メタファー判断における LLM-as-a-Judge の安定性分析：モデル・プロンプト・人間判断の比較

王　略丞 [1]　林　克彦 [1]　上垣外　英剛 [2]
[1] 東京大学　[2] 奈良先端科学技術大学院大学
{wanglc,katsuhiko-hayashi}@g.ecc.u-tokyo.ac.jp　kamigaito.h@is.naist.jp

概要

近年，大規模言語モデル（LLM）を判定器（LLM-as-a-Judge）として用い，データセット構築や評価を行う研究が数多く報告されている．しかし，これらの手法は判定の安定性を暗黙に前提としており，メタファー判断のような主観性の強いタスクにおいて，その前提が成立するかは十分に検証されていない．判定の不安定性やプロンプト依存性が大きい場合，LLM-as-a-Judge に基づく評価の信頼性や再現性に影響を及ぼす可能性がある．本研究では，英語メタファー判断を対象に，Instruction tuning されたパラメータ規模の近い複数のオープンソース LLM を LLM-as-a-Judge として用いた場合の判定特性を分析する．意味的に等価な複数プロンプトに対する判断の一致性と，人手アノテーションに付随する判断自信度との関係に着目し，モデル・プロンプト・人間判断の対応関係を検討する．MOH-X データセットを用いた実験の結果，判定の安定性はモデルおよびプロンプトの選択により大きく異なり，一部のモデルでは，人間判断の自信度が高い事例ほど LLM による判定との整合性が高まる傾向が確認された．

1　はじめに

近年，大規模言語モデル（LLM）を判定器として用い，データセット構築，注釈支援，自動評価を行う LLM-as-a-Judge の枠組みが広く用いられている [1, 2]．これらの手法は，人手コストの削減や評価の効率化に寄与し [3]，自然言語処理分野における多様なタスクへの応用が進められている．一方で，LLM-as-a-Judge による判定はモデル内部の判断基準に依存しており，その判定が一貫した基準に基づいて安定に行われているかどうかは明らかではない．

特に，判定基準を定義しにくい主観的判断タスクにおいては，問い方などの表層的な違いが判定結果に影響を及ぼす可能性が指摘されている [4]．このような不安定性やプロンプト依存性が大きい場合，LLM-as-a-Judge に基づく評価やアノテーションは体系的な偏りを含み，再現性や信頼性の観点から問題となり得る．そのため，LLM を判定器として用いる際には，性能指標に加えて判定の安定性や人間判断との対応関係を検証することが重要である．

メタファー判断は，文脈に基づく意味解釈を必要とし，人間アノテータ間においても判断のばらつきが生じやすい言語現象である [5]．この特性から，主観的解釈を伴う状況下における LLM の判定安定性や人間判断との対応関係を検討する上で，厳しい評価対象であると考えられる．

本研究では，英語メタファー判断タスクを対象に，LLM-as-a-Judge として用いた場合の判定挙動を，判定の安定性の観点から分析する．具体的には，Instruction tuning されたパラメータ規模の近い複数のオープンソース LLM を対象として，モデル間の判定性能差，意味的に等価な複数プロンプトに対する判定一致性，および人手アノテーションに付随する判断自信度との対応関係を検討する．これにより，主観的判断を伴うタスクにおいて LLM-as-a-Judge を用いる際に生じ得る判定の不安定性やプロンプト依存性の実態を明らかにする．

2　関連研究

LLM を人手判断の代替あるいは補助として用いる LLM-as-a-Judge の枠組みは，自然言語生成タスクの自動評価や大規模データセット構築における注釈コスト削減を目的として広く用いられてきた [6]．近年では，生成結果の品質評価やモデル間比較といった評価設定において，LLM による判定結果を評価指標として用いる研究が多数報告されている [7, 8, 9]．これらの研究は，人間評価との相関や運用上の効率性の観点から，LLM-as-a-Judge の実用的有

効性を示している．

一方，プロンプトの設計や表現の違いが LLM の判断挙動や意思決定に影響を与える可能性も指摘されており [10, 11]，近年の研究では，プロンプト表現の差異により評価結果が変動し得ることや，自動評価における再現性・信頼性の課題が議論されている [12, 13, 14]．しかし，これらの分析は主に生成品質評価や事実性判定を対象としており，意味解釈や主観的判断が強く関与する言語現象に対する検討は限定的である．

メタファー判断は，文脈に基づく意味解釈を必要とする代表的な意味論的タスクであり，人間アノテータ間においても判断のばらつきが生じやすいことが知られている [15, 16]．このような主観的判断のばらつきは，評価や注釈における信頼性の問題としても指摘されており [17]，近年では BERT や LLM を用いたメタファー検出手法も提案されている [18, 19, 20]．しかし，これらの研究は主に性能向上を目的としており，LLM を判定器として用いた場合の判断の安定性やプロンプト依存性に焦点を当てた分析は行われていない．

以上を踏まえ，本研究は LLM-as-a-Judge の枠組みをメタファー判断タスクに適用し，モデル間差異，プロンプト表現の違い，および人間判断との対応関係を同一の実験設定下で体系的に分析する．特に，意味的に等価な複数プロンプトに対する判定の一致性と，人間アノテーションに付随する判断自信度との関係に着目すること．

3 実験設定

3.1 タスクとデータセット

本研究では，英語文におけるメタファー使用の有無を判定する二値分類タスクを扱う．各入力文に対し，当該文がメタファー表現を含むか否かを yes / no の形式で判断する．

実験には，Mohammad ら [5] により構築された MOH-X メタファー判断データセットを用いる．本データセットは動詞を中心とした英語文から構成され，各文に対してメタファー用法（metaphorical）または字義的用法（literal）のラベルが付与されている．さらに，各事例には複数の人間アノテータによる判断に基づく判断自信度（confidence）が付随情報として付与されている．

本研究で用いた MOH-X は全 1,639 文からなり，字義的用法が 1,229 文，メタファー用法が 410 文である．クラス不均衡を含むため，評価においては正解率に加え，クラス間のバランスを考慮した指標を用いる．また，判断自信度は，人間判断における確信度の指標として扱い，LLM の判定挙動との関係を分析する．

3.2 モデル

LLM-as-a-Judge として，Instruction tuning されたパラメータ規模の近い複数のオープンソース LLM を用いる．具体的には，Llama-3.1-8B-Instruct[21], Gemma-2-9B-it[22], Mistral-7B-Instruct-v0.3[23] の 3 種類を対象とする．これらはいずれも指示追従能力を目的として調整されたモデルであり，分類や判断タスクへの適用が想定されている．また，パラメータ規模はいずれも 7–9B 程度と近接しており，モデルサイズ差に起因する影響を抑えた比較が可能である．本研究では，LLM-as-a-Judge として用いた場合の判定挙動の安定性およびプロンプト表現に対する感度に着目するため，すべてのモデルに対して同一の条件を適用する．

3.3 プロンプト設計

メタファー判断を行わせるため，質問応答（QA）形式のプロンプトを採用する．各プロンプトは，入力文がメタファー表現を含むか否かを直接問う質問文と，*yes* / *no* の二値応答のみを生成するよう指示する回答指示から構成される．この形式により，生成の自由度を抑え，判定結果の比較を容易にする．

プロンプト設計では，判断対象となる意味内容を固定したまま質問文の表現のみを変化させることで，問い方の違いが LLM の判定挙動に与える影響を検証する．具体的には，意味的には等価であるが表現の異なる 3 種類の質問文を用意した．各プロンプトでは，質問文以外の構成要素（判断対象文および回答形式）をすべて共通とした．

```
Q: <prompt>
Answer only "yes" or "no".
Sentence: "{s}"
A:
```

ここで，`{s}` は判断対象となる入力文を表す．用いた質問文（`<prompt>`）は以下の 3 種類である．

- P1: Is the following sentence metaphorical?
- P2: Does the following sentence contain a metaphor?
- P3: Is a metaphor used in the following sentence?

本研究では，特定のプロンプト表現の最適化ではなく，意味的に等価な問い方に対する判定結果の一貫性および安定性を評価対象とする．

3.4 推論設定

すべての実験は Apple Silicon 環境において Metal Performance Shaders（MPS）を用いて実行した．数値精度には，計算の安定性および再現性を考慮し，32-bit 浮動小数点（fp32）を用いた．推論時の生成設定として，最大生成トークン数を 5，温度パラメータを 0 とし，確率的サンプリングを行わない決定論的な出力を得た．また，すべてのモデルに対して同一の推論設定および乱数シードを適用した．

3.5 評価方法

判定性能の評価には，人手アノテーションとの一致率として，正解率（Accuracy）および Macro-F1 スコアを用いる．MOH-X はクラス不均衡を含むため，多数クラスへの偏りを考慮し，クラス間を等重みで評価する Macro-F1 を併せて報告する．各モデルの性能は，3 種類のプロンプトに対する判定結果を平均して算出する．

また，プロンプト表現に対する判定の安定性を評価する指標として，プロンプト間一致率（prompt agreement）を用いる．これは，同一文に対する各プロンプトの判定結果の一致度を表す指標であり，あるプロンプト p_i に対する一致率は，残りの 2 つのプロンプト p_j, p_k との pairwise 一致率の平均として次式で定義する：

$$\mathrm{Agreement}(p_i) = \frac{1}{2}\left(\mathbb{I}[p_i = p_j] + \mathbb{I}[p_i = p_k]\right).$$

モデルごとの一致率は，全事例にわたって平均した値として算出する．

さらに，人間判断との対応関係を分析するため，判断自信度に基づき事例を区分し，各区間における LLM の判定正解率を比較する．

4 結果と分析

4.1 モデル間の判定性能比較

表 1 は，LLM-as-a-Judge として用いた各モデルのメタファー判断性能を示している．ここでは，意味的に等価な 3 種類のプロンプトに対する判定結果を平均することで，モデル固有の判定特性を比較する．評価指標としては，正解率に加え，クラス不均衡の影響を考慮した Macro-F1 スコアを用いる．

表 1 モデル間のメタファー判断性能（プロンプト平均）

モデル	正解率 (%)	Macro-F1 (%)
Llama-3.1-8B-Instruct	**78.8**	66.2
Gemma-2-9B-it	70.2	**67.6**
Mistral-7B-Instruct-v0.3	55.1	52.9

Llama は，3 モデル中で最も高い正解率を示しており，全体として比較的高い判定性能を示した．一方，Gemma は正解率では Llama に及ばないものの，Macro-F1 スコアは同程度，あるいはわずかに高い値を示している．このことは，Gemma がクラス間の偏りを抑えた比較的バランスの取れた判定を行っている可能性を示唆する．

これに対し，Mistral は，正解率および Macro-F1 ともに他の 2 モデルを大きく下回った．この結果は，モデル間で LLM-as-a-Judge としての判定挙動に大きな差異が存在することを示している．モデル別・プロンプト別の詳細な性能指標は付録の表 4 に示す．

4.2 プロンプト表現に対する安定性分析

意味的には等価であるが表現の異なる複数のプロンプトに対して，LLM の判定結果がどの程度安定しているかを分析する．表 2 は，各プロンプトに対する判定性能を 3 種類のモデルにわたって平均した結果である．

表 2 プロンプト別メタファー判断性能（モデル平均）

プロンプト	正解率 (%)	Macro-F1 (%)
Prompt 1	60.0	55.7
Prompt 2	70.7	63.5
Prompt 3	**73.4**	**67.5**

Prompt 1 は，正解率および Macro-F1 とともに他のプロンプトより低い値を示しており，プロンプト表現の違いが判定性能に影響を与え得ることが確認された．一方，Prompt 2 および Prompt 3 は正解率の観点では比較的高い性能を示すが，Macro-F1 スコアには差が見られる．特に Prompt 2 では，正解率に比して Macro-F1 が低く，クラス間で偏りのある判定が生じている可能性が示唆される．各モデルごとのプロンプト別挙動は付録の表 4 に示されている．

これに対し，Prompt 3 は正解率と Macro-F1 の双方が高く，クラスバランスの観点からも比較的安定した判定性能を示した．この結果は，正解率のみを用いた評価では，プロンプト表現に起因する判定の偏りを捉えきれない可能性があることを示している．

次に，LLM によるメタファー判断と人間判断との対応関係を検討する．表 3 は，判断自信度の区間ごとに，各モデルの判定正解率を示したものである．

表 3　判断自信度区間ごとの判定正解率 (%)

自信度区間	Llama	Gemma	Mistral
0.5–0.7	50.1	59.0	58.9
0.7–1.0	85.2	70.7	53.5
1.0	**94.7**	**86.3**	56.8

Llama および Gemma では，判断自信度が高い事例ほど正解率が高くなる明確な対応関係が一貫して確認された．なお，低自信度区間（0.5–0.7）には，メタファー用法の割合が相対的に高く，人間アノテータ間でも判断が分かれやすい境界的な事例が多く含まれている（付録の表 6）．一方，Mistral では，自信度区間による判定性能の差が他のモデルと比べて小さいことが確認された．

4.3　人間判断との対応関係

前節の結果から，判断自信度が低い境界的な事例においては，全体として正解率が低下する傾向が観察され，人間にとって判断が困難な文は LLM にとっても判定が難しい場合が多いことが確認された．この結果は，主観的解釈を伴う事例において，LLM による判定の信頼性を評価する際には，安定性を併せて検討する必要があることを示唆している．

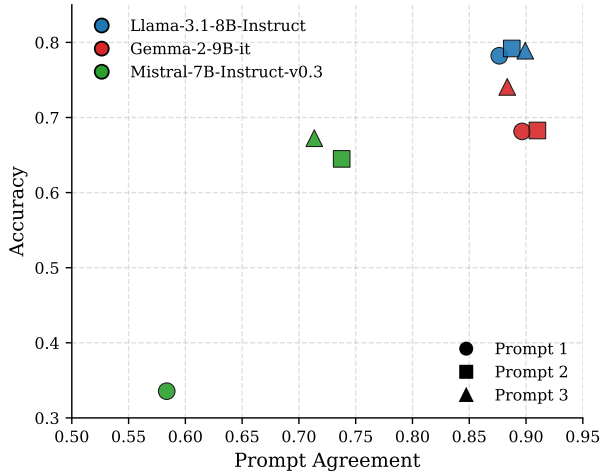

図 1　各モデル・各プロンプトにおける判定正解率とプロンプト間一致率の関係．各点は単一モデルにおける単一プロンプトの判定性能を表す．

さらに，各モデルにおける判定正解率とプロンプト間一致率との関係を図 1 に示す．全体として，プロンプト間一致率が高いほど正解率も高くなる傾向が観察され，問い方の違いに対してより安定した判定を行うモデルほど，人間判断との整合性が高い可能性が示された．プロンプト間の pairwise 一致率の詳細は付録の表 5 に示す．

5　考察

本研究の結果は，LLM-as-a-Judge によるメタファー判断が，モデル選択およびプロンプト表現に大きく依存することを示している．特に，プロンプト間一致率と判定正解率との間に観察された対応関係から，問い方の違いに対して頑健なモデルほど，人間判断とも整合的な判定を行う傾向があることが確認された．このことは，判定の安定性が単なる副次的な性質ではなく，LLM-as-a-Judge の信頼性を評価する上で本質的な観点として位置づけられることを意味する．

また，一部のプロンプトは正解率を向上させる一方で，クラス不均衡に起因する判定バイアスを強める場合があり，単一の性能指標のみに基づく評価では，判定特性の違いを十分に捉えられないことが明らかとなった．本研究の結果は，Macro-F1 などの補助的指標を併用することにより，プロンプト表現に起因する系統的な偏りをより適切に把握できる可能性を示唆している．

これらの知見は，主観的判断を伴うタスクにおいて LLM を判定器として利用する際には，性能指標に加えて判定の安定性や再現性を事前に検証することが方法論的に要請されることを示している．

6　おわりに

本研究では，英語メタファー判断タスクを対象として，LLM-as-a-Judge による判定の安定性を，モデル，プロンプト表現，および人間判断との対応関係の観点から分析した．実験の結果，意味的に等価な問い方であっても，プロンプト表現の違いが判定性能および判定の一貫性に系統的な影響を与え得ることが示された．また，一部のモデルでは，人間判断の自信度が高い事例ほど，LLM による判定との一致率が高まる傾向が確認された．これらの結果は，主観的判断を含むタスクにおいて LLM-as-a-Judge を用いる際には，単一の性能指標のみに依拠するのではなく，判定の安定性やプロンプト依存性を併せて評価する必要があることを示唆している．

今後は，これらの判定特性が生じる要因を内部処理の観点から考察するとともに，主観的評価タスクでの LLM-as-a-Judge の利用指針をさらに検討する．

参考文献

[1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In **Advances in Neural Information Processing Systems**, Vol. 36, pp. 46595–46623. Curran Associates, Inc., 2023.

[2] Yang Liu, Dan Iter, Yichong Xu, et al. G-eval: NLG evaluation using gpt-4 with better human alignment. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[3] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. **Proceedings of the National Academy of Sciences**, Vol. 120, No. 30, July 2023.

[4] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics.

[5] Saif Mohammad, Ekaterina Shutova, and Peter Turney. Metaphor as a medium for emotion: An empirical study. In **Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics**, pp. 23–33, Berlin, Germany, August 2016. Association for Computational Linguistics.

[6] Jiawei Gu, Xuhui Jiang, Zhichao Shi, et al. A survey on llm-as-a-judge, 2025.

[7] Jiaan Wang, Yunlong Liang, Fandong Meng, et al. Is ChatGPT a good NLG evaluator? a preliminary study. In **Proceedings of the 4th New Frontiers in Summarization Workshop**, pp. 1–11, Singapore, December 2023. Association for Computational Linguistics.

[8] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, et al. Agent-as-a-judge: Evaluate agents with agents, 2024.

[10] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[11] Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. Cognitive bias in decision-making with LLMs. In **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 12640–12653, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] Peiyi Wang, Lei Li, Liang Chen, et al. Large language models are not fair evaluators. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[13] Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. In **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 297–312, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[14] Hui Wei, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates, 2025.

[15] Ekaterina Shutova, Simone Teufel, and Anna Korhonen. Statistical metaphor processing. **Computational Linguistics**, Vol. 39, No. 2, pp. 301–353, June 2013.

[16] Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. **A Method for Linguistic Metaphor Identification: From MIP to MIPVU**. No. 14 in Converging Evidence in Language and Communication Research. John Benjamins Publishing Company, Amsterdam, 2010.

[17] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In **Proceedings of the 24th Annual Conference of the European Association for Machine Translation**, pp. 193–203, Tampere, Finland, June 2023. European Association for Machine Translation.

[18] Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. Neural metaphor detection in context. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 607–613, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[19] Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, et al. A report on the 2020 VUA and TOEFL metaphor detection shared task. In **Proceedings of the Second Workshop on Figurative Language Processing**, pp. 18–29, Online, July 2020. Association for Computational Linguistics.

[20] Rui Mao, Chenghua Lin, and Frank Guerin. End-to-end sequential metaphor identification inspired by linguistic theories. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3888–3898, Florence, Italy, July 2019. Association for Computational Linguistics.

[21] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024.

[22] Gemma Team, Morgane Riviere, Shreya Pathak, et al. Gemma 2: Improving open language models at a practical size, 2024.

[23] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b. In **The Twelfth International Conference on Learning Representations (ICLR)**, 2024.

A　付録

本付録では，主文中で報告した結果を補足するため，モデル別・プロンプト別の詳細な性能指標，プロンプト間一致率の内訳，および判断自信度区間ごとのサンプル分布を示す．

表 4　モデル別・プロンプト別メタファー判断性能（正解率および Macro-F1）

Prompt 1		
モデル	正解率 (%)	Macro-F1 (%)
Llama-3.1-8B-Instruct	78.2	69.1
Gemma-2-9B-it	68.2	66.0
Mistral-7B-Instruct-v0.3	33.6	32.0
Prompt 2		
モデル	正解率 (%)	Macro-F1 (%)
Llama-3.1-8B-Instruct	79.2	62.0
Gemma-2-9B-it	68.3	66.2
Mistral-7B-Instruct-v0.3	64.5	62.4
Prompt 3		
モデル	正解率 (%)	Macro-F1 (%)
Llama-3.1-8B-Instruct	78.9	67.6
Gemma-2-9B-it	74.1	70.5
Mistral-7B-Instruct-v0.3	67.2	64.4

表 5　モデル別プロンプト（P）間一致率（%）

モデル	P1-P2	P1-P3	P2-P3
Llama-3.1-8B-Instruct	86.5	88.8	91.0
Gemma-2-9B-it	92.3	87.0	89.7
Mistral-7B-Instruct-v0.3	60.8	56.0	86.8

表 6　判断自信度区間ごとのサンプル分布およびメタファー・字義用法の割合

自信度区間	総数	Metaphorical (%)	Literal (%)
0.5–0.7	357	56.0	44.0
0.7–1.0	1070	18.3	81.7
1.0	212	6.6	93.4