには各ラベル 100 件ずつ（計 300 件），テストデータには各ラベル 200 件ずつ（計 600 件）を使用した．

5.3 比較手法

本検証では，提案手法である Polar2 の有効性を実証するため，以下の比較手法の性能と比較した．

Baseline: 本手法では、ベースモデルをそのまま用いたセンチメント分類を行う．

LoRA-single/LoRA-all: 本手法では、まず、LLM に対し、軽量適応の既存手法である LoRA を用いたドメイン適応を行う．その後、適応後の LLM を用いたセンチメント分類を行う．また，提案手法 Polar 2 との直接的な比較のため，一般的な LoRA の適用方法である，全ての層に対して更新を行う全層 LoRA を用いた結果 (LoRA-all) に加え，介入層 L のみに対して更新を行う 単一層 LoRA を用いた場合の結果（LoRA-single）とも比較した．

5.4 その他の設定

TweetFinSent におけるベースモデルは `meta-llama/Meta-Llama-3-8B` とし，介入層は Layer 16 に固定した．EWS におけるベースモデルは `pfnet/plamo-2-8b` とし，介入層は Layer 15 に固定した．Polar2 の軸・強度およびサンプル単位ゲートのベスト設定は A.1 に示す．また，単一層 LoRA 及び全層 LoRA におけるパラメータ r はぞれぞれ r=4/8 及び r=4 とした（詳細は A.3 に記載）．

5.5 結果・考察

各手法の評価結果は表 1 及び表 2 の通りである．また，参考として，各手法における更新パラメータ数も記載する．これらの結果より，提案手法である Polar2 を用いた手法が，比較手法である Baseline や LoRA-single/LoRA-all に比べ，高い Accuracy・Macro-F1 値を出せていることがわかる．また，Baseline より高いスコアであることより，Polar2 による適応がうまく動いていることもわかる．

表 1 評価結果: TweetFinSent

手法	更新パラメータ数	Acc	Macro-F1
Baseline	0	0.4120	0.3179
LoRA-single (r=4)	32,768	0.4153	0.3272
LoRA-single (r=8)	65,536	0.4121	0.3208
LoRA-all (r=4)	1,703,936	0.4281	0.3636
Polar2（提案手法）	8,192	**0.5144**	**0.5109**

表 2 評価結果: EWS

手法	更新パラメータ数	Acc	Macro-F1
Baseline	0	0.578	0.5450
LoRA-single (r=4)	32,768	0.577	0.5424
LoRA-single (r=8)	65,536	0.578	0.5412
LoRA-all (r=4)	524,288	0.610	0.6077
Polar2 （提案手法）	8,192	**0.632**	**0.6272**

更に，LoRA の結果と比較してみると，Polar2 では LoRA に比べ，更新するパラメータ数（すなわち，計算コスト）を抑えつつも，高い正解率・F1 値を出すことに成功していることがわかる．これより，大規模な追加学習を伴わないドメイン適応手法として Polar2 が有効であることが示唆される．

6 結論

本研究では，金融センチメント分析において有効な，LLM の低コストなドメイン適応手法 Polar2 を提案した．Polar2 では，LLM 重みを固定したまま層末尾の内部表現へ 2 軸ベクトルを微小加算する軽量介入を行う．更に，Polar2 の有効性を二つのデータセットを用いて評価した結果，Polar2 によるドメイン適応が実際にできること，そして，Polar2 が既存手法に比べ，低コストかつ高性能にドメイン適応できることを実証できた．併せて，提案手法 Polar2 を用いることで，少ないパラメータ数の更新でも，既存手法に比べ， 高い正解率、及び Macro F1 値を出せることも実証した．

今後の展開として，センチメント分類以外のタスクや言語を跨いだドメイン適応における有効性の検証に加え，介入層・介入位置の一般化（末尾トークンに限らず，全トークン／特定スパンへの介入や層 L の自動選択）により，性能と計算コストのトレードオフの体系化が挙げられる．また，頑健性・安全性の観点から，分布外データに対する安定性や，介入が安全性に与える影響の評価も重要である．

参考文献

[1] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pages 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[3] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale, 2022.

[4] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, volume 35, pages 17359–17372. Curran Associates, Inc., 2022.

[5] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In **The Eleventh International Conference on Learning Representations**, 2023.

[6] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale, 2022.

[7] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In **International Conference on Learning Representations**, 2020.

[8] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. GeDi: Generative discriminator guided sequence generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Findings of the Association for Computational Linguistics: EMNLP 2021**, pages 4929–4952, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[9] Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pages 6691–6706, Online, August 2021. Association for Computational Linguistics.

[10] Tsubouchi-K. Sakaji H. Yamashita T. Ito, T. and K. Izumi. Csnn: Contextual sentiment neural network. In **IEEE ICDM 2019**, 2019.

[11] Tsubouchi-K. Sakaji H. Yamashita T. Ito, T. and K. Izumi. Word-level contextual sentiment analysis with interpretability. In **AAAI 2020**, 2020.

[12] T. Ito, H. Sakaji, K. Tsubouchi, K. Izumi, and T. Yamashita. Text-visualizing neural network model: Understanding online financial textual data. In **PAKDD 2018**, 2018.

[13] Tsubouchi-K. Sakaji H. Yamashita T. Ito, T. and K. Izumi. Sentiment shift neural network. In **SDM 2020**, 2020.

[14] Yulong Pei, Amarachi Mbakwe, Akshat Gupta, Salwa Alamir, Hanxuan Lin, Xiaomo Liu, and Sameena Shah. TweetFinSent: A dataset of stock sentiments on Twitter. In Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen, editors, **Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)**, pages 37–47, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[15] Masahiro Suzuki and Hiroki Sakaji. Economy watchers survey provides datasets and tasks for japanese financial domain. In **Companion Proceedings of the ACM on Web Conference 2025**, WWW '25, page 805–808, New York, NY, USA, 2025. Association for Computing Machinery.

A　Appendix：追加の実験設定と詳細結果

A.1　Polar2 のベスト設定

Polar2 のベスト設定（軸・強度およびサンプル単位ゲート）を表 3 に示す．

表 3　steering（Polar2）のベスト設定（追加情報）．`gate_sdiff` は τ_t，`gate_ndist_r` は τ_r に対応．

言語	軸・強度（α）	ゲート（サンプル単位）
英語	r 軸: boundary（$k = -1500$, $a_{\max} = 6.0$），θ 軸: boundary（$k = -750$, $a_{\max} = 7.0$）	$gate\_sdiff = 0.4$, $gate\_ndist\_r = 1.5$
日本語	r 軸: const（$a_{const} = 0.02$），θ 軸: boundary（$k = -70$, $a_{\max} = 1.6$）	$gate\_sdiff = 2.0$, $gate\_ndist\_r = 1.5$

A.2　詳細結果（クラス別 F1・trainable params）

本文では省略したクラス別 F1 を含む詳細結果を，英語は表 4，日本語は表 5 にまとめる．生成は `do_sample=False` とし，決定的デコードで評価した．

表 4　英語実験：詳細結果（生成ベース評価，クラス別 F1）

Method	trainable params	Acc	Macro-F1	POS-F1	NEU-F1	NEG-F1
Baseline	0	0.4120	0.3179	0.5160	0.4380	0.0000
Polar2 (steering)	8,192	0.5144	0.5109	0.5000	0.4457	0.5872
LoRA-single (r=4)	32,768	0.4153	0.3272	0.5464	0.2767	0.1584
LoRA-single (r=8)	65,536	0.4121	0.3208	0.5474	0.2564	0.1584
LoRA-all (r=4)	1,703,936	0.4281	0.3636	0.5418	0.4182	0.1308

表 5　日本語実験：詳細結果（生成ベース評価，クラス別 F1）

Method	trainable params	Acc	Macro-F1	POS-F1	NEU-F1	NEG-F1
Baseline	0	0.5780	0.5450	0.7990	0.5400	0.2960
Polar2 (steering)	8,192	0.6320	0.6272	0.7940	0.5080	0.5790
LoRA-single (r=4)	32,768	0.5770	0.5424	0.7960	0.5410	0.2900
LoRA-single (r=8)	65,536	0.5780	0.5412	0.8040	0.5440	0.2750
LoRA-all (r=4)	524,288	0.6100	0.6077	0.8040	0.5310	0.4880

A.3　LoRA 学習設定の詳細

英語

- 共通設定：epochs=8，lr=2×10^{-5}，`batch_size=1`，`grad_accum=8`，max_length=256
- LoRA-all（r=4）：`q_proj`，`v_proj`（全層）
- LoRA-single（r=4/8）：`o_proj`（層 16 のみ），trainable params=32,768/65,536

日本語

- 共通設定：lr=2×10^{-5}，batch_size=1，grad_accum=8，max_length=256
- LoRA-single（r=4）：`o_proj`（層 15 のみ），epochs=1，`alpha=1`，`dropout=0.50`（trainable params=32,768）
- LoRA-single（r=8）：`o_proj`（層 15 のみ），epochs=1，`alpha=1`，`dropout=0.50`（trainable params=65,536）
- LoRA-all（r=4）：`o_proj`（全 Attention 層），epochs=2，`alpha=2`，`dropout=0.30`（trainable params=524,288）

A.4　プロンプト例

Prompt Example (watcher_survey_jp)

次の文章を `POSITIVE`, `NEUTRAL`, または `NEGATIVE` として分類してください：２月から始まっている新型コロナウイルスの影響は依然として脅威である。全体的にそれがしみわたって、なかなか状況の変化が見られない。終息に向かっていれば良くなるのだろうが、現状ではそれは考えられない。回答：`Answer with one of: Positive, Neutral, Negative. Answer:`

Prompt Example (TweetFinSent)

```
Target ticker: GME. Classify stock sentiment (POSITIVE/NEUTRAL/NEGATIVE) for this tweet. Answer with exactly one token: POSITIVE, NEUTRAL, or NEGATIVE. Tweet: \$GME stunning move +\$23 \$66+ last Answer:
```