ミーム画像に対する自己スコアリング選別による推論根拠付き回答の検証

鼎 凌太朗　坂井 優介　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学
kanae.ryotaro.kr8@naist.ac.jp
{sakai.yusuke.sr9,kamigaito.h,taro}@is.naist.jp

概要

ミームにおけるヘイト検出には，過剰検知は表現を萎縮させ得る一方で，過小検知は被害の拡散を助長し得る，といった本質的なトレードオフがある．したがって，精度の向上だけでなく，意思決定過程を明示し，その推論の妥当性を検証することが必要不可欠である．本研究では，文脈依存的なヘイトを含意するミームのヘイト検出に向けて，視覚言語モデルの説明可能性と検出感度の双方の向上を試みる．具体的には，推論強化の枠組みである REPS の着想を取り入れ，生成モデルが出力した複数の根拠候補を，独立な選別モデルによって妥当性の観点から選別し，その結果得られたデータで学習するパイプラインを構築する．実験の結果，説明の妥当性は約 60% の勝率で改善し，再現率は 0.960 に到達した．これらの結果は，本手法が有害コンテンツの見逃しに対する強力なセーフガードになりうる．一方，モデルが生成した根拠を分析したところ，モデルは過剰検知の傾向が強いことも明らかとなった．

1　はじめに

ソーシャルメディア上のミームは，図 1 中の画像のように，テキストと視覚的要素の組み合わせであり，文脈依存性が高く，また急速に拡散しやすい [1]．一方，この性質により，人種，ジェンダー，宗教などに関する差別的含意を伝達するため，自動検出の必要性が高まっている [2, 3]．ミームにおけるヘイトは画像–テキスト間の相互作用から生じることが知られている [4]．しかし，マルチモダリティに基づく手法には，このような相互作用の扱いの難しさから依然として課題が残る [5]．また，特に視覚言語モデル（VLM）では，検出過程の説明生成において，自身の回答の正当化のために幻覚や事後的な根拠を生成し得るため [6, 7, 8]，意思決定過程の観測可能性が低い．

図 1　ヘイトを含むミーム画像の例と本研究における学習用データセットの作成手順，及び推論手順．

本研究では，Rationale Enhancement through Pairwise Selection（REPS）[9] を参考に，複数の VQA 候補を生成し，最も妥当性の高いものを学習データとして選択する．このデータを用いて学習することで，モデルの説明性能を向上を試みる．なお，本研究の選別は REPS での生成モデルが評価も兼ねる，という自己評価を忠実に再現するものではなく，外部の選別モデルによって候補を評価する REPS-inspired filtering として位置付ける．また比較対象として，生成された回答とその根拠を選好せずにそのまま学習したモデルと，回答とその根拠をランダムに一つ選別したモデルによりベースラインを作成する．

本研究の実験より，REPS の手法を取り入れた学習データ選別は，ベースラインと比較して生成される根拠（rationale）の妥当性を有意に向上させることが確認された．また分類性能においても，高い再現率を達成し，ヘイトの兆候に対する検出感度の向上が示された．一方で，感度の上昇に伴う過剰検知の傾向が観測されたことから，説明品質と識別精度の間にトレードオフが存在することが示唆された．

2 背景

2.1 Hateful Memes Dataset とマルチモーダル推論

Hateful Memes Dataset [10] は，ヘイト検出における，マルチモーダル推論能力を測定するために構築されたデータセットである．このデータセットの特筆すべき特徴は，画像またはテキストの単一モダリティでは無害に見えるものの，両者を組み合わせることで初めて差別的な意味合いが生じるサンプルを含んでいる点である．したがって，このデータセットに対しヘイト検出を行うには，VLM が視覚情報とテキスト情報の複雑な相互作用を捉え，背景知識や文化的文脈を正確に理解し判断する能力が必要不可欠となる．こうした背景から，モデルが「どの要素に基づき，どのような論理を持ってミーム画像を有害と判断したのか」という根拠を明示することは，モデルの信頼性および誤判定要因の分析において重要となる．

2.2 VLM の推論過程と REPS

VLM の推論過程における説明可能性の向上は，重要な研究課題である．近年では，Chain-of-Thought (CoT) プロンプティング [11] をはじめとして，回答とともにその導出過程を生成させる手法が一般的となっている．本研究は，QA 形式を利用して分類ラベルとその根拠を同時に学習させる枠組みを Hateful Memes Dataset [10] に応用するものである．

また，生成される説明の質を向上させるための基盤として，川端らが提案した Rationale Enhancement through Pairwise Selection（REPS）[9] が存在する．REPS は，モデルが生成した複数の推論候補（回答と根拠のペア）に対し，ペアワイズ比較を行うことで，より論理的整合性の高い根拠を選別することで，良質な学習用データを構築する手法である．本研究では，REPS の中核である「候補生成 → 選好に基づく選別 → 学習」という設計思想をマルチモーダル設定に取り入れる一方で，REPS の自己評価を再現するのではなく，外部の VLM を選別器として用い，スコアリングを通じた選好を行う REPS-inspired selection として実装する．この設定により作成された学習データの違いによる，最終的な説明品質と検出性能がどの程度変化するかを検証する．

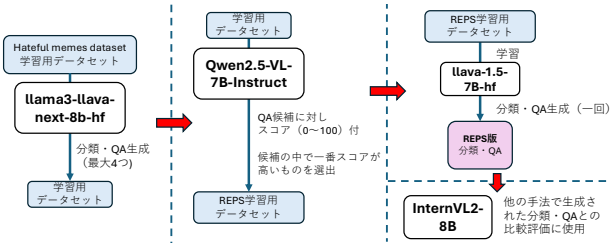

図 2 各工程で使用するモデルおよび手順

3 提案手法と実験設定

3.1 データセット

2 節で述べた Hateful Memes Dataset [10] を使用する．学習用データ（8,500 件）をモデルの学習に用い，検証用データ（500 件）を最終評価にのみ用いる．なお，テストデータは正解ラベルが非公開であるため，検証データにおける評価値を最終結果として報告する．

3.2 モデル構成

本実験では，図 2 で示されるように，役割の異なる以下の 4 つのモデルを使用する．複数のモデルを採用した主な理由として，単一モデルに全工程を担わせた場合に生じ得る系統的な偏りを低減し，評価の独立性を確保するためである．具体的には，(i) 生成モデル自身が候補の良否を判定すると，そのモデル固有の文体・推論形式への選好が「高品質」として固定化され，選別を通じて自己強化される可能性がある [9]．また (ii) 学習対象モデルと同一の評価器を用いると，学習対象の癖に沿った説明が高く評価されるなど，評価が自己循環し得る．これらの課題を踏まえて，本研究では工程ごとに異なるモデルを以下のように割り当てて実験を行う．

- QA 生成モデル (M_{gen})：Hateful memes dataset の学習用データを入力とし，判定結果とその根拠を含む QA ペアを生成を行う．モデルは `llama3-llava-next-8b-hf` [12] を使用する．
- QA 選別モデル (M_{sel})：M_{gen} により生成された QA 候補の中から，妥当性が高いものを選別す

る．本選別は REPS [9] から着想を得ているが，REPS とは異なり外部モデルを評価器に用いる．`Qwen2.5-VL-7B-Instruct` [13] を使用する．

- **学習対象モデル (M_{stu}):** 学習対象モデルには，`llava-1.5-7B-hf` [14] を学習対象として固定し，データ構築手法の差分のみが性能差として現れるようにした．
- 最終評価用モデル (M_{eval})：各手法で学習したモデルが生成した「説明の質」を，定性的に評価を行うモデルである．人手評価では，評価の再現性や評価者確保が難しいため，`InternVL2-8B` [15] を使用した自動評価を行う．

3.3 データ構築および比較手法

本研究では，REPS [9] の着想を取り入れた候補生成・選別に基づく学習データ構築（REPS-inspired selection）を検証するため，まず M_{gen} を用いて，学習用データ 8,500 件に対する QA 候補の生成を行った．一つの画像に対し図 1 の流れで，分類とその根拠の組み合わせを，QA 形式で生成する．このとき，1 画像あたり最大 4 つの QA 候補を確保することを目標に，最大で 10 回の生成を試行した．その結果，モデルによる生成が困難であった一部の画像（約 3%）を除く **8,245 枚**のユニーク画像に対し，合計 **29,886 件**の QA ペアが得られた．本実験ではこの候補プールを基に，以下の 3 つの基準で構築した学習データを使用する．

1. **Baseline (Random)**: 各画像に対して生成された QA 候補の中から，**ランダムに 1 つ**を選択して学習データとする（データ数:$N = 8,245$）．これは，選別を行わずに教師モデルの出力をそのまま蒸留した場合のベースラインに相当する．
2. **Baseline (Augmentation)**: 図 1 の 2 段目の工程で生成された候補プールを，**すべて**学習データとして利用する（データ数:$N = 29,886$）．
3. **REPS-inspired selection**: 図 1 の 3 段目のように，各画像の QA 候補に対し「回答の適切さ」「妥当性」などを基準にプロンプトによりスコアリングを行い，最もスコアが高いものを選別して学習データに利用する（データ数:$N = 8,245$）．

3.4 実装詳細

学習対象モデル（M_{stu}）は LoRA[16] を用いて学習を行う．学習データの違いによる性能への影響を評価するため，ハイパーパラメータは全て同一で学習を行う．詳細は付録 C に示す．

表 1 各手法の分類性能．提案手法は最も高い Recall と F1 Score を記録した．

Method	Accuracy	Precision	Recall	F1 Score
Baseline (Random)	0.526	0.514	0.928	0.662
Baseline (Augmentation)	0.518	0.511	0.872	0.644
REPS-inspired selection	0.524	0.513	0.960	0.669

3.5 評価指標

モデルの性能を，定量的な分類性能と，定性的な説明能力の二つの観点から評価を行う．

分類性能 (Classification Metrics) 学習したモデルが生成した回答テキストから判定ラベル（Hateful/Not Hateful）を抽出し，Accuracy，Precision，Recall，F1 Score を算出する．

説明の妥当性 回答の根拠の比較を行う．本研究では，M_{eval}（`InternVL2-8B`）を最終評価用モデルに用いて自動評価をして，比較を行なった．具体的には，検証用データ（500 件）に対し，異なる学習データで学習した 3 つのモデルが出力した根拠を，最終評価モデル M_{eval} にペアワイズで比較する．そして提案手法を利用して学習を行なったモデルにより生成された根拠が勝利した割合である Win Rate を算出する．

4 結果と考察

4.1 定量評価: 分類性能

表 1 に，各手法における分類性能を示す．Accuracy および Precision については，手法間で大きな差は見られなかった．川端らの先行研究 [9] でも，説明品質の改善を主眼とする枠組みでは必ずしも分類精度を大きく向上させるものではないことが報告されており，本結果はその傾向と整合している．

一方で，提案手法では Recall が 0.960 まで向上した反面，False Positive の増加が確認された．混同行列の内訳を見ると，検証データ内の Hateful 250 件のうち，Not Hateful と誤判定した件数は REPS-inspired selection が 10 件，Augmentation が 32 件，Random が 20 件であった．このことは，提案手法で学習したモデルが Hateful を取りこぼしにくい判定傾向へシフトした一方で，過剰検知とのトレードオフが生じたことを示唆している．

Method	Setting	n	Word count		Hedge count	
			mean±std	median[Q1,Q3]	mean±std	median[Q1,Q3]
Baseline	augmentation	500	147.59 ± 24.99	145 [130, 160]	2.848 ± 1.983	3 [1, 4]
Baseline	random	500	138.15 ± 23.41	138 [121, 153.25]	2.532 ± 1.742	2 [1, 4]
REPS-inspired selection	reps	500	138.68 ± 23.33	137.5 [122, 154]	2.066 ± 1.605	2 [1, 3]

表 2　生成された根拠（rationale）のスタイル統計量．平均 ± 標準偏差および，四分位数 $[Q1, Q3]$ を伴う中央値を報告する．

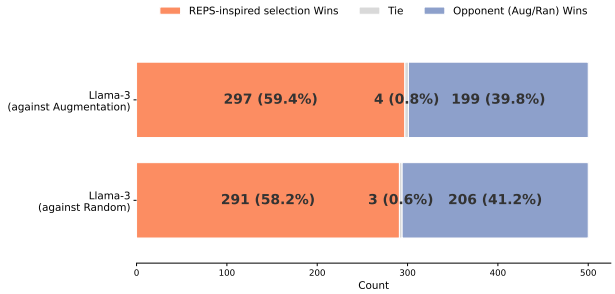

図 3　M_{eval} によるペアワイズ評価の結果（Win Rate）．オレンジ色は REPS が勝利した割合，青色は比較対象が勝利した割合を示す．

図 4　過剰検知が発生した画像 (ID:14865)

4.2　定性評価: 説明の妥当性

次に，生成された回答根拠の質を評価するため，最終評価用モデル（M_{eval}）によるペアワイズ評価を行う．その結果は図 3 に示されるように，提案手法は二つの Baseline に対してそれぞれ 58.2%/59.4%の割合でより妥当と判定された．この結果から，少なくとも本研究のモデル構成において，候補生成 → 選別 → 学習という選別プロセスが説明品質の改善に寄与し得ることを示唆する．

4.3　考察

Recall 向上と過剰検知増加の要因分析　生成された根拠の表層的性質として，(i) 根拠の語数と (ii) ヘッジ語の使用回数の点で分析する．本研究では，ヘッジ語を不確実性・推量を表す語句（*may, might, could, possibly, probably, potential, likely, unlikely, seem, appear, suggest, can be*）に加え，解釈の余地を示す語（*interpretation, context, intent*）として定義し，その出現回数をヘッジ数として数える．その分析結果が表 2 であり，各設定における平均 ± 標準偏差および中央値 $[Q1, Q3]$ を示す．根拠の長さは，提案手法は Baseline（random）と同水準であり，また Baseline（augmentation）と比べるとわずかに短い．一方，ヘッジ語の使用に関しては提案手法が二つの Baselin と比べて，使用回数が比較的少なかった．このことから，提案手法を利用して生成された回答の根拠は，文章の長さを増やすことなく，より断定的な言い回しになっていることが示唆される．

過剰検知の具体例　過剰検知の具体例として，画像 4 を挙げる．本例に対し，Baseline（Random）は無害であると判断できた一方，提案手法では，歴史的文脈に関する倫理的リスクを強調する根拠を生成し，結果として誤判定を起こした．この事例は，無害なものであっても，センシティブな情報（特定の歴史・人物など）に重み付けが生じることで，判定が Hateful 側へシフトする傾向を示唆している．

5　おわりに

本研究では，VLM の説明品質と検出感度の向上を目的に，候補生成 → 妥当性に基づく選別 → 学習からなる学習データ構築パイプラインを構築し，Hateful Memes Dataset を対象に，ミーム画像のヘイト検出を行った．

実験の結果，提案手法によって学習されたモデルから生成された根拠は，両ベースラインに対して説明の妥当性が約 6 割の勝率で選好され，説明品質の改善が観測された．また分類性能の分析では，再現率が 0.960 に到達する一方で，False Positive の増加も伴っており，説明品質の改善および高再現率化は，過剰検知とのトレードオフを生じ得ることが示唆された．今後の課題として，過剰検知を抑制するための選別基準の改良が挙げられる．加えて本稿の結果は当該モデル構成に依存し得るため，選別モデルや評価器の変更に対する頑健性の検証も行いたい．

参考文献

[1] Limor Shifman. Memes in a digital world: Reconciling with a conceptual troublemaker. **Journal of Computer-Mediated Communication**, Vol. 18, No. 3, pp. 362–377, 04 2013.

[2] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In Lun-Wei Ku and Cheng-Te Li, editors, **Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media**, pp. 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics.

[3] Delfina Sol Martinez Pandiani, Erik Tjong Kim Sang, and Davide Ceolin. Toxic memes: A survey of computational perspectives on the detection and explanation of meme toxicities, 2024.

[4] Gokul Karthik Kumar and Karthik Nandakumar. Hateclipper: Multimodal hateful meme classification based on cross-modal interaction of clip features, 2022.

[5] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications, 2019.

[6] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Transactions on Information Systems**, Vol. 43, No. 2, p. 1–55, January 2025.

[7] Zujing Liu, Junwen Pan, Qi She, Yuan Gao, and Guisong Xia. On the faithfulness of visual thinking: Measurement and enhancement, 2025.

[8] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics.

[9] Akira Kawabata and Saku Sugawara. Rationale-aware answer verification by pairwise self-evaluation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 16178–16196, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[10] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes, 2021.

[11] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[12] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild, May 2024.

[13] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024.

[15] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. **arXiv preprint arXiv:2412.05271**, 2024.

[16] 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.

[17] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

A　プロンプトテンプレート

本研究のパイプラインで用いたプロンプトテンプレートを示す．具体的には，(i) 判定ラベルと最小限の根拠を厳密な JSON 形式で出力させる構造化 QA，(ii) 判断根拠（rationale）を生成する説明生成，(iii) 画像に基づく事実性と論理的一貫性から 0–100 点で評価する採点，の 3 種類である．いずれも，利用可能な場合は OCR 抽出テキストを追加コンテキストとして付与する．これらは再現性の確保と，要求する出力の明確化のために提示する．

A.1　構造化 QA プロンプト (JSON)

```
You are a careful multimodal QA assistant. Output JSON
    only.

1) Q1: Does the meme express hate toward a protected
    group?
   - Answer with one of: "Yes", "No", or "Uncertain".
2) Q2: If yes or uncertain, which group appears
    targeted?
   - Short phrase, or "None".
3) Evidence: Quote minimal text or describe minimal
    visual elements (<=2 sentences).

Return STRICT JSON with keys:
    {"q1":"...","q2_group":"...","evidence":"..."}
Do not add any text outside JSON.
[OCR/text in meme]: "<OCR_TEXT>"
```

A.2　説明生成プロンプト

```
Provide a concise, fact-grounded rationale in 2-4
    sentences about whether the meme is hateful and
    why.
[OCR/text in meme]: "<OCR_TEXT>"
```

A.3　スコアリングプロンプト

```
You are a helpful assistant who evaluates
    explanations' factual accuracy and logical
    consistency.

Look at the image and the provided information.
    Evaluate the following explanation and assign a
    score from 0 to 100
based on how factually grounded it is in the image and
    how logically valid it is.

Scoring rubric:
- 90-100: Fully grounded in the image/information,
    logically coherent, no unsupported claims.
- 70-89: Mostly grounded; minor speculation or small
    logical gaps.
- 40-69: Partially grounded; several unsupported
    claims or notable inconsistencies.
- 10-39: Largely ungrounded; major hallucinations or
    faulty reasoning.
- 0-9: Completely ungrounded or contradicts the
    image/information.

Output format:
Score: <integer 0-100>
Rationale: <1-2 sentences explaining the main evidence
    and the main issue, if any>
```

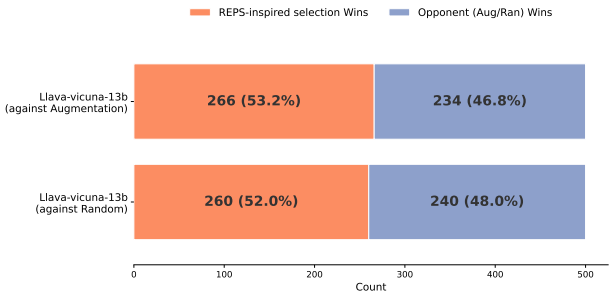

図 5　LLaVA-v1.6-Vicuna-13B におけるペアワイズ勝率．REPS-inspired selection は，Baseline（random）に対して 266/500（53.2%），Baseline（augmentation）に対して 260/500（52.0%）勝利した．

Hyperparameter	Value
Per-device train batch size	16
Gradient accumulation steps	2
Effective batch size	32
Epochs	3
Learning rate	2×10^{-4}
Precision	bfloat16
Gradient checkpointing	enabled
Logging steps	10
Checkpoint saving	every epoch
Dataloader workers	4

表 3　使用した学習用ハイパーパラメータの詳細．

B　LLava を用いた win rate の算出

さらに，LLaVA-v1.6-Vicuna-13B[17] で作成した学習データを用いた場合の最終的なペアワイズの選好結果を報告する．選好前の学習データのデータ数は 29886 件で，選好後は 8255 件となった．図 5 に示すとおり，REPS-inspired selection の勝率はチャンスレベルをわずかに上回った．Llama での結果と比べると差分は小さく，選好の改善幅は学習データの作成に利用したモデルに依存し得ることを示唆する．

C　Training Hyperparameters

公平な比較を担保するため，すべてのモデルを同一条件で学習した．表 3 に，共通して用いた設定をまとめる．