

Said Aloud, Read Different: Cross-Modal Instability in Multimodal Models

Basel Mousi, Fahim Dalvi, Shammur Chowdhury, Firoj Alam, Nadir Durrani

Qatar Computing Research Institute, HBKU, Qatar

{bmousi, faimaduddin, shchowdhury, fialam, ndurrani}@hbkku.edu.qa

Abstract

Multimodal foundation models are increasingly used in speech-first assistants that must interpret spoken queries and produce visually grounded decisions. Yet it remains unclear whether semantically equivalent queries yield consistent judgments across modality (text vs. speech) and language (English vs. Arabic). We introduce a *speech-augmented visually grounded contrastive triplet benchmark* spanning 10,150 culturally grounded images from 18 MENA countries, where each image is paired with one supported statement and two plausible but unsupported alternatives. We define contrastive instability as the conditional rate at which a model fails to resolve all statements within a triplet, isolating fragmented reasoning from complete failure. Evaluating recent multimodal models under text and speech in English and Arabic, we find that modality and language shifts introduce substantial triplet-level inconsistencies that are not fully captured by aggregate accuracy, with speech amplifying partial failures. We make the benchmark publicly available to the community.¹

Index Terms: Multimodal models, cross-modal consistency, speech-text grounding, multilingual evaluation

1. Introduction

Multimodal foundation models are increasingly accessed through speech interfaces [1, 2]. In principle, a spoken query that preserves semantic meaning should yield the same visually grounded decision as its textual counterpart. However, modality is not merely a neutral delivery channel [3, 4]. The pathway through which language enters the system may reshape how linguistic signals interact with visual evidence. This raises a central question: *are grounded decisions consistent when equivalent inputs are presented across different modalities and languages?*

Recent work shows that speech-capable systems can exhibit hallucination-like failures, including fluent ASR transcripts weakly grounded in audio [5, 6], increased hallucination under distribution shift [7], and cross-modal grounding failures in audio-visual LLMs that ignore audio and rely on visual priors [8, 9]. However, most studies evaluate hallucination *within a single channel* (speech-to-text) or focus on audio-visual description tasks. Our work instead asks whether *visually grounded decisions remain consistent* when the *same intended statement* is delivered via text vs. speech across languages.

Evaluating behavioral consistency across modality requires more than aggregate accuracy [10]. A model may answer many queries correctly while still exhibiting instability in how it distinguishes supported from unsupported interpretations [11]. To

probe consistency directly, we adopt a **contrastive evaluation design** in which each image is paired with one visually supported statement and two plausible but unsupported alternatives. This structure requires the model to make coherent distinctions within a semantically matched set, rather than succeed on isolated statements. By evaluating performance at the level of the triplet, we can assess whether equivalent inputs lead to consistent grounded decisions across modalities and languages.

To our knowledge, there is no publicly available dataset that supports *contrastive evaluation* in an image-grounded multimodal setting spanning text and speech in Arabic and English. We therefore construct a culturally grounded multimodal dataset spanning 18 countries across the Middle East and North Africa (MENA). The dataset covers visually distinctive themes including architecture, traditional clothing, cuisine, religious settings, and public spaces. This diversity enables the creation of semantically plausible yet visually unsupported alternatives, ensuring that distractors remain contextually coherent rather than trivial negatives. For each image, we generate one visually supported statement and two plausible alternatives, forming the contrastive triplets used in our analysis. An example is shown in Figure 1.

Contrastive Instability To quantify decision consistency, we evaluate performance at the triplet level. A model achieves full consistency only if it answers all three statements correctly. However, models may exhibit partial success, resolving some statements correctly while failing others. We therefore define contrastive instability as the rate at which a model fails to achieve full consistency among triplets where it answers at least one statement correctly. This conditional formulation isolates fragmented reasoning from complete failure and reveals instability that may remain hidden under strong aggregate accuracy.

We evaluate recent multimodal foundation models namely: Qwen2.5 (3B and 7B Omni), Qwen3-30B-Omni [12], and Phi-4-multimodal-instruct [13], all supporting both text and speech inputs. For speech evaluation, we synthesize English and Arabic queries, enabling controlled comparisons across modality, language, and speaker variation. We also test robustness under speech perturbations to assess stability under realistic acoustic conditions. Our findings reveal:

- Speech inputs increase triplet-level instability that is not fully captured by aggregate Q^+ and Q^- accuracy, indicating degraded contrastive discrimination rather than performance.
- Instability is substantially higher in Arabic than in English, and speech amplifies this cross-lingual gap.
- Model scaling improves stability and robustness to acoustic noise, but does not eliminate cross-modal or cross-lingual inconsistency; joint speech-text input mitigates instability without fully restoring text-only coherence.

¹<https://huggingface.co/datasets/QCRI/M2CQA-S>



Figure 1: Q^+ (👉) The image shows national animal of Qatar.
 Q^- (👈) The image shows national animal of Morocco.
 Q^- (👈) The image shows national animal of Algeria.

We make the following contributions:

- We introduce a speech-augmented culturally grounded contrastive benchmark of 10,150 triplets, enabling evaluation of consistency across modality and language.
- We propose contrastive instability (CI), a triplet-level metric isolating decision incoherence beyond aggregate accuracy.
- We provide a cross-modal, cross-lingual evaluation of multimodal foundation models under text and speech, showing modality is not a neutral input channel.

2. Dataset

We curated a culturally grounded multimodal image dataset spanning 18 Arab countries, building on the OASIS image collection and M2CQA, to support cross-modal consistency evaluation [14, 15]. The dataset covers visually distinctive themes including architecture, traditional clothing, cuisine, religious settings, and public spaces, reflecting the importance of cultural context in multimodal understanding and evaluation [16, 17]. Images were collected from the open web using country-specific queries derived from a predefined taxonomy in a human-in-the-loop setup. Retrieval employed geolocalized search settings, image quality constraints, and Creative Commons license filtering, followed by exact and embedding-based near-duplicate removal to ensure diversity and regional appropriateness.

2.1. Contrastive Triplet Construction

For each image in the curated repository, we generate a multiple-choice question (MCQ) consisting of one visually supported answer and two culturally plausible distractors within the same thematic category. To ensure that items genuinely require visual grounding, we apply an image-blind filtering step using strong language models and discard MCQs that can be answered without the image, as they are likely solvable from linguistic or cultural priors alone. Each retained MCQ is converted into a contrastive triplet by rewriting its options as standalone statements: the correct answer becomes a visually supported statement (Q^+), and the distractors become visually unsupported alternatives (Q^-). The statements are minimally contrastive, differing along fine-grained visual or cultural attributes to ensure reliance on image evidence rather than lexical cues (see example in Figure 1).

Statement	Need-Image		Q-Correctness	
	AC1	Raw Agr.	AC1	Raw Agr.
Q^+	0.996	0.996	0.752	0.784
Q^-	0.997	0.997	0.775	0.803
Q^-	0.996	0.996	0.695	0.738
Avg	0.996	0.996	0.741	0.775

Table 1: Inter-annotator agreement for Need-Image and Q-Correctness labels. We report AC1 and percent agreement.

2.2. Human Verification and Annotation Reliability

To assess the reliability of the contrastive annotations, we conduct a human verification study on a randomly sampled subset comprising 15% of the dataset. The subset is stratified to ensure proportional coverage across countries and taxonomic categories, with images and their corresponding triplet statements sampled uniformly from each country. Annotators are presented with an image and its associated statements and assign two independent labels: (i) *Need-Image*, indicating whether answering the statement requires access to the image (*Yes*, *No*, *Unsure*), and (ii) *Q-Correctness*, indicating whether the statement is visually supported by the image (*Correct*, *Incorrect*, *Unsure*).

In cases of disagreement between annotators, a third annotator performs adjudication. We report both raw percent agreement and Gwet’s AC1 coefficient [18] to account for potential class imbalance. Table 1 summarizes the results. Agreement is near-perfect for the Need-Image label (AC1 $\approx$ 0.996; raw agreement $\approx$ 0.996), indicating consistent identification of statements that require visual evidence. For Q-Correctness, agreement is substantial overall (AC1 $\approx$ 0.741; raw agreement $\approx$ 0.775), with slightly higher consistency for Q^+ statements than for Q^- alternatives. Bootstrap 95% confidence intervals computed via item-level resampling are strictly above zero across conditions, confirming reliability well above chance.

Beyond aggregate agreement statistics, we examine disagreement cases to understand annotation variance. Qualitative inspection shows most disagreements involve visually subtle counterfactual distinctions or fine-grained cultural interpretations rather than systematic errors. Instances of image insufficiency are rare, consistent with the high Need-Image agreement. These findings indicate that the dataset provides a reliable foundation for evaluating triplet-level contrastive consistency.

2.3. Cross-Lingual and Spoken-Modality Construction

To evaluate grounded consistency across language and modality, we translate all English statements into Arabic using an in-house machine translation system [19]. Translation is performed at the statement level to preserve declarative structure and semantic equivalence of the contrastive triplets, avoiding paraphrasing or lexical restructuring that could alter the Q^+/Q^- relationship. We then synthesize spoken versions of each statement in English and Arabic using zero-shot neural speech synthesis,² conditioned on natural reference recordings from human speakers [20]. This setup enables prosodically natural speech while avoiding artifacts associated with fixed proprietary TTS personas. For each statement, we generate matched male and female voice variants, keeping the lexical content strictly identical to the original text (no reformulation or punctuation changes that affect meaning), ensuring speech remains

²<https://huggingface.co/coqui/XTTS-v2>

a modality-preserving transformation. To evaluate robustness under realistic acoustic conditions, we further augment the synthesized speech with controlled perturbations, including non-stationary background noise and synthetic reverberation to simulate diverse physical environments. This stress test allows us to determine whether modality-induced instability persists beyond idealized clean audio.

3. Contrastive Instability

Evaluating grounded reasoning using independent statements does not test whether a model can distinguish between closely competing interpretations. A model may correctly accept a visually supported statement in isolation while also accepting a plausible but unsupported alternative. Such behavior reflects inconsistent discrimination rather than coherent grounding.

Contrastive triplets explicitly introduce minimal semantic competition by pairing one visually supported statement (Q^+) with two plausible but unsupported alternatives (Q_1^- , Q_2^-). A triplet is considered *fully consistent* if the model correctly accepts Q^+ and rejects both Q_1^- and Q_2^- . Triplet accuracy measures the proportion of such fully correct triplets. However, triplet accuracy alone does not reveal how errors arise. A decrease in triplet accuracy may reflect complete misunderstanding (all statements incorrect) or partial inconsistency (some correct, but not all). These behaviors correspond to different failure modes. To isolate fragmented reasoning, we define **contrastive instability (CI)**. Let N_{partial} denote the number of triplets with at least one correct statement, and let $N_{\text{consistent}}$ denote the number of those triplets that are fully consistent. We define:

$$CI = 1 - \frac{N_{\text{consistent}}}{N_{\text{partial}}}$$

CI measures how often a model fails to resolve a triplet coherently given that it demonstrates at least partial success. Importantly, CI is not a monotonic transformation of triplet accuracy. Two models may exhibit identical triplet accuracy yet differ substantially in CI if one fails primarily through global errors while the other fails through internal inconsistency. By conditioning on partial success, CI separates decision-boundary instability from overall performance degradation. In cross-modal evaluation, this distinction is crucial. Modality shifts may preserve aggregate statement accuracy while increasing internal inconsistency across alternatives. Contrastive instability thus provides a complementary and more diagnostic view of grounded decision behavior across speech and text inputs.

4. Results

Setup We evaluate Qwen2.5-Omni-3B, Qwen2.5-Omni-7B, Qwen3-Omni-30B-A3B-Instruct, and Phi-4-multimodal-instruct using the vLLM-Omni framework.³ We consider two settings: (1) comparing text versus audio inputs while keeping task instructions textual, and (2) evaluating modality effects under joint signaling. All models are prompted to decide whether a statement is True or False using a constrained response format (“The final answer is: <True/False>”). All experiments use greedy decoding for reproducibility, and code and data are released.⁴

Evaluation We evaluate grounded decision consistency across modality, language, and model scale using Q^+ and Q^-

Model	Lg	Mode	$Q^+ \uparrow$	$Q^- \uparrow$	F1 $\uparrow$	CI $\downarrow$
Q2.5-3B	en	Text	0.93	0.91	0.92	0.20
	en	Speech	0.93	0.90	0.91	0.22
	ar	Text	0.66	0.92	0.77	0.43
	ar	Speech	0.35	0.94	0.51	0.71
Q2.5-7B	en	Text	0.94	0.92	0.93	0.17
	en	Speech	0.95	0.92	0.93	0.18
	ar	Text	0.90	0.82	0.86	0.35
	ar	Speech	0.88	0.62	0.73	0.63
Q3-30B	en	Text	0.93	0.91	0.92	0.20
	en	Speech	0.92	0.92	0.92	0.19
	ar	Text	0.92	0.86	0.89	0.28
	ar	Speech	0.92	0.73	0.81	0.47
Phi-4M	en	Text	0.57	0.94	0.71	0.50
	en	Speech	0.61	0.83	0.70	0.58
	ar	Text	0.45	0.77	0.57	0.75
	ar	Speech	0.04	0.96	0.08	0.97

Table 2: Triplet-level performance across language and modality. Q denotes Qwen models. $\uparrow$ indicates higher is better; $\downarrow$ indicates lower is better. CI denotes contrastive instability.

accuracy, F1, and Contrastive Instability (CI). While Q^+ and Q^- summarize isolated correctness and F1 captures aggregate performance, CI measures conditional incoherence within contrastive triplets. Table 2 reports results across English and Arabic under text and speech inputs. Based on these findings, we address the following research questions.

RQ1: Does CI provide information beyond Q^+ and Q^- accuracy? Q^+ and Q^- accuracies summarize aggregate correctness, but they do not capture whether a model coherently distinguishes supported interpretations from closely competing alternatives. As shown in Table 2, there are cases where aggregate accuracy remains relatively strong while CI remains high, revealing instability that is not visible from statement-level metrics alone. For example, Q2.5-3B in Arabic text achieves an F1 score of 0.77, yet its contrastive instability (CI) is 0.43. This indicates that although the model often answers individual statements correctly, it frequently fails to resolve the full triplet consistently. Such conditional incoherence is not visible from Q^+/Q^- accuracy alone. More generally, Table 2 reveals settings in which aggregate correctness remains high while internal decision consistency degrades. These observations demonstrate that standard accuracy metrics do not fully characterize grounded stability. CI provides a complementary diagnostic measure that isolates contrastive instability within semantically matched triplets.

RQ2: How does decision stability vary between text and speech? Table 2 shows that speech can substantially increase contrastive instability (CI) relative to text, particularly in Arabic and across multiple models. While English results remain largely stable across modalities, Arabic exhibits pronounced instability under speech. For example, Q2.5-3B’s CI rises from 0.43 in Arabic text to 0.71 in Arabic speech, indicating that many triplets that are partially correct become internally inconsistent when the same statements are delivered through speech. Similar trends appear across model scales: Q2.5-7B increases from 0.35 to 0.63, and Q3-30B from 0.28 to 0.47. These changes occur alongside moderate declines in F1, suggesting that speech introduces instability beyond simple accuracy degradation. Overall, these results indicate that speech is not a neutral input channel, and that modality effects are partic-

³<https://docs.vllm.ai/projects/vllm-omni/en/latest>

⁴<https://github.com/baselmousi/cfhr-ci>

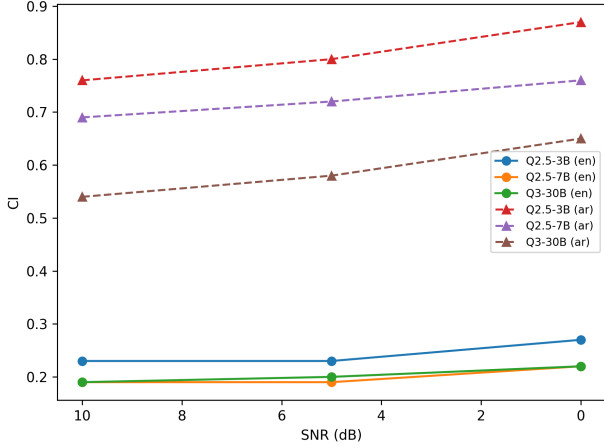


Figure 2: Contrastive instability (CI) as a function of SNR for Qwen models under speech input. Circles (solid) denote English and triangles (dashed) denote Arabic. Lower SNR indicates higher acoustic noise. Results are averaged across voices.

ularly pronounced in cross-lingual settings.

RQ3: How does instability vary across languages?

Across all models, contrastive instability is consistently higher in Arabic than in English (Table 2), and the gap widens under speech. For example, for Q2.5-3B under text, CI is 0.20 in English versus 0.43 in Arabic; under speech, CI is 0.22 in English versus 0.71 in Arabic. Similar gaps appear for larger models (e.g., Q3-30B: 0.20 vs. 0.28 in text; 0.19 vs. 0.47 in speech), suggesting that cross-lingual grounding introduces structural instability that is further amplified by spoken input.

RQ4: Does model scaling improve grounded stability?

Contrastive instability generally decreases with model size, particularly in Arabic (Table 2). Under text inputs, Arabic CI drops from 0.43 for Q2.5-3B to 0.35 for Q2.5-7B and further to 0.28 for Q3-30B, indicating improved discrimination between supported and unsupported statements as model capacity increases. A similar trend appears under speech input, where CI decreases from 0.71 to 0.63 and 0.47 across the same models. In English, instability remains comparatively low and largely stable across scales (e.g., $CI \approx 0.19$ – 0.22). Overall, scaling improves robustness and reduces contrastive instability, but does not fully eliminate cross-modal or cross-lingual inconsistency.

RQ5: How robust are models to acoustic degradation under speech input? Figure 2 reports contrastive instability (CI) as a function of SNR for Qwen models under speech input. CI increases monotonically as SNR decreases, indicating that acoustic noise exacerbates grounded decision instability. In English, the degradation is gradual, suggesting robustness under additive noise. In Arabic, the effect is markedly stronger, with sharper CI increases as SNR decreases, indicating that acoustic degradation compounds cross-lingual instability. Model scaling improves robustness, as Q3-30B maintains lower CI across SNR levels. However, the language gap persists even at higher SNRs, suggesting that noise amplifies pre-existing cross-lingual instability rather than causing it.

RQ6: Does joint speech–text signaling reduce instability? We also evaluate a joint-input setting where models receive both speech and text versions of the same statement. Figure 3 shows contrastive instability (CI) for Qwen models in Arabic under text-only, audio-only, and joint input. Audio-only yields the highest instability, text-only the lowest, and joint input consistently reduces CI relative to audio-only, indicating a stabiliz-

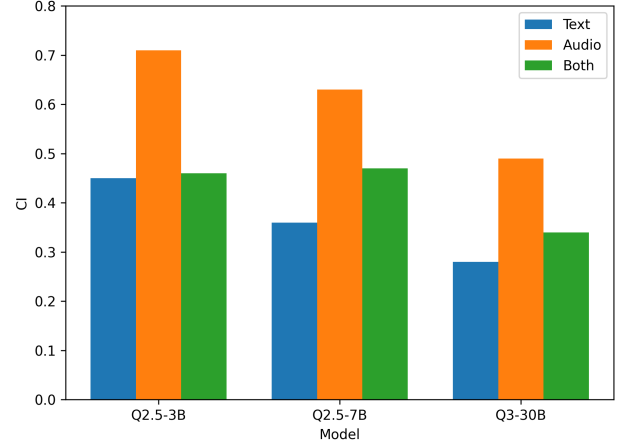


Figure 3: Contrastive instability (CI) under text-only, audio-only, and joint speech–text input for Qwen models in Arabic. Joint signaling reduces instability relative to audio-only input, but does not fully recover text-only stability.

ing effect from the textual signal. However, joint input does not fully recover text-only stability, particularly for smaller models. In English (not shown), differences across input conditions are minimal. Overall, multimodal redundancy mitigates but does not eliminate cross-lingual speech-induced instability.

5. Related Work

Recent work on multimodal evaluation has shown that aggregate accuracy alone is insufficient to characterize model reliability, motivating behavioral testing and contrastive probing frameworks that assess robustness under controlled variation [11, 21, 22]. Multimodal benchmarks such as MME [23], SEED-Bench [24] and Hal-Eval [25] evaluate grounded correctness across diverse tasks, yet primarily report independent-sample performance rather than decision coherence. In multilingual settings, culturally grounded benchmarks and dialect-aware models [16, 17, 26, 27] demonstrate that linguistic and regional variation can affect visual grounding behavior. Parallel work on speech-enabled systems highlights robustness challenges under transcription noise and distribution shift [2, 28], but does not examine whether semantically equivalent inputs delivered via text vs. speech yield consistent grounded decisions. Our work addresses this gap by introducing a conditional triplet-level measure of cross-modal instability that isolates internal discrimination failures across modality and language.

6. Conclusion

We examined whether multimodal foundation models make consistent visually grounded decisions when semantically equivalent inputs are presented through speech and text. Using a culturally grounded contrastive benchmark and introducing contrastive instability (CI), a conditional triplet-level measure of internal decision incoherence, we show that modality is not a neutral channel of input. Speech often preserves aggregate Q^+ and Q^- accuracy while increasing triplet-level instability, particularly in Arabic and under acoustic degradation. Although model scaling and joint speech–text signaling improve robustness, they do not fully eliminate cross-modal or cross-lingual instability. These findings highlight the importance of contrastive evaluation for diagnosing grounded reasoning stability beyond standard accuracy metrics.

7. Use of Generative AI Disclosure

Generative AI tools were used in data construction, including question generation, translation, and speech synthesis, and for writing assistance. All research design, analysis, and conclusions are the authors' own..

8. References

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, PMLR, 2023, pp. 28 492–28 518.
- [2] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15 757–15 773. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.1055/>
- [3] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, p. 423–443, Feb. 2019. [Online]. Available: <https://doi.org/10.1109/TPAMI.2018.2798607>
- [4] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao, "Multimodal foundation models: From specialists to general-purpose assistants," 2023. [Online]. Available: <https://arxiv.org/abs/2309.10020>
- [5] A. Koenecke, A. S. G. Choi, K. X. Mei, H. Schellmann, and M. Sloane, "Careless whisper: Speech-to-text hallucination harms," in *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, 2024, pp. 1672–1681.
- [6] R. Frieske and B. E. Shi, "Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models," *arXiv preprint arXiv:2401.01572*, 2024.
- [7] H. Atwany, A. Waheed, R. Singh, M. Choudhury, and B. Raj, "Lost in transcription, found in distribution shift: Demystifying hallucination in speech foundation models," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 23 181–23 203. [Online]. Available: <https://aclanthology.org/2025.findings-acl.1190/>
- [8] T. Nishimura, S. Nakada, and M. Kondo, "On the audio hallucinations in large audio-video language models," *arXiv preprint arXiv:2401.09774*, 2024.
- [9] K. Sung-Bin, O. Hyun-Bin, J. Lee, A. Senocak, J. S. Chung, and T.-H. Oh, "Avhbench: A cross-modal hallucination benchmark for audio-visual large language models," in *The Thirteenth International Conference on Learning Representations*.
- [10] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shueb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research*, 2023.
- [11] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 4902–4912. [Online]. Available: <https://aclanthology.org/2020.acl-main.442/>
- [12] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, Y. Lv, Y. Wang, D. Guo, H. Wang, L. Ma, P. Zhang, X. Zhang, H. Hao, Z. Guo, B. Yang, B. Zhang, Z. Ma, X. Wei, S. Bai, K. Chen, X. Liu, P. Wang, M. Yang, D. Liu, X. Ren, B. Zheng, R. Men, F. Zhou, B. Yu, J. Yang, L. Yu, J. Zhou, and J. Lin, "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025. [Online]. Available: <https://arxiv.org/abs/2509.17765>
- [13] M. Abdin *et al.*, "Phi-4 technical report," Microsoft Research, Tech. Rep., 2024. [Online]. Available: <https://www.microsoft.com/en-us/research/wp-content/uploads/2024/12/P4TechReport.pdf>
- [14] F. Alam, A. E. Shahroor, M. A. Hasan, Z. S. Ali, H. H. Bhatti, M. B. Kmainasi, S. A. Chowdhury, B. Mousi, F. Dalvi, N. Durrani, and N. Milic-Frayling, "EverydayMMQA: A multilingual and multimodal framework for culturally grounded spoken visual qa," *arXiv preprint arXiv:2510.06371*, 2025. [Online]. Available: <https://arxiv.org/abs/2510.06371>
- [15] B. Mousi, F. Dalvi, S. Chowdhury, F. Alam, and N. Durrani, "Once correct, still wrong: Counterfactual hallucination in multilingual vision-language models," 2026. [Online]. Available: <https://arxiv.org/abs/2602.05437>
- [16] S. Nayak, K. Jain, R. Awal, S. Reddy, S. V. Steenkiste, L. A. Hendricks, K. Stanczak, and A. Agrawal, "Benchmarking vision language models for cultural understanding," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5769–5790. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.329/>
- [17] A. Vayani, D. Dissanayake, H. Watawana, N. Ahsan, N. Sasikumar, O. Thawakar, H. B. Ademtey, Y. Hmaiti, A. Kumar, K. Kukreja *et al.*, "All languages matter: Evaluating llms on culturally diverse 100 languages," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 565–19 575.
- [18] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *British Journal of Mathematical and Statistical Psychology*, vol. 61, no. 1, pp. 29–48, 2008.
- [19] F. TEAM, U. Abbas, M. S. Ahmad, M. Ahmad, A. Al-Homaid, A. Al-Nuaimi, E. Altinisik, E. Asgari, S. Chawla, S. Chowdhury, F. Dalvi, K. Darwish, N. Durrani, M. Elfeky, A. Elmagarmid, M. Eltabakh, A. Ersoy, M. Fatehkia, M. Q. Hashim, M. Hawasly, M. Hefeeda, M. Husaini, K. Isufaj, S.-G. Jung, H. Lachemat, J. K. Lucas, A. Mohamed, T. Mohiuddin, B. Mousi, H. Mubarak, A. Musleh, M. Ouzzani, A. Sadeghi, H. T. Sencar, M. Shinoy, O. Sinan, and Y. Zhang, "Fanar 2.0: Arabic generative ai stack," 2026. [Online]. Available: <https://arxiv.org/abs/2603.16397>
- [20] Z. S. Ali, H. H. Bhatti, R. N. Nandi, S. A. Chowdhury, and F. Alam, "Menaspeechbank: A reference voice bank with persona-conditioned multi-turn conversations for audiollms," *arXiv preprint arXiv:2602.07036*, 2026.
- [21] M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, N. Gupta, H. Hajishirzi, G. Ilharco, D. Khashabi, K. Lin, J. Liu, N. F. Liu, P. Mulcaire, Q. Ning, S. Singh, N. A. Smith, S. Subramanian, R. Tsarfaty, E. Wallace, A. Zhang, and B. Zhou, "Evaluating models' local decision boundaries via contrast sets," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1307–1323. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.117/>
- [22] L. Li, J. Qu, L. Song, Y. Zhou, Y. Qin, T. Yang, and Y. Zhao, "Treble counterfactual VLMs: A causal approach to hallucination," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 18 423–18 434. [Online]. Available: <https://aclanthology.org/2025.findings-emnlp.1000/>
- [23] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023.

- [24] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan, "Seed-bench: Benchmarking multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13 299–13 308.
- [25] C. Jiang, H. Jia, M. Dong, W. Ye, H. Xu, M. Yan, J. Zhang, and S. Zhang, "Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 525–534. [Online]. Available: <https://doi.org/10.1145/3664647.3680576>
- [26] F. Alwajih, G. Bhatia, and M. Abdul-Mageed, "Dallah: A dialect-aware multimodal large language model for Arabic," in *Proceedings of the Second Arabic Natural Language Processing Conference*, N. Habash, H. Bouamor, R. Eskander, N. Tomeh, I. Abu Farha, A. Abdelali, S. Touileb, I. Hamed, Y. Onaizan, B. Alhafni, W. Antoun, S. Khalifa, H. Haddad, I. Zitouni, B. AlKhamissi, R. Almatham, and K. Mrini, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 320–336. [Online]. Available: <https://aclanthology.org/2024.arabicnlp-1.27/>
- [27] F. Alwajih, A. El Mekki, S. M. Magdy, A. A. Elmadany, O. Nacar, E. M. B. Nagoudi, R. Abdel-Salam, H. Atwany, Y. Nafea, A. M. Yahya, R. Alhamouri, H. A. Alsayadi, H. Zayed, S. Shatnawi, S. Sibae, Y. Ech-chammakhy, W. Al-Dhabyani, M. M. Ali, I. Jarraya, A. O. El-Shangiti, A. Alraeesi, M. A. AL-Ghrawi, A. S. Al-Batati, E. Mohamed, N. T. Elgindi, M. Saeed, H. Atou, I. A. Yahia, A. Bouayad, M. Machrouh, A. Makouar, D. Alkawi, M. Mohamed, S. T. Abdelfadil, A. Z. Ounnoughene, A. Rouabhia, R. Assi, A. Sorkatti, M. C. Tourad, A. Koubaa, I. Berrada, M. Jarrar, S. Shehata, and M. Abdul-Mageed, "Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 32 871–32 894. [Online]. Available: <https://aclanthology.org/2025.acl-long.1579/>
- [28] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. de Chaumont Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rozenberg, T. Sainath, J. Schalkwyk, M. Sharifi, M. T. Ramanovich, M. Tagliasacchi, A. Tudor, M. Velimirović, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. Frank, "Audiopalm: A large language model that can speak and listen," 2023. [Online]. Available: <https://arxiv.org/abs/2306.12925>