MLLM-as-a-Judgeは自己を選好する

小山 修生[1,a)]　平野 愼之助[1]　松田 一起[1]　杉浦 孔明[1,b)]

概要

本研究では，画像キャプション生成において，マルチモーダル大規模言語モデル (MLLM) が自身の生成文を不当に高く評価する傾向 (自己選好バイアス) の程度を明らかにする．テキストのみを入力とする LLM を対象とする自己選好バイアスについての既存研究においては，自己選好の定量化が不十分であるため，自己選好バイアスの度合いが不明である．そこで本研究では，MLLM の組に互いにキャプションの生成と評価を行わせ，評価値を標準化することでバイアスの程度を定量化し，MLLM の自己選好バイアスの度合いの検証を行った．検証の結果，画像キャプション生成において，検証対象の各 MLLM は，自己選好バイアスを有する可能性が高いことが明らかになった．また，同系列の各 MLLM がお互いの生成文を選好するバイアスの存在が示唆された．

1. はじめに

マルチモーダル大規模言語モデル (MLLM) は，自動運転やロボティクスを含む多様な分野において，研究および社会応用が進んでいる [12]．MLLM の性能評価を人手で行うことは時間・コスト面から不利であるため，性能評価に MLLM を用いるアプローチが一般的になりつつある [7, 5]．このアプローチは，LLM-as-a-Judge または MLLM-as-a-Judge と呼ばれる．一方，MLLM-as-a-Judge では，MLLM が自身の生成した文を不当に高く評価する自己選好バイアスが存在する可能性がある．これは，MLLM が生成した文が，評価時において生成時と同様に尤度が高いトークン系列であると考えられるためである．

画像キャプション生成における標準的な MLLM-as-a-Judge(例：G-VEval[17]) は，特定の MLLM を用いて評価を行う．仮に MLLM-as-a-Judge が特定の MLLM による生成文を過大評価するのであれば，評価の客観性が損なわれるおそれがある．

既存研究 [9, 14, 16] における自己選好バイアスの知見は，以下の点で不十分である．まず，我々の知る限り，複数のモダリティを入力とするタスクにおいて MLLM における自己選好バイアスの検証はほとんど行われていない．次に，既存研究においては，自己選好の定量化が不十分であるため，自己選好の度合いが不明である．

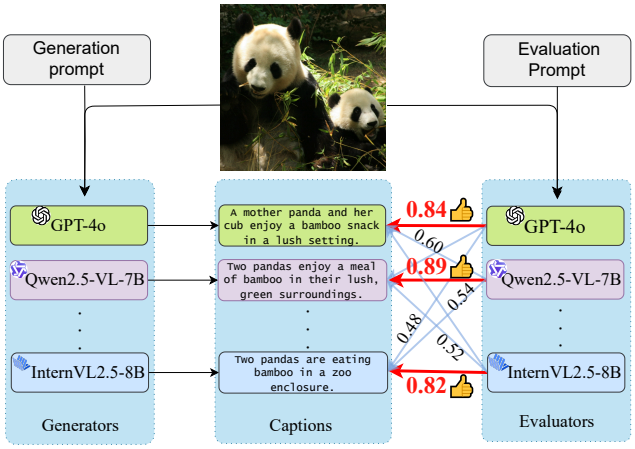

図 1 Generator と Evaluator の関係.

本研究では，提案手法である Cross-Model Preference Evaluation（X-PrefEVAL）を用いて，画像キャプション生成における MLLM の自己選好バイアスの検証を行う．X-PrefEVAL は，MLLM の組にキャプションの生成と評価 (例：FLEUR[10], G-VEval) を互いに行わせ，評価値を標準化することでバイアスの程度を定量化する．既存研究との主要な違いは画像キャプションにおいて，MLLM の自己選好バイアスを X-PrefEVAL を用いて定量的に比較する点である．本研究の貢献は以下である．

- 画像キャプション生成における MLLM の自己選好の程度を定量化する手法として，X-PrefEVAL を提案する．
- 画像キャプション生成において，各 MLLM の自己選好バイアスがどの程度生じるのかを明らかにする．

2. 問題設定

本論文において使用する用語を以下のように定義する．

- 生成文：MLLM よって生成された画像説明文.
- Evaluator：生成文の評価に使用する MLLM.
- Generator：生成文を生成する MLLM.
- 自己選好：Evaluator が自身の生成文をそれ以外の文よりも有利に評価すること.

MLLM-as-a-Judge では，MLLM が自身の生成した文を不当に高く評価する自己選好バイアスが存在する可能性が

[1] 慶應義塾大学
[a)] koyamashu3@keio.jp
[b)] komei.sugiura@keio.jp

ある．このような仮説が立てられるのは，以下の理由による．Generator の出力するキャプションは，画像と生成用プロンプトを条件とした際に最も高い尤度を持つトークン系列である．Generator と Evaluator が同一モデルである場合，同一画像と評価用プロンプトで条件付けられた当該トークン系列の尤度は評価時にも高くなる可能性がある．実際に，文生成タスクにおける LLM を用いた評価では，評価用モデルにとって高い尤度のトークン系列に高い評価値を割り当てる傾向が報告されている [15]．したがって，画像を条件とした確率的言語モデルである MLLM には，画像キャプション生成における自己選好バイアスがあると予想される．

本研究では，MLLM の性能を測る標準的なタスクであることから，画像キャプション生成を対象に選択する．ここで，ある Evaluator が自身の生成文を他 Evaluator より相対的に高く評価する傾向が強いほど，自己選好バイアスが大きいとみなす．

3. 提案手法

我々は，Evaluator による生成文の評価値の統計的傾向を用いて自己選好バイアスを検証する．本手法は，MLLM-as-a-Judge に限らず，生成モデルによる評価における自己選好バイアスの検証に広く適用可能である．図 1 に，Generator と Evaluator の関係を示す．Generator は画像のキャプションを生成し，Evaluator はその生成文の評価を行う．ここで，ある Evaluator が自身の生成文を他 Evaluator より相対的に高く評価する傾向が強いほど，自己選好バイアスが大きいと言える．

3.1 X-PrefEVAL

本手法は，3 つのステップから構成される．まず，各 Generator によって画像群に対して生成文を生成する．次に各 Evaluator を用いて生成文を評価する．最後に，評価値の統計的傾向を捉える．Algorithm 1 に本手法の疑似コードを示す (詳細は Appendix B 参照)．

Step1. 本ステップでは，各 Generator を用いて画像に対する生成文を生成する．Generator は画像 $\boldsymbol{x}_{\text{img}}$，生成用プロンプト $\boldsymbol{x}_{\text{gprompt}}$ を入力として，生成文のトークン系列 $\hat{\boldsymbol{y}}_{\text{g}}$ を出力する．

Step2. 本ステップでは，各 Evaluator を用いて，$\hat{\boldsymbol{y}}_{\text{g}}$ に対する評価値 $\psi \in [0, 1]$ を算出し，以下で定義される行列 $\boldsymbol{S}$ を得る．我々は標準的な手法 [10, 17] に則って，トークンの出力確率分布を用いて ψ を算出する．i 番目の Generator による生成文に対する，Evaluator が出力する評価値の平均を次のように定義する．

$$\Psi_{\text{Evaluator}}(\text{Generator}) = \frac{1}{N}\sum_{k=1}^{N} \psi_{\text{Evaluator}}(\hat{\boldsymbol{y}}_g^{(i,k)}) \quad (1)$$

ここで $\psi_{\text{Evaluator}}(\boldsymbol{y}_g^{(i,k)})$ は，Evaluator が i 番目の Generator による k 番目の生成文 $\hat{\boldsymbol{y}}_g^{(i,k)}$ に与えた評価値を表す．

Algorithm 1 X-PrefEVAL

Input: Generators $\mathcal{G} = \{G^{(i)}\}$, Evaluators $\mathcal{E} = \{E^{(j)}\}$, Images $\{\boldsymbol{x}_{\text{img}}^{(k)}\}$, References $\{\boldsymbol{X}_{\text{ref}}^{(k)}\}$, Generation prompt $\boldsymbol{x}_{\text{gprompt}}$, Evaluation prompt $\boldsymbol{x}_{\text{eprompt}}$
$\hat{\boldsymbol{y}}_g^{(i,k)} \leftarrow \textsc{Generate}(G^{(i)}, \boldsymbol{x}_{\text{img}}^{(k)}, \boldsymbol{x}_{\text{gprompt}}), \forall i, k$
for each $(G^{(i)}, E^{(j)}) \in \mathcal{G} \times \mathcal{E}$ **do** ▷ Caption evaluation
  $\boldsymbol{S}[i, j] \leftarrow \textsc{Average}_k(\textsc{Eval}(\boldsymbol{x}_{\text{img}}^{(k)}, \boldsymbol{X}_{\text{ref}}^{(k)}, \hat{\boldsymbol{y}}_g^{(i,k)}, \boldsymbol{x}_{\text{eprompt}}))$
end for
$\tilde{\boldsymbol{S}}[:, j] \leftarrow \textsc{Standardize}(\boldsymbol{S}[:, j]), \forall j$ ▷ Evaluator-wise
$\tilde{\boldsymbol{S}}[i, :] \leftarrow \textsc{Standardize}(\boldsymbol{S}[i, :]), \forall i$ ▷ Generator-wise
return $\tilde{\boldsymbol{S}}$

また，N は画像の枚数を表す．いま，M 種類の MLLM があるとき，Evaluator と Generator を入れ替えた全組み合わせに対する行列を $\boldsymbol{S} = [s_{ij}] \in \mathbb{R}^{M\times M}$ と定義する．ここで，$s_{ij} = \Psi_{\text{model-}j}(\text{model-}i)$ である．ただし，model-i, model-j はそれぞれ i，j 番目のモデルを表す．また，model-j が全モデルに対してつけた式 (1) における Ψ の平均値を次のように定義する．

$$\Psi_{\text{model-}j} = \frac{1}{M}\sum_{i=1}^{M} s_{ij} \quad (2)$$

Step3. 本ステップでは $\boldsymbol{S}$ を，全成分を直接の比較を可能な行列 $\tilde{\boldsymbol{S}} = [\tilde{s}_{ij}] \in \mathbb{R}^{M\times M}$ に変換する．Evaluator および Generator によって式 (1) で定義した Ψ の分布（平均および分散）が大きく異なる場合があることから，$\boldsymbol{S}$ の各成分は比較できない．そこで，我々は $\boldsymbol{S}$ を，以下で定義する直接各成分を比較できる行列 $\tilde{\boldsymbol{S}}$ に変換する．$\tilde{\boldsymbol{S}}$ は以下の 2 段階で計算される．まず，第一段階として Evaluator 間における $\boldsymbol{S}$ の成分の直接比較を可能にするため，Evaluator ごとに標準化を行う．次に，第二段階として，Generator 間における成分の直接の比較を可能にするため，Generator ごとに標準化を行う．

3.2 Self-Preference Index

以下で本手法における選好の評価尺度を定義する．$\tilde{\boldsymbol{S}}$ の各成分 $\tilde{s}_{ij}$ は，model-j が，model-i による生成文をどのように相対的に評価したかを表す値である．そこで，選好の評価尺度として $\tilde{S}_{\text{model-}j}(\text{model-}i)$ を以下のように定義する．

$$\tilde{S}_{\text{model-}j}(\text{model-}i) = \tilde{s}_{ij} \quad (3)$$

特に，$\tilde{S}_{\text{model-}i}(\text{model-}i)$ は自己選好の度合いを表し，これを Self-Preference Index と定義する．本手法で式 (3) の $\tilde{S}_{\text{model-}j}(\text{model-}i)$ を選好の評価尺度として用いるのは，Evaluator と Generator を入れ替えた全組み合わせの選好を定量的に直接比較できるためである．

4. 実験設定

上述の手法で画像キャプション生成における MLLM の自己選好バイアスを検証するためには，生成文とそれらに対する Evaluator および人間による評価が必要である．しかし，我々の知る限りこのようなデータセットは存在しない．そのため，我々は Generator による候補文，Evaluator

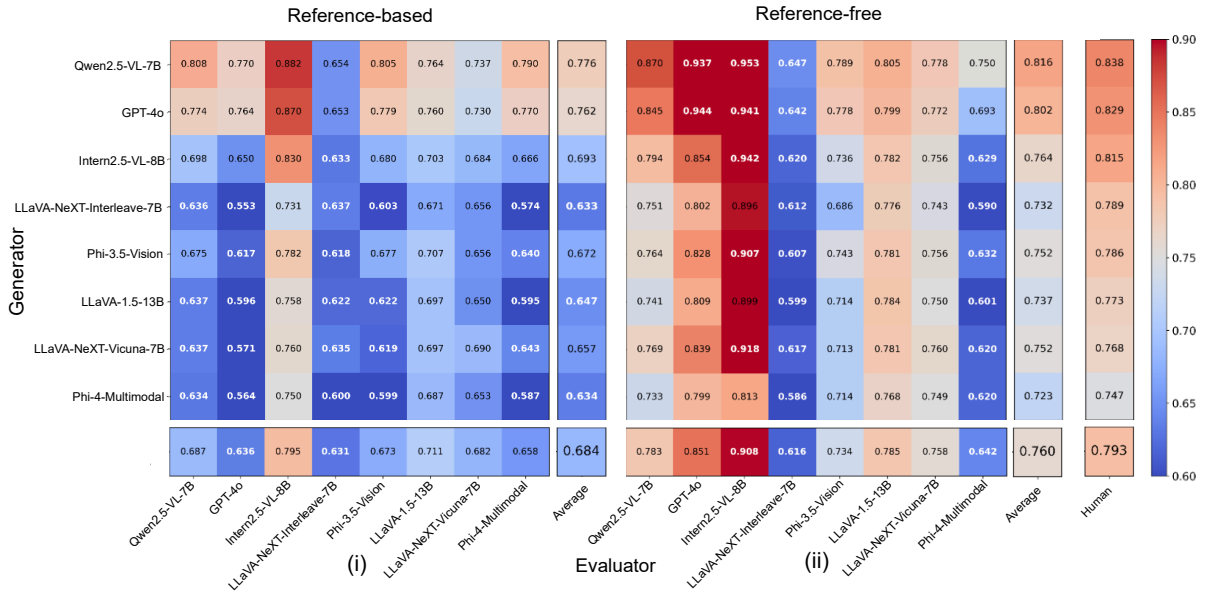

図 2 検証対象のモデルにおける $\boldsymbol{S}$ の定量的結果.

および人間による評価値を含む，自己選好バイアスを検証するためのデータセット SelfPref-Cap を構築した．

本研究では, Generator および Evaluator として Qwen2.5-VL-7B[3], GPT-4o[8], InternVL2.5-8B[6]，LLaVA-NeXT-Interleave-7B[11]，Phi-3.5-vision[1]，LLaVA-1.5-13B[13]，LLaVA-NeXT-Vicuna-7B[4]，Phi-4-Multimodal[2] を用いた．各モデルはいずれも，代表的な MLLM として位置付けられるものであるため採用した．また，様々な MLLM における自己選好バイアスの程度を明らかにするため，多様な開発元およびモデル系列に由来する MLLM を選定した．生成用プロンプトは既存研究 [13] において採用されているプロンプトを使用し，評価時は G-VEval の評価用プロンプトを使用した．

5. 実験結果

5.1 補正なしの評価値の定量的分析

図 2 に，SelfPref-Cap における $\boldsymbol{S}$ を示す．図 2-(i), (ii) はそれぞれ reference-based 設定，reference-free 設定における結果である．ここで，自己選好の検証を行うにあたり，Evaluator 間の成分を直接比較することは不適切である．実際，図 2-(ii) に示されるように，$\Psi_{\text{InternVL2.5-8B}}$ と $\Psi_{\text{LLaVA-NeXT-Interleave-7B}}$ はそれぞれ 0.908 および 0.616 であり，0.292 ポイント差がある．したがって，列間の成分を直接比較できるようにするためには，各 Evaluator 列ごとに標準化する必要がある．同様に，自己選好の検証を行うにあたり生成品質が異なる生成文の評価の各成分を直接比較することは不適切であるため，Generator ごとにも標準化する必要がある．

Phi-4-Multimodal を除く MLLM において，reference-based 設定では reference-free 設定と比較して評価値が低下する傾向が確認された．この要因として，reference-based 評価では，画像のみならず参照文の内容も考慮されるため，生成文における参照文と整合しない記述が低く評価されることが挙げられる．

図 2 において GPT-4o，Qwen2.5-VL-7B, InternVL2.5-8B による生成文は，人間による評価の平均値が高い傾向があった．一方で，Phi-4-Multimodal は他の Generator と比較して人間による評価の平均値が低い傾向があった．これらの傾向は，各 Evaluator においても同様に確認された．

5.2 自己選好の定量的分析

図 3 に，SelfPref-Cap における $\tilde{\boldsymbol{S}}$ を示す．対角成分は Self-Preference Index に対応し，自己選好バイアスの度合いを示す．Reference-based 設定では，非対角成分の平均値が -0.21 であるのに対し，対角成分の平均値が 1.51 と大きく上回っている．Reference-free 設定においても，非対角成分の平均値は -0.22 であるのに対し，対角成分の平均値は 1.57 と，同様に高く評価する傾向にあった．これらの結果から，MLLM による自己選好バイアスが存在することが示唆される．両設定において，対角成分と非対角成分の値には統計的に有意な差が認められた ($p < 0.05$).

図 3-(i) より，reference-based 設定において，Self-Preference Index は LLaVA-NeXT-Interleave-7B が 2.49 と最も高かった．これにより，当該モデルの自己選好バイアスが他のモデルと比較して最も顕著であることが示唆される．一方で，Phi-4-Multimodal における Self-Preference Index は 0.21 と最も低く，他の Generator による生成文を評価した際の値と比較して高くなかった. このことから，Phi-4-Multimodal の自己選好は，他のモデルと比較して小さいことが示唆される．

Reference-free 設定においては，InternVL2.5-8B の Self-Preference Index が 2.20 と最も高かった．よって，当該モデルの自己選好バイアスが他のモデルと比較して最も大きいと考えられる．一方で，Qwen2.5-VL-7B は Self-Preference Index が 0.52 と最も低く，他の Generator による生成文を評価した際の値と比較して高くはなかった．したがって，

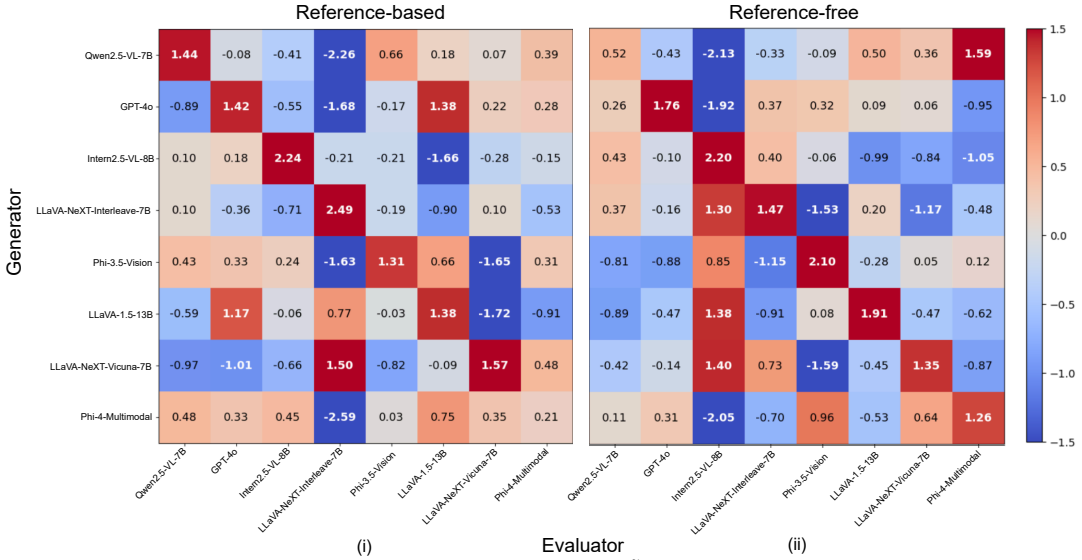

図 3 検証対象のモデルにおける $\tilde{S}$ の定量的結果.

Phi-4-Multimodal が自己選好は他のモデルと比較して相対的に小さいと考えられる.

また，図 3 より，LLaVA-NeXT-Interleave-7B は，同系列のモデルに対して高い評価を与える傾向がみられた．実際，図 3-(i) より，reference-based 設定において，$\tilde{S}_{\text{LLaVA-NeXT-Interleave-7B}}(\text{LLaVA-1.5-13B})$ および $\tilde{S}_{\text{LLaVA-NeXT-Interleave-7B}}(\text{LLaVA-NeXT-Vicuna-7B})$ はそれぞれ 0.77, 1.50 であり，いずれも他の Evaluator と比較して高い値であった．同様の傾向は reference-free 設定でも確認され，$\tilde{S}_{\text{LLaVA-NeXT-Interleave-7B}}(\text{LLaVA-NeXT-Vicuna-7B}) = 0.73$ と高い値であった．さらに，図 3 から，Phi-3.5-Vision および Phi-4-Multimodal もお互いを他モデルより相対的に高く評価する傾向が確認された．具体的には，reference-free 設定において $\tilde{S}_{\text{Phi-3.5-Vision}}(\text{Phi-4-Multimodal}) = 0.96$ であった．また，reference-based 設定において，$\tilde{S}_{\text{Phi-3.5-Vision}}(\text{Phi-4-Multimodal}) = 0.31$ であった．これらの同系列のモデルは他のモデルと比較して共通の学習データを多く含む [1, 2, 13, 4, 11] ため，学習データに起因する選好である可能性が示唆される.

加えて，図 3-(i) より reference-based 設定において $\tilde{S}_{\text{LLaVA-1.5-13B}}(\text{GPT-4o})$, $\tilde{S}_{\text{GPT-4o}}(\text{LLaVA-1.5-13B})$ はそれぞれ 1.38 および 1.17 であり，GPT-4o と LLaVA-1.5-13B は互いの生成文を選好する傾向にあった．LLaVA-1.5-13B の学習データは GPT-4 による合成データセットを含む [13] ことから，同様に学習データに起因する選好である可能性がある.

一方で，自己選好バイアスが他のモデルに比べて小さい MLLM もみられた．具体的には，Phi-4-Multimodal は reference-based 設定において，自己選好バイアスが他モデルと比べて小さかった．これは，Phi-4-Multimodal は同規模の他の MLLM と比較して，評価用プロンプトに従い参照文を重視する評価を行ったためであると考えられる．実際に，Phi-4-Multimodal は，instruction following の性能

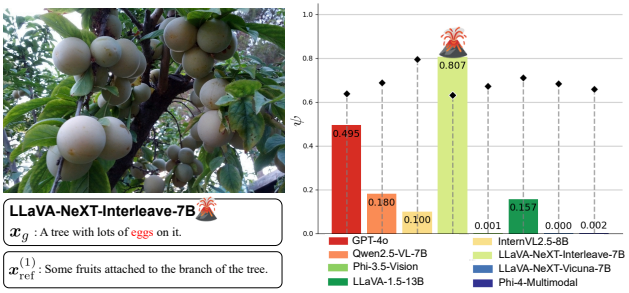

図 4 定性的結果．◆は，各 Evaluator による評価値の平均を表す．が高いことが報告されている [2].

5.3 自己選好の定性的結果

図 4 に，自己選好の定性的結果を示す．図 4 は，reference-based 設定における，LLaVA-NeXT-Interleave-7B の自己選好が顕著な例である．ただし，$\boldsymbol{x}_g$ における赤文字は重大な誤りを表す．図 4-(i) より，LLaVA-NeXT-Interleave-7B による生成文 “A tree with lots of eggs on it.” は果物を卵と誤認しており，ハルシネーションを含んでいた．しかし，reference-based 設定において，$\psi_{\text{LLaVA-NeXT-Interleave-7B}}(\boldsymbol{x}_g) = 0.807$ であった．これは $\Psi_{\text{LLaVA-NeXT-Interleave-7B}} = 0.606$ と比較して 0.201 ポイント高かった．一方，他の Evaluator は，一貫して当該生成文を各 Evaluator における $\Psi_{\text{Evaluator}}$ よりも低く評価しており，自身による評価のみが高かった.

6. おわりに

本研究では，画像キャプション生成において，MLLM による自己選好バイアスがどの程度生じるのかを検証した．本研究で得られた知見は次の通りである．(i) 画像キャプション生成において，検証対象の各 MLLM は，自己選好バイアスを有する可能性が高い．(ii)Phi-3.5-Vision および Phi-4-Multimodal など同系列の各 MLLM がお互いの生成文を選好するバイアスの存在が示唆された.

謝辞

本研究の一部は，JSPS 科研費 23K28168，JST ムーンショットの助成を受けて実施されたものである.

参考文献

[1] Abdin, M. et al.: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, *arXiv preprint arXiv:2404.14219* (2024).

[2] Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N. et al.: Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs, *arXiv preprint arXiv:2503.01743* (2025).

[3] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S. et al.: Qwen2.5-VL Technical Report, *arXiv preprint arXiv:2502.13923* (2025).

[4] Bo, L., Kaichen, Z., Hao, Z., Dong, G., Renrui, Z., Feng, L., Yuanhan, Z. et al.: LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild (2024).

[5] Chen, D., Chen, R. et al.: MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark, *ICML*, pp. 6562–6595 (2024).

[6] Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E. et al.: Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling, *arXiv preprint arXiv:2412.05271* (2024).

[7] Ge, W., Chen, S., Chen, G. H. et al.: MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria, *NAACL*, pp. 4951–4974 (2025).

[8] Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A. et al.: GPT-4o System Card, *arXiv preprint arXiv:2410.21276* (2024).

[9] Laurito, W. et al.: AI AI Bias: Large Language Models Favor Their Own Generated Content, *arXiv preprint arXiv:2407.12856* (2024).

[10] Lee, Y. et al.: FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model, *ACL*, pp. 3732–3746 (2024).

[11] Li, F. et al.: LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models, *arXiv preprint arXiv:2407.07895* (2024).

[12] Liang, Z., Xu, Y., Hong, Y., Shang, P., Wang, Q., Fu, Q. and Liu, K.: A Survey of Multimodel Large Language Models, *CAICE*, pp. 405–409 (2024).

[13] Liu, H. et al.: Improved Baselines with Visual Instruction Tuning, *CVPR*, pp. 26296–26306 (2024).

[14] Liu, Y., Moosavi, S. and Lin, C.: LLMs as Narcissistic Evaluators: When Ego Inflates Evaluation Scores, *ACL*, pp. 12688–12701 (2024).

[15] Ohi, M., Kaneko, M., Koike, R., Loem, M. and Okazaki, N.: Likelihood-based Mitigation of Evaluation Bias in Large Language Models, *ACL*, pp. 3237–3245 (2024).

[16] Panickssery, A., Bowman, S. and Feng, S.: LLM Evaluators Recognize and Favor Their Own Generations, *NeurIPS*, Vol. 37, pp. 68772–68802 (2024).

[17] Tong, T. C., He, S. et al.: G-VEval: A Versatile Metric for Evaluating Image and Video Captions Using GPT-4o, *AAAI*, Vol. 39, No. 7, pp. 7419–7427 (2025).