大規模視覚言語モデルにおける言語タスクに対する視覚情報の影響

吉田 大城 [1]　林 和樹 [1]　坂井 優介 [1]　上垣外 英剛 [1]　林 克彦 [2]　渡辺 太郎 [1]
[1] 奈良先端科学技術大学院大学　[2] 東京大学
{yoshida.daiki.ye6, hayashi.kazuki.hl4}@naist.ac.jp
{sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp
katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp

概要

大規模視覚言語モデル（LVLM）は言語情報に加えて視覚情報も扱うことができる言語モデルである。一般的にLVLMは、視覚と言語の両方を同時に扱うようなタスクで用いられることが前提とされているが、言語情報のみで解決可能なタスク（言語タスク）に限定して使用することも可能である。ただし、LVLMの構築過程では画像情報と整合が取られているため、視覚情報が追加的に与えられる状況において、その画像がタスクに密接に関連する場合だけでなく、逆に敵対的であったり無関係であったりする場合に応答がどのように変化するのかは明らかではない。本研究では、心理学的効果検証の一種であるプライミングを参考に、LVLMを用いて言語タスクを解く際に、視覚情報を追加的に挿入することで言語タスクへの影響を調査する。実験の結果、言語タスクにおいて視覚情報を追加することで、精度と確信度の両方に変化が観測され、LVLMでもプライミング効果を確認でき、LVLMが言語タスクを扱う際でも視覚情報の影響を受けることが明らかとなった。

1　はじめに

大規模視覚言語モデル（LVLM）[1, 2, 3, 4, 5]は、言語情報を扱うための大規模言語モデル（LLM）[6, 7, 8]と視覚情報を扱うためのビジョンエンコーダ[9, 10, 11]を統合することで構築される。視覚と言語の2つのモダリティを統合的に処理することで、視覚情報と言語情報の両方を同時に扱うような高度な推論を行うことができる。この統合は、視覚情報と言語情報を同時に用いるタスクを通じて実現され、LLMの推論過程において視覚情報を活用することを可能にしている。しかし、訓練プロセスでは常に関連性の高い画像と言語が提供されるため、言語タスクに追加的な視覚情報が提供される場面は存在しない。そのため、言語のみで解決可能なタスク（言語タスク）や視覚情報が不要な状況、言語情報が主導的役割を果たす問題解決において、視覚情報の寄与や相互作用がLVLMの応答にどの程度影響を与えるのか定かではない[12, 13, 14, 15]。

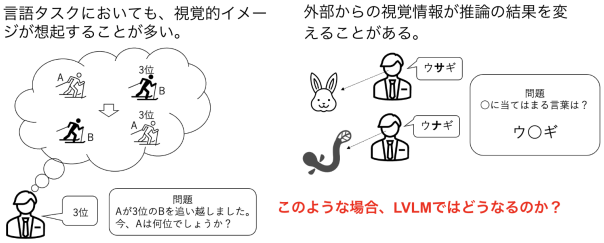

図1　視覚情報によるプライミングの例：図の左側は言語タスクにおいても視覚情報を補助的に用いている例。一方、右側は視覚情報によって推論が変化する例である。

この影響を理解することは、モデルの性能向上にとどまらず、モデルがどのように機能しているかを深く理解する上でも重要である。視覚と言語といった異なるモダリティが言語タスクの解決にどのように影響を及ぼすかを明確にすることで、モデルが情報を統合して意思決定を行う仕組みの解明に貢献できる。さらに、視覚と言語情報の相互作用は、人間の認知メカニズムと密接に関連しており[16]、人間の認知における視覚の役割を踏まえると、LVLMが視覚情報をどのように活用して言語的推論を行うかを明らかにすることで、人間の認知と類似したプロセスが見られるかを探ることができる。

本稿では、心理学的効果検証の一種であるプライミングを参考に、言語タスクにおいて視覚情報がLVLMの精度と確信度にどのような影響を与えるかを調査した。実験では、感情分類タスクである

表 1 EQ（上段）と SST2（下段）の例。Text 列の括弧を除く部分は、モデルに与えるプロンプトである

Task	Text	サポート画像	敵対画像	ノイズ画像
EQ	Question: What position does ***Roman Turek*** play? (Answer: goaltender)			
SST2	Sentence: Equals the original and in some ways even betters it. Is this sentence negative or positive? (Answer: positive)			

SST2 [17] と、特定のエンティティに関する質問応答タスクである Entity Questions（EQ）[18] の 2 つの言語タスクを対象とした。実験では、答えに導くこと意図したサポート画像、回答を惑わすことを意図した敵対画像、特定の意味を持たないノイズ画像の 3 種類を LVLM に与え、それに伴う精度と確信度の変化を分析した。実験の結果、言語タスクにおいても視覚情報を追加することで、精度と確信度の両方に変化が観測された。この結果から、LVLM でもプライミング効果が生じることがわかり、LVLM が言語タスクを扱う際でも視覚情報の影響を受けることが明らかとなった。

2 プライミング

プライミングとは、図 1 に示すような、先行刺激の受容が後続刺激の処理に**促進的**な効果をもたらす現象を指す [19, 20, 21, 22]。また、特定の条件下ではプライマーが**抑制的**な効果をもたらすことがあり、これはネガティブプライミングと呼ばれる [23, 24]。

本研究では、言語タスクにおける問い（テキスト）と共に、プライミング効果やネガティブプライミング効果の誘発が期待される画像を LVLM に追加で入力することで、プライミング効果が LVLM でも発生するか検証する。付録 D に EQ タスクにおける LVLM のプライミング効果の実施例を示す。

3 視覚情報の影響の調査方法

LVLM に対する言語タスクにおける視覚情報の影響を調べるため、Entity Questions [18] と SST2 [17] という性質の異なる 2 つの言語タスクを対象とする。

3.1 データセットの詳細

Entity Questions（EQ）は表 1 上段の Text に示すように、特定のエンティティに関する解答を導出する質問応答タスクである。質問文の形式はあらかじめ定められた複数の質問パターンに基づいて決定される [1)]。EQ では、テストデータとして 13,307 件の質問を用い、画像情報が事実抽出にどのような影響を与えるかを検証する。一方、SST2 は感情分類タスクであり、表 1 下段の Text に示すように、与えられた文章を positive または negative に分類する。SST2 では、1,821 件のテストデータを用いて、画像情報が感情判断にどのように影響するか調査する。各データセットの事例は付録 C に掲載している。

3.2 プライミング効果の検証用画像

言語タスクにおける視覚情報の影響を調べるため、LVLM にはテキストに加えて、以下の 3 種類の画像が与えられる：

- **サポート画像**: タスクに関連する視覚的なヒントや情報を含む画像で、プライミング効果により、正解へ導くことを促進する。
- **敵対画像**: 意図的に誤解を招くような視覚情報を含む画像で、ネガティブプライミング効果により、正しい答えを導き出すことを妨害する。
- **ノイズ画像**: ランダムに生成された画素で構成された画像で、プライミング効果の対照実験として、モデルが無関係な視覚情報にどの程度影響されるかを検証するために使用される。

各データセットにおける各画像の例を表 1 に示す。各画像の選定方法として、EQ では、各事例の回答となるエンティティを Wikpedia で検索し、該当記事に掲載されている画像をサポート画像として用いた。一方、敵対画像については、質問パターンごとに適切な敵対画像を 2 枚ずつ選定した。各質問パターンごとの敵対画像の一覧は付録 C の表 3 に記載している。一方 SST2 については、感情予測タスクであるため、絵文字をプライミング用画像として使用した。絵文字は Kaggle データセット [2)] から取得し、見た目が明らかに positive なものと negative

1) 詳細は Entity Quetions のリポジトリを参照。
2) 詳細は Full Emoji Image Dataset を参照

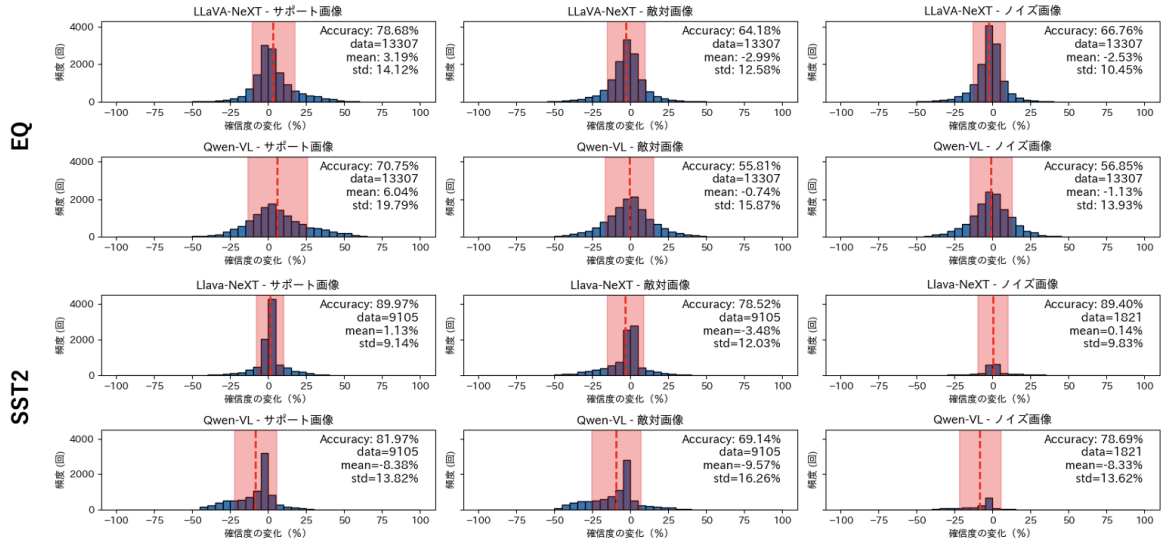

図 2　各 LVLM における各画像ごとの確信度の変化。テキストのみの条件と比較して、回答の選択肢に対する確信度の変化率を集計した。平均値は確信度の全体的な変化率を表す。確信度が最も高い選択肢を回答としたときの精度を示す。

なものをそれぞれ 5 種類ずつ手作業で選定した。選定画像の一覧は付録 C の表 4 に掲載している。

3.3　実験設定

実験に使用した LVLM　本研究では、Qwen-VL [3] [2]、LLaVA-NeXT [4] [1]、mPLUG-Owl2 [5] [4] の 3 つの LVLM を使用した。さらに、それぞれの LVLM の基盤となる LLM である Qwen-7B [6] [25]、Mistral-7B [7] [26]、Llama2-7B [8] [6] についても言語タスクの評価を行った。なお、LLM に関しては instruct-tuning されたモデルを使用している。

評価指標　評価方法として、各タスクのテキスト情報に加え、各画像を同時に入力し、LVLM から出力を得る。基本的に出力と正解ラベルの完全一致を行うが、モデルが生成した文には正解ラベル以外の情報も付与されている。そのため付録 A に記載したルールに従って、生成文に後処理を加えることで、回答候補を抽出している。

確信度　単にモデルが正答を出すか否かだけでなく、視覚情報がモデルの判断に与える影響の深さやその性質を捉えるために、前述の精度に加えて応答の確信度（の出るが特定の選択肢を正しいと判断する度合い）についても調査する。たとえば、画像によって確信度が大きく向上する場合、視覚情報が言語タスクの補完や強化に寄与している可能性が高い。一方、確信度が低下する場合、視覚情報が逆に判断を曖昧にしている可能性がある。本家級では、EQ では 4 つの選択肢、SST2 では positive と negative の 2 つの選択肢を用意し、各選択肢に対するモデルの確信度を計算し（詳細については付録 B を参照）、画像による確信度の変化を分析した。

3）　Qwen-VL: Qwen/Qwen-VL-Chat
4）　LLaVA-NeXT: llava-hf/llava-v1.6-mistral-7b-hf
5）　mPLUG-Owl2: MAGAer13/mplug-owl2-llama2-7b
6）　Qwen-7B: Qwen/Qwen-7B-Chat
7）　Mistral-7B: mistralai/Mistral-7B-Instruct-v0.2
8）　Llama2-7B: meta-llama/Llama-2-7b-chat-hf

表 2　画像別の精度：LLM は各 LVLM で使用されている LLM を示す。mPLUG-Owl2 については、テキストのみで推論を行うことができないため、欠損値としてある。また、LVLM の中でスコアが最大と最小のケースにはそれぞれ赤と青のセルで表す。

Task	Input	Qwen-VL	LLaVA-NeXT	mPLUG-Owl2
EQ	LLM	30.9	35.3	35.6
	テキストのみ	27.8	33.2	-
	サポート画像	34.5	38.4	32.9
	敵対画像	21.4	24.6	22.1
	ノイズ画像	25.4	28.4	26.6
SST2	LLM	74.5	65.1	83.5
	テキストのみ	80.5	81.4	-
	サポート画像	67.7	82.7	90.7
	敵対画像	63.8	79.4	81.6
	ノイズ画像	61.8	82.2	88.2

4　実験結果

画像の種類による、精度変化の実験結果を表 2 に示す。EQ と SST2 の両方において、サポート画像の場合はテキストのみの場合よりスコアの上昇、敵対画像の場合はスコアの低下が観察された。したがって、LVLM が言語タスクを扱う際でも、視覚情報の影響によるプライミング効果が現れることが示された。このことは、言語課題における LVLM のパ

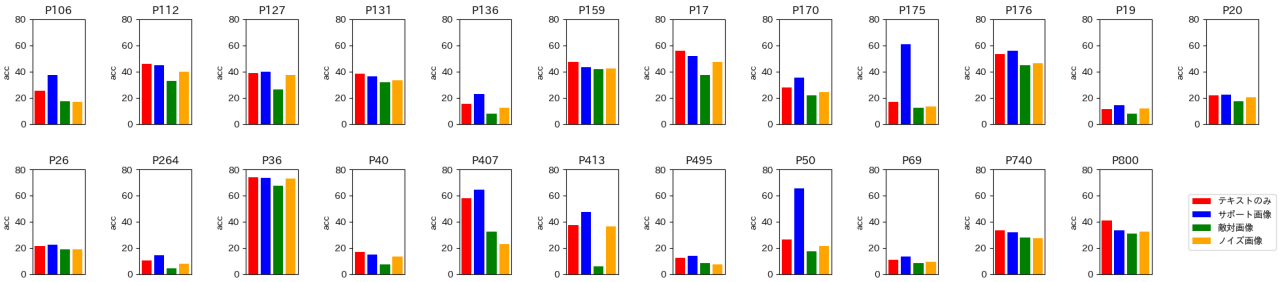

図 3　EQ の質問パターン別に与えられた画像の種類に基づく精度の変化を可視化。画像の種類による精度の変化はモデル間で同様の傾向を示したため、紙面の都合上 LLaVA-NeXT の結果のみを表示している。

フォーマンスにおいて、視覚情報が重要な役割を果たしていることを示している。

ベースの LLM と比較しても、一部を除きサポート画像を与えた場合の方がスコアが高いため、LVLM は人間と同じように、図 1 の左側部分にあるような、言語タスクにおいても視覚情報を補助的に用いることで推論の精度が向上していることが確認できる。また、ノイズ画像については、多くの場合で敵対画像ほどではないものの、テキストのみと比較してスコアが低下しており、無意味な視覚情報によって応答が惑わされていることが確認できた。

4.1　Entity Questions

どの LVLM においても、与えられた画像が、サポート画像の場合で精度が最も高く、敵対画像の場合で精度が最も低くなるという明確な傾向が見られた。これは、視覚情報が事実の抽出を伴う言語タスクに大きく影響することを示している。

図 3 は EQ で問われる質問パターンごとに対する結果を示しており、質問パターンによっては精度が 50%以上変動している。具体的に、サポート画像が与えられた場合に精度が著しく向上したのは、本や歌の作者について尋ねる問題（P50, 175）で、次いで人物の職業を尋ねる問題（P106）であった。これらの質問パターンで与えられる画像には、タスクの答えが直接的に含まれていることが多いという特徴があった。また、スポーツ選手のポジションを問う問題（P413）では、モデルが敵対的な画像に映し出された選手を問題の対象として誤って認識し、誤答になることが頻繁に観察された。これら結果は、視覚情報によって、モデルが答えを生成する際に周辺知識を利用できる可能性を示している。

4.2　SST2

SST2 においても、画像情報が精度に影響することが確認され、視覚情報が感情分類タスクにも影響することが示された。しかし、EQ の場合とは異なり、モデル間で一貫した傾向は観察されなかった。具体的には、LLaVa-NeXT や mPLUG-Owl2 では、提供された画像の意図に沿って精度が変化したが、Qwen-VL では、画像が与えられると種類に関係なく精度が低下した。EQ で明らかになったように、視覚情報がタスクの答えを直接含むかどうかが、精度の変化を決定することからも、このモデルにとって、絵文字画像によって表現される感情は、パフォーマンスに有意な影響を与えるには間接的すぎた可能性を示唆している。

4.3　プライミングによる確信度の変化

図 2 に確信度の結果を示す。EQ、SST2 の両方において、画像が与えられた際の確信度の変化は精度の変化と同じ傾向を示し、特に画像が確信度を向上させるケースでは精度も向上する傾向が顕著であった。これにより、視覚情報がモデルの判断過程に影響を及ぼすことがより強固に示された。

5　結論

本研究では、言語タスクにおいて、視覚情報が LVLM に与える影響を検討した。その結果、LVLM の応答が画像の意図する方向にシフトする傾向が見られ、LVLM におけるプライミング効果の存在が確認された。具体的には、視覚情報がタスクの解答と関連性の高い内容を含む場合、モデルは積極的に視覚情報を利用し、応答の精度や確信度に変化をもたらした。これらの結果から、LVLM は言語課題においても視覚情報の影響を受けると結論付ける。

本研究により、LVLM は言語タスクであっても視覚情報の影響を受けることが明らかになった。不正確な応答を引き起こす悪意のあるプライミングのリスクに対処するために、画像情報を柔軟に利用する手法の開発については今後の研究課題としたい。

謝辞

本研究は JSPS 科研費 JP23H03458 の助成を受けたものです.

参考文献

[1] 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.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[3] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. **arXiv preprint arXiv:2409.12191**, 2024.

[4] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 13040–13051, June 2024.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[6] Hugo Touvron, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[7] OpenAI. Gpt-4 technical report. **ArXiv**, Vol. abs/2303.08774, , 2023.

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **International Conference on Machine Learning**, 2021.

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In **International conference on machine learning**, pp. 19730–19742. PMLR, 2023.

[12] Feiqi Cao, Soyeon Caren Han, Siqu Long, Changwei Xu, and Josiah Poon. Understanding Attention for Vision-and-Language Tasks. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 3438–3453, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[13] Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. Real-World Robot Applications of Foundation Models: A Review, February 2024.

[14] Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. Towards artwork explanation in large-scale vision language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 705–729, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[15] Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. Towards cross-lingual explanation of artwork in large-scale vision language models. **ArXiv**, Vol. abs/2409.01584, , 2024.

[16] David Swinney, Edgar Zurif, Penny Prather, and Tracy Love. Neurological distribution of processing resources underlying language comprehension. **Journal of Cognitive Neuroscience**, Vol. 8, No. 2, pp. 174–184, 1996.

[17] 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 David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[18] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6138–6148, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[19] John A. Bargh and Tanya L. Chartrand. The mind in the middle: A practical guide to priming and automaticity research. In Heinrich T. Reis and Chris M. Judd, editors, **Handbook of research methods in social and personality psychology**, pp. 253–285. New York: Cambridge, 2000.

[20] Marco Zorzi, Ivilin Peev Stoianov, and Carlo Umilta. Computational modeling of numerical cognition. **The Handbook of Mathematical Cognition**, 2004.

[21] Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. Leveraging hand-object interactions in assistive egocentric vision. **IEEE Trans. Pattern Anal. Mach. Intell.**, Vol. 45, No. 6, p. 6820–6831, June 2023.

[22] Mandar Sharma, Rutuja Taware, Pravesh Koirala, Nikhil Muralidhar, and Naren Ramakrishnan. Laying anchors: Semantically priming numerals in language modeling. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 2653–2660, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[23] Steven P. Tipper. The negative priming effect: Inhibitory priming by ignored objects. **The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology**, Vol. 37A, No. 4, pp. 571–590, 1985.

[24] Bruce Milliken and Adrienne Rock. Negative priming, attention, and discriminating the present from the past. **Consciousness and Cognition**, Vol. 6, No. 2, pp. 308–327, 1997.

[25] Jinze Bai, et al. Qwen technical report, 2023.

[26] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

A　モデルの応答の処理

モデルによって生成された文章は、個々の単語に分割され、小文字に変換された後、各タスクについてそれぞれ以下のフローで分類し、正誤判定を行う。

Entity Question

- 生成されたテキストに答えの単語が存在する。
  → **正解**とする。
- 生成されたテキストに答えに単語が存在しない。
  → **不正解**とする。

SST2

生成されたテキストの中に positive または negative という単語の含まれ方で、positive, negative, improper（不適切な応答）に分類する。

- 生成されたテキストに “positive” という単語存在し、“negative” という単語が存在しない。
  → **“positive”** と分類し、正解ラベルと比較する。
- 生成されたテキストに “negative” という単語存在し、“positive” という単語が存在しない。
  → **“negative”** と分類し、正解ラベルと比較する。
- “positive” も “negative” 存在しない、あるいはどちらも存在している。
  → **“improper”** と分類し、不正解とする。

B　確信度の計算

本研究では、各選択肢の確信度を、モデルが生成する語彙全体の対数確率に基づいて計算した。具体的には、各選択肢の対数確率を標準確率に変換し、選択肢全体の確率が 1 になるよう正規化した。また、EQ の 4 つの選択肢においては、ダミー選択肢を同じ質問パターンからランダムに選定した。正規化の数式は以下のように表される：

1. 対数確率を指数関数に変換

$$S = \sum_{n=1}^{\infty} a_n \tag{1}$$

2. 正規化定数（合計確率）を計算

$$Z = \sum_{i=1}^{N} \exp(\log p_i) \tag{2}$$

3. 正規化後の確率を計算

$$\tilde{p}_i = \frac{\exp(\log p_i)}{Z} \tag{3}$$

4. すべての選択肢の確率の合計

$$\sum_{i=1}^{N} \tilde{p}_i = 1 \tag{4}$$

C　各タスクと使用画像の詳細

表 3 に EQ の質問パターンごとにあらかじめ設定した 2 種類の敵対画像を、表 4 に SST2 で使用した絵文字画像を掲載している。

D　LVLM のプライミング効果の例

図 4 では EQ タスクにおいて、視覚情報により（ネガティブ）プライミング効果が引き起こされている一例を掲載している。

表 3　EQ の各設問パターンに対する敵対画像の設定。敵対画像が質問の答えと一致しないように各質問パターンに対して、2 種類ずつ敵対画像を準備している。

質問パターン		敵対画像 1	敵対画像 2
P17: P19: P20: P69: P131: P159: P276: P495: P740:	Which country is [E] located in? Where was [E] born? Where did [E] die? Where was [E] educated? Where is [E] located? Where is the headquarter of [E]? Where is [E] located? Which country was [E] created in? Where was [E] founded?		
P26: P40: P50: P112: P127: P170: P175:	Who is [E] married to? Who is [E]'s child? Who is the author of [E]? Who founded [E]? Who owns [E]? Who was [E] created by? Who performed [E]?		
P800:	What is [E] famous for?		
P136:	What type of music does [E] play?		
P264:	What music label is [E] represented by?	SONY	UNIVERSAL MUSIC GROUP
P407:	Which language was [E] written in?	漢語 汉语 中文	العربية
P413:	What position does [E] play?		
P36:	What is the capital of [E]?		
P106:	What kind of work does [E] do?		
P176:	Which company is [E] produced by?	facebook	SONY

表 4　SST2 で使用した絵文字

感情ラベル	絵文字画像
positive	😄 😊 🥰 🤩 🤗
negative	😱 😵 😭 😫 😡

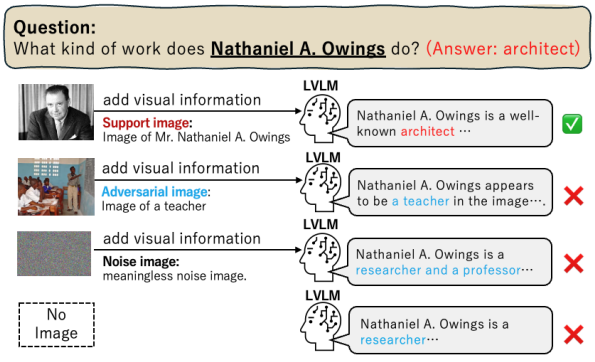

図 4　視覚情報によるプライミングの例：テキストのみの場合では誤答してしまうような問題（ある人物の職業を問う問題）において、質問文で出現する人物の画像を与えた場合、モデルはその人物の職業を回答できているが、答えと異なる職業を表す画像を与えると、モデルはその視覚情報にある職業を参考にしてしまい、誤った応答をしている。