# How Foundation Models behave for Arabic Image Captioning?

**Khaoula Dahimi**[1], **Hadda Cherroun**[1], **Amel Belabbaci**[1] **and Abdelhamid Haouhat**[1]
[1] Laboratoire d'Informatique et de Mathématiques LIM
*Amar Telidji University*, Laghouat, Algeria
{k.dahimi, hadda_cherroun, a.belabbaci, a.haouhat}@lagh-univ.dz

**Abstract**

Image captioning plays a crucial role in numerous applications, including educational systems. However, ensuring caption quality remains a significant challenge, particularly for morphologically rich, low-resource languages such as Arabic. We investigate an evaluation of Arabic image captioning using state-of-the-art multimodal foundation models. We systematically assess the performance of leading models—Gemini, Gemma, LLaMA, and Fanar. Our evaluation framework employs a diverse set of metrics spanning rule-based, learnable, visually-grounded, and LLM-based approaches to capture semantic accuracy, linguistic fluency, and hallucination detection. Experiments are conducted on two benchmark datasets: Flickr8k-Arabic and JEEM. Our findings reveal significant performance variations across models and evaluation metrics, highlighting the need for Arabic-specific optimization in multimodal architectures.

**Keywords:** Multi-Modal Foundation Models, Arabic Image Captioning, Image Caption Evaluation, Rule-based Metrics, Learnable Metrics, LLM-based Metrics, Hallucination Detection.

## 1. Introduction

The emergence of large-scale foundation models has fundamentally transformed the landscape of natural language processing, with multimodal architectures demonstrating unprecedented performance across diverse tasks (Qin et al., 2024). Vision-language models, including Gemini (Team et al., 2023), Gemma (Team et al., 2025b), and LLaMA (Dubey et al., 2024), have exhibited substantial capabilities in image-to-text generation tasks, particularly in automatic caption generation. Despite these advances, empirical evaluations have predominantly concentrated on high-resource languages, particularly English (Zhu et al., 2023), resulting in limited understanding of model behavior and performance characteristics when applied to morphologically rich, low-resource languages such as Arabic (Haouhat et al., 2025).

The automatic generation of textual descriptions from visual content constitutes a critical component in numerous domains, including assistive technologies for visually impaired users, multimedia content retrieval, and computer-assisted language learning systems (Dongare et al., 2024; Hasnine et al., 2019; Delassi et al., 2025). For Arabic—a language with over $400$ million native speakers and significant morphological complexity—the development of robust image captioning systems presents considerable challenges attributable to dataset scarcity (Team et al., 2025a), dialectal variation, and intricate grammatical structures. Within AI-assisted educational frameworks, automatically generated captions offer potential as reference standards for learner output evaluation and language proficiency assessment (Sarto et al., 2025), yet the reliability of such baselines remains underexplored.

This study systematically investigates the performance of a set of state-of-the-art multimodal foundation models—Gemini, Gemma, LLaMA, and Fanar in generating Arabic image captions using the Arabic Flickr8k dataset (ElJundi et al., 2020).

Our investigation employs established quantitative metrics (Papineni et al., 2001; Banerjee and Lavie, 2005; Hessel et al., 2022) to evaluate caption quality, semantic accuracy, and the prevalence of hallucination phenomena (Petryk et al., 2024; Rohrbach et al., 2019).

The investigation pursues two primary objectives: first, to assess whether model-generated captions exhibit sufficient quality and consistency to serve as reliable reference baselines for automated evaluation in AI-driven educational contexts; second, to examine whether existing quantitative metrics adequately capture the quality of generated descriptions, particularly in educational settings involving visual assignments.

The rest of this paper is organized as follows. Section 2 is dedicated to introducing some basic concepts on image captioning evaluation metrics. Section 3 reviews related work on Arabic image captioning. Section 4 presents our experimental methodology, including the selected metrics. Section 5 details the experimental setup, presents the results, and provides a comparative analysis of the targeted models. Finally, we conclude the paper with a summary of our findings and directions for future work.



*Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks*, pages 49–58

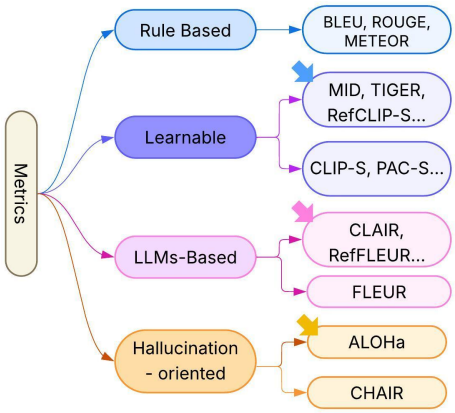



Figure 1: Taxonomy of Image Captioning Evaluation Metrics. Highlighted boxes indicate reference-based metrics that require ground-truth captions for comparison

## 2. Background

This section surveys the landscape of image captioning evaluation metrics and some existing Arabic image captioning datasets.

### 2.1. Image Captioning Evaluation Metrics

Image captioning evaluation metrics play a crucial role in assessing the performance of generative models. These metrics measure the quality of the generated text, ensuring that captions are fluent, accurate, and faithful to the image content while penalizing hallucinations or misleading information. Over time, many evaluation metrics have been proposed and refined alongside the evolution of large language models and the image captioning task itself. As illustrated in Figure 1, these metrics are organized into four main categories according to their computational foundations, following the taxonomy proposed by Sarto et al.'s comprehensive survey (Sarto et al., 2025).

#### Rule Based Metrics

These metrics were originally designed for evaluating machine translation systems. They rely on the principle that "the more the matches, the better the candidate translation." BLEU (Papineni et al., 2001) measures how many n-grams (words or sequences of words) in the generated text also appear in the reference text, using a weighted average approach. While BLEU focuses primarily on n-gram precision, METEOR (Banerjee and Lavie, 2005) extends the evaluation by also considering recall, stemming, synonyms, and word order, making it more linguistically informed. It operates at the word level, aligning the generated and reference texts to find the best possible matches, even when the words are not identical (e.g., *run* ↔ *running* or *big* ↔ *large*)

#### Learnable Metrics

Rule-based metrics depend on surface-level n-gram overlap and may fail when different words conveys the same meaning. Learnable metrics address this limitation by evaluating semantic similarity.

BERTScore (Zhang et al., 2020) metric evaluates the semantic similarity between the generated and reference text (Hanna and Bojar, 2021). They perform this by representing each token with contextual embeddings from a pretrained BERT model and comparing them using cosine similarity. Given reference tokens $r$ and predicted tokens $p$, using the recall, $R_{\text{BERT}}$, and precision $P_{\text{BERT}}$, BERTScore computes:

$$\text{BERTscore} F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad (1)$$

Where :

$$R_{\text{BERT}} = \frac{1}{|r|}\sum_{i \in r} \max_{j \in p} \vec{i}^{T}\vec{j}, \quad P_{\text{BERT}} = \frac{1}{|p|}\sum_{j \in p} \max_{i \in r} \vec{i}^{T}\vec{j}$$

However, these metrics still ignore the visual content and are sensitive to the style and quality of reference captions (Sarto et al., 2025). To address this gap, TIGEr (Jiang et al., 2019) and CLIPScore (Hessel et al., 2022), visually grounded metrics, were introduced. They assess how visually and semantically aligned a caption is with an image. Specifically, for a given image-caption pair. TIGEr (Jiang et al., 2019) computes the matching between textual tokens and image regions using a Text-to-Image Grounding approach. Similarly, CLIPScore (Hessel et al., 2022) uses the CLIP model to generate visual-text embeddings for each image-caption pair and evaluates their alignment via cosine similarity as follows:

$$\text{CLIP\_S}(c, v) = w \cdot \max\big(\cos(c, v), 0\big) \quad (2)$$

where: $c$ is the textual CLIP embedding, $v$ the visual CLIP embedding and $w = 2.5$ an empirical value (Hessel et al., 2022).

#### LLMs Based Metrics

With the emergence of powerful LLMs, metrics such as CLAIR (Chan et al., 2023) and FLEUR (Lee et al., 2024) incorporate them directly into the evaluation pipeline. These metrics convert caption evaluation into a text completion task solved using a structured prompt.

CLAIR (Chan et al., 2023) evaluates how well a candidate caption describes an image by comparing it to reference captions. It outputs both a numerical score and a brief explanation.

FLEUR (Lee et al., 2024) extends this by incorporating the actual image into the evaluation prompt, improving accuracy. It also introduces score smoothing, which uses the token probabilities of each digit generated by the model to produce less noisy and more interpretable scores.

### Hallucination Oriented Metrics

Unlike the aforementioned metrics that focus on evaluating caption quality, CHAIR (Rohrbach et al., 2019) and ALOHa (Petryk et al., 2024) are designed specifically to measure object hallucination in generated captions. Hallucination occurs when a caption includes information or objects not actually present in the image (Sarto et al., 2025).

CHAIR (Rohrbach et al., 2019) detects hallucinations by comparing objects mentioned in the generated caption with those found in the reference captions and segmentation labels of the image, allowing it to determine accurately whether an object is truly present. However, CHAIR is limited to the fixed object vocabulary of the MS COCO dataset.

ALOHa (Petryk et al., 2024) addresses this limitation by moving to an open-vocabulary setting. By using LLMs to extract object mentions and semantic similarity to match them. ALOHa provides a more flexible and generalizable way to detect hallucinations.

## 2.2. Image Captioning Datasets

Arabic image captioning datasets can be categorized into two primary groups:

**Translation-Based Datasets** comprises datasets derived by translating established English-language benchmarks into Arabic. These include Arabic Flickr8k (ElJundi et al., 2020) and Multilingual COCO (JP, 2025), which adapt their respective originals with Modern Standard Arabic (MSA) captions, and the Arabic Image Captioning (Hennara et al., 2025) dataset, which incorporates translated samples from UCSC-VLAA[1] and Recap-DataComp-1B(Li et al., 2024).

**Natively Collected Datasets** consists of resources specifically curated by native Arabic speakers to ensure linguistic authenticity and cultural relevance. Notable examples are JEEM (Kadaoui et al., 2025), FutureBeeAI's Arabic Image Caption (FutureBee AI, 2025) Dataset, and the ImageEval 2025 dataset from the shared task (Bashiti et al., 2025).

[1] UCSC-VLAA Lab.

## 3. Related Work

Many studies have addressed the challenges of Arabic Image Captioning (AIC) using Deep Learning techniques, typically employing CNN-based encoders and LSTM-based decoders. The authors in (ElJundi et al., 2020) use an LSTM sequence model that outperforms traditional English-caption translation approaches. Meanwhile, (Afyouni et al., 2021) introduce a novel architecture that combines object detection with attention-based caption generation.

On the other hand, (Hejazi and Shaalan, 2021) present a comparative study analyzing factors that influence caption generation—such as preprocessing choices, Deep Learning techniques, and feature extraction methods—and offer practical recommendations. Additionally, (Al-Malki and Al-Aama, 2023) suggest that classifying image attributes before feeding them into the decoder can further improve Arabic caption quality.

Only a few studies incorporate transformer-based models. Specifically, (Emami et al., 2022) employ a pre-trained transformer to model relationships between object tags extracted by a CNN encoder. Similarly, (Alsayed et al., 2023) explore the impact of different Arabic text-preprocessing methods alongside various image recognition models using a transformer-based approach.

Authors in (Elchafei and Fashwan, 2025) introduce an interpretable Arabic image captioning approach that retrieves visual concepts using multilingual CLIP models and feeds them as prompts to vision-language decoders. This two-stage design improves semantic grounding and leads to better caption quality than traditional end-to-end models.

Recent efforts have aimed to advance Arabic image captioning through shared tasks and community-driven initiatives. ImageEval'2025 (Bashiti et al., 2025) introduced the first dedicated shared task for Arabic image captioning, combining collaborative dataset creation with competitive model evaluation to address challenges such as cultural specificity, morphological complexity, and dataset scarcity.

## 4. Methodology

To systematically examine how multimodal foundation models handle the Arabic language in the context of image caption generation, we have drawn out a comprehensive evaluation pipeline fed by images from Arabic datasets, designed to assess model performance under different experimental

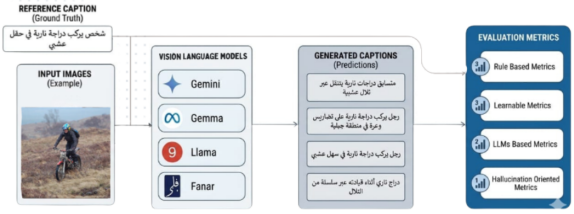



Figure 2: Overview of the comprehensive evaluation pipeline for Arabic image captioning.

conditions and using diverse evaluation criteria and metrics.

The proposed pipeline is illustrated in Figure 2. It comprises four sequential stages: dataset selection and preparation, model-based caption generation, preprocessing, and multi-metric evaluation.

### Dataset Selection

A curated set of image-caption pairs from different Arabic image datasets serves as the evaluation corpus, providing sufficient visual and linguistic diversity while maintaining computational feasibility.

Each image is independently processed through multiple models to generate corresponding Arabic captions, which are subsequently compared against human-generated reference captions. To generate captions of similar length, we provide the model with an explicit length prompt. By controlling the length, we can fairly evaluate the models' ability to produce descriptive captions across datasets with varying caption styles and levels of detail.

### Model Selection

In this investigation a panoply of models is targeted, intentionally chosen to represent different architectural paradigms and training methodologies. The selected models are categorized into two primary groups based on their design focus and characteristics of training data. The first category comprises *state-of-the-art (SOTA) multilingual multimodal models*, including *Gemini*, *Gemma*, and *LLaMA*, which have demonstrated exceptional performance across multiple languages and vision-language tasks. These models are predominantly trained on large-scale multilingual corpora that include Arabic among other languages, but without specific optimization for Arabic linguistic features. The second category encompasses *Arabic-specific multimodal models*, which have been either pre-trained or fine-tuned specifically on Arabic datasets to capture the morphological richness, dialectal variations, and syntactic structures inherent to the Arabic language. This categorization enables comparative analysis between general-purpose multilingual models and language-specific architectures, providing insights into whether specialized Arabic training yields superior performance for Arabic caption generation.

### Caption Preprocessing

Before evaluation, all model-generated captions undergo a systematic preprocessing phase to ensure consistency and comparability across models. This preprocessing step is critical to ensure that the subsequent evaluation focuses purely on the semantic and linguistic quality of the generated content rather than superficial formatting differences. The caption are normalized. This process involved removing unnecessary elements and eliminating any non-Arabic words to ensure consistency in the analysis.

### Multi-Metric Evaluation Framework

The resulted image captions are then evaluated using a comprehensive set of metrics designed to capture different dimensions of caption quality. These metrics are multi-purpose, each addressing specific aspects of the caption generation task. We employ both *reference-based metrics* that compare generated captions against human-written references and *reference-free metrics* that assess caption quality based on intrinsic properties.

This multi-purpose evaluation strategy provides a holistic assessment of model performance, enabling identification of specific strengths and weaknesses in Arabic caption generation across different model categories and experimental conditions.

## 5. Experiments and Results

To achieve our research objectives, we designed a comprehensive series of experiments with two main goals in mind. The first goal is to identify which foundation models produce the most accurate and high quality Arabic image captions based on our evaluation metrics. We systematically compare outputs from SOTA multilingual models and Arabic-specific architectures, testing them under different conditions—both with and without prompting. The second goal addresses a critical question: which metrics are best suited for evaluating Arabic linguistic quality in the context of language learning? Our experimental design carefully examines two types of metrics: those that rely on human-written reference captions for comparison, and those that can assess caption quality independently. This distinction matters greatly in real-world educational settings, where reference captions may not always be available.

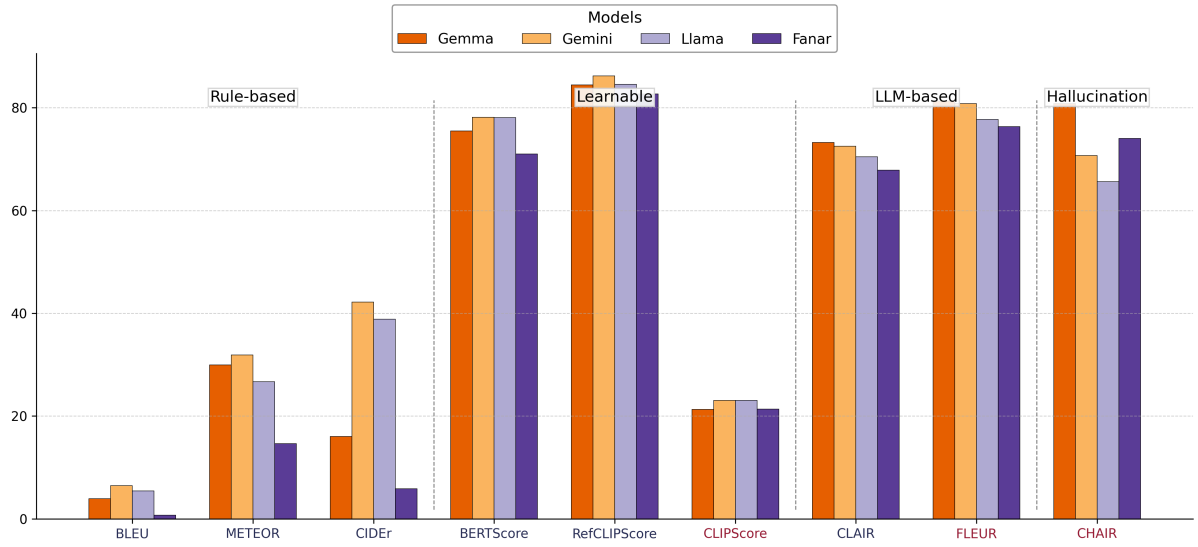



Figure 3: Performance Assessment of Multimodal Models across Nine Metrics on Arabic Flickr8K dataset

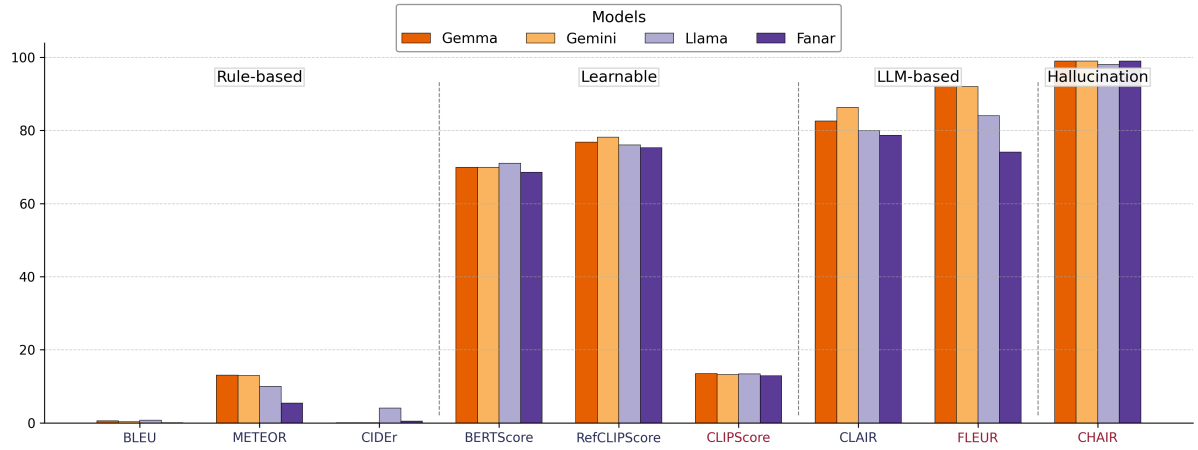



Figure 4: Performance Assessment of Multimodal Models across Nine Metrics on JEEM dataset

We start by describing the experiment testbed namely the used dataset, foundation models and metrics. All our experimentation scripts are publicly accessible on GitHub.

### 5.1. Datasets

We utilized two distinct Arabic image captioning datasets: the Arabic version of the well-known Flickr8k dataset, and the natively collected JEEM dataset, which covers a broader range of domains and contains longer, more detailed descriptions that convey richer information about the images. Both resources provide standardized collections of images paired with descriptive captions, enabling consistent evaluation of vision-language models. A detailed summary of these datasets is presented in Table 1. It shows that JEEM has a large domain covarage. Its captions are more long, so that they give more detailed on the image.

### 5.2. Targeted Vision Language Models

Concerning the SOTA models, we have used:

**Gemini-2.5-Flash**[2]: Developed by Google AI, it is a generative model designed for high-quality text generation.

**Gemma-3-27B-IT**[3] :It is an advanced generative model introduced by Google AI (Team et al., 2025b). It is widely recognized for its instruction-tuned capabilities in text generation.

**LLaMA-4-Scout-17B-16E-Instruct**[4] : Developed by Meta AI, it is a state-of-the-art generative model commonly used for various language generation tasks.

Concerning the Arabic multimodal models and due to their scarcity, we have narrowed the study

[2]API version of the model Gemini Model
[3]Gemma Model : the API version of the model
[4]LlaMa Model: : the API version of the model

Table 1: A Detailed Overview of Arabic Vision-Language Datasets Utilized in This Study

| Dataset | Size | Captions per Image | Avg. Words | Resolution | Domain | Experimental Subset |
|---|---|---|---|---|---|---|
| Arabic Flickr8k | 8,092 | 3 references | 9–12 | 640×480 | Everyday objects, people, animals, scenes | 1,173 |
| JEEM | 2,196 | 1 reference | 44–63 | Variable | Places, events, arts, nature, education, transport, food, trade, technology, characters, games | 1,704 |

to Fanar Vision version **Fanar-Oryx-IVU-1**[5]. It is developed by Fanar.ai, and represents an Arabic multimodal model designed for visual and linguistic understanding, including image captioning.

## 5.3. Results and Discussion

The performances of model-generated Arabic image captions are reported across all evaluation metrics in Figures 3 and 4.

The evaluation reveals clear differences among multimodal LLMs across lexical, semantic, and hallucination-sensitive metrics.

The overall performance of the models is quite similar, with Gemini & Gemma achieving the highest scores across both datasets. Additionally, Fanar model performs poorly compared to the others, which is unexpected for an Arabic model, this may be due to the diversity of the selected dataset, and the Fanar model may not have been trained on it.

Regarding the **rule-based metrics**, the models show low scores for the Arabic Flickr8k dataset, suggesting a limited lexical overlap between the generated captions and the reference texts. There are also challenges in capturing semantic adequacy and linguistic variation, especially for morphologically rich languages such as Arabic. The scores are even worse for the JEEM dataset, as expected, due to its longer descriptions.

For the **learnable metrics**, BERTScore and RefCLIPScore evaluate the similarity between generated captions and reference texts. The models achieve high scores on both datasets, indicating strong alignment with the references. This suggests that the generated captions effectively paraphrase or semantically match the human-provided captions.

In contrast, CLIPScore measures the similarity between the image and the generated caption within CLIP's joint embedding space. The models obtain relatively low CLIPScores on both datasets. This may reflect limitations in CLIP's representations, such as sensitivity to non–internet-style phrasing, domain mismatch, or reduced robustness to long and complex sentences (Zur et al., 2024; Chen et al., 2025).In summary, although the generated captions appear strong according to BERTScore and RefCLIPScore, a low CLIPScore does not necessarily indicate poor accessibility quality.

For the **LLM-based metrics**, the models achieve high scores on CLAIR and FLEUR, demonstrating their ability to generate descriptive, visually grounded captions that closely reflect the image content. Notably, Gemma and Gemini consistently outperform the other models across both datasets.

### Metric Reliability Analysis

We have observed that all models exhibit relatively poor performance on the CHAIR metric (where lower values indicate better performance). A deep manual analysis of over $100$ generated descriptions revealed consistently captions' good quality, suggesting a discrepancy between the metric's hallucination indicators and the actual output quality, as illustrated in Table 2.

The misleading CHAIR scores may explain the seemingly poor performance of the models, as the metric is built on a limited object vocabulary derived from the MS COCO dataset, as illustrated in the first image of Table 2. In addition, CHAIR relies on strict object matching and does not account for deeper caption understanding. Consequently, it may classify contextually inferred objects—rather than physically visible ones—as hallucinations, as shown in the second image in Table 2.

Poor CHAIR performance should therefore be attributed to metric limitations rather than model failure.

### Linguistic and Semantic Error Analysis of Generated Captions

Table 3 illustrates representative examples of linguistic errors observed in model-generated Arabic captions. Two primary error types are identified. The first involves object-level mismatches combined with grammatically incorrect word usage, as seen in the first example where the generated caption incorrectly refers to horses and uses a morphologically ill-formed term. The second error type concerns grammatical agreement viola-

[5] Fanar Model : the API version of the model

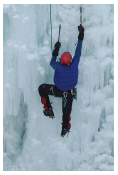
| Image | Generated Caption | CHAIR Score | Observation |
|---|---|---|---|
|  | رجل يتسلق الثلج بواسطة معدات التسلق، مرتديا قفازات و أحذيه الثلوج و قبعه حمراء<br>En: A man climbs the snow using climbing equipment, wearing gloves, snow boots, and a red hat. | 1 | These objects are visually present in the image. |
|  | طفله صغيره تتسلق شبكه حبال حمراء في ملعب وتنظر الي الكاميرا<br>En: A little girl climbs a red rope net in a playground and looks at the camera. | 1 | The object is not visually detected, although it is implicitly inferred from the photographic context. |

Table 2: Illustration of false hallucination cases for generated captions.

| Image | Reference | Incorrect caption | CLAIR Score | Observation |
|---|---|---|---|---|
|  | ثلاثة كلاب من أحجام مختلفة في فناء خارجي<br>En: Three dogs of different sizes in an outdoor courtyard. | خيول صغيرة نشطة و ممرحلة في مساحة خضراء<br>En: Small, active and —— horses in a green open space. | 0.8 | The caption contains an object-level mismatch and also includes a grammatically incorrect word. |
|  | كلبان بنيان يركضان جنبا إلى جنب<br>En: Two brown dogs are running side by side. | كلبان يركضون في الحديقة<br>En: Two dogs are running in the park. | 0.8 | The caption contains a grammatical agreement error related to the Arabic dual form, a morphological category absent from English. |
|  | صبيان يركضان و يلعبان بالماء<br>En: Two boys are running and playing with water. | طفل يستمتع باللعب بالماء في حديقة<br>En: A child is enjoying playing with water in a garden. | 0.9 | The caption incorrectly represents the number of subjects. |
|  | طبق من الكنافة النابلسية الناعمة التي تقدم علي شكل مربعات بحسب الوزن، وهي من أشهر الحلويات في الأردن، وتتميز بوجهها المحمر المغطى بطبقه من الفستق الحلبي المفروم للزينة، يظهر بعض القطر الذي يضفي عليها تكهة حلوة غنية إضافة إلى طبقة الجبنة الذائبة الموجودة تحت العجينة. | طبق من الحلوى الشرقية، تحديدا البقلاوة، مقدمة على طبق أبيض دائري. يقع الطبق على سطح طاولة ذات ملمس مميز، ويبدو أن الإضاءة في المكان خافتة ودافئة. البقلاوة قطعة مربعة الشكل، ذات لون ذهبي بني، تبدو مقرمشة ومليئة بالطبقات الرقيقة من العجين، مزينة بكمية صغيرة من الفستق الحلبي المفروم. | 0.8 | The caption contains an error in object identification, as it incorrectly labels the dessert. |

Table 3: Examples of grammatical and semantic errors in automatically generated captions that go unpenalized by evaluation metrics.

tions, where the generated caption fails to produce the correct verb form with respect to number, as demonstrated in the second example. Despite these linguistic inaccuracies, both captions receive a relatively high CLAIR score of $0.8$, indicating that the metric assigns favorable scores without penalizing such errors. This reveals a critical limitation of existing evaluation metrics, which fail to capture fine-grained linguistic quality, particularly morphological and grammatical agreement violations inherent to Arabic. Such overestimation poses a serious concern in educational contexts, where linguistically flawed captions could be incorrectly validated as correct references, potentially misleading learners and undermining the reliability of automated evaluation systems.

#### 5.3.1. Comparative Image Captioning Performance Across Arabic and English

For comparative purposes, we evaluated both Gemma and LLaMA models on the English and Arabic versions of the Flickr8k dataset using $400$ samples. The results are reported in Figure 5.

Overall, both models consistently achieve higher scores in English than in Arabic across nearly all evaluation metrics. This indicates that the models generate more accurate and reliable captions in English compared to Arabic.

For the LLM-based metrics, the performance is relatively comparable across languages, with higher CLAIR scores in Arabic and higher FLEUR scores in English. This suggests that the Arabic-generated captions are more closely aligned with the reference texts, while the English captions

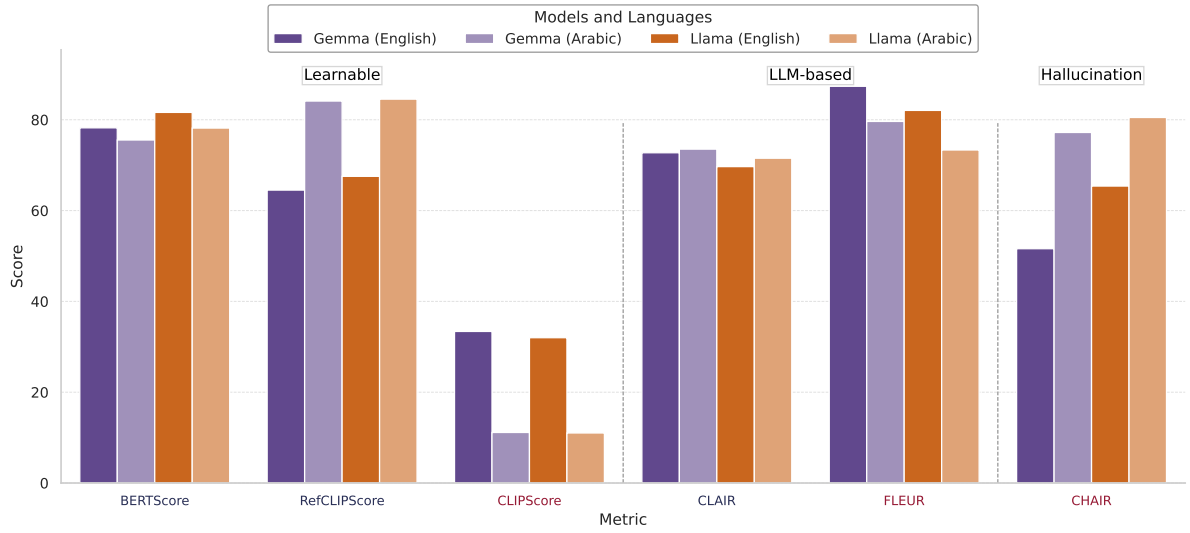



Figure 5: Performances of Gemma and LLaMA, across six metrics, on $400$ samples from both Flickr8k versions : English and Arabic.

demonstrate stronger visual grounding.

A notable difference appears in the CHAIR scores across languages, with English consistently achieving better scores. This suggests that both models are more prone to hallucinations when generating captions in Arabic than in English. However, these results should be interpreted with caution. CHAIR relies on object lists derived from the MS-COCO dataset, which is predominantly English-centric and lacks adequate Arabic lexical coverage. As discussed earlier, this renders CHAIR a poorly suited metric for Arabic caption evaluation, and its scores in this language cannot be considered fully reliable.

## 6. Conclusion

We evaluated Multimodal Large Language Models for Arabic image captioning using the Arabic Flickr8k and JEEM datasets and nine metrics. Gemma and Gemini models achieved the strongest performance across standard metrics, demonstrating robustness in this domain. However, while all models effectively captured core semantic content, they consistently struggled with exact lexical alignment to reference captions — a limitation largely attributable to Arabic's rich morphology and syntactic complexity.

Moreover, the tendency of the models to generate overly detailed or speculative captions highlights the persistent challenge of hallucination in generative MLLMs. While the evaluation is confined to two datasets, the results provide a significant baseline for evaluating Arabic vision–language models and offer valuable insights into their current strengths and weaknesses.

Future research will explore prompt engineering strategies grounded in the linguistic competencies framework — targeting morphological, syntactic, and lexical dimensions of Arabic — alongside metrics that better reflect true caption quality. We also plan to address hallucination and integrate linguistically-aware captioning systems into educational tools to foster richer, more accessible learning experiences.

## 7. Acknowledgements

The authors gratefully acknowledge FANAR.ai for granting access to their Vision Model API. This work is supported by the Directorate-General for Scientific Research and Technological Development (DGRSDT) - Algeria, and performed under the PRFU Project: C00L07N030120220002.

Imad Afyouni, Imtinan Azhar, and Ashraf Elnagar. 2021. Aracap: A hybrid deep learning architecture for arabic image captioning. *Procedia Computer Science*, 189:382–389.

Rasha Saleh Al-Malki and Arwa Yousuf Al-Aama. 2023. Arabic captioning for images of clothing using deep learning. *Sensors*, 23(8):3783.

Ashwaq Alsayed, Thamir M. Qadah, and Muhammad Arif. 2023. A performance analysis of transformer-based deep learning models for arabic image captioning. *Journal of King Saud University - Computer and Information Sciences*, 35(9):101750.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ahlam Bashiti, Alaa Aljabari, Hadi Khaled Hamoud, Md. Rafiul Biswas, Bilal Mohammed Shalash, Mustafa Jarrar, Fadi Zaraket, George Mikros, Ehsaneddin Asgari, and Wajdi Zaghouani. 2025. Imageeval 2025: The first arabic image captioning shared task. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, page 376–389, Suzhou, China. Association for Computational Linguistics.

David M. Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell, and John Canny. 2023. Clair: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 13638–13646, Singapore. Association for Computational Linguistics.

Xiaofu Chen, Israfel Salazar, and Yova Kementchedjhieva. 2025. Specs: Specificity-enhanced clip-score for long image caption evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9406–9418.

Khaled Bachir Delassi, Lakhdar Zeggane, Hadda Cherroun, Abdelhamid Haouhat, and Kaoutar Bouzouad. 2025. Vqa support to arabic language learning educational tool.

Yashwant Dongare, Bhalchandra M. Hardas, Rashmita Srinivasan, Vidula Meshram, Mithun G. Aush, and Atul Kulkarni. 2024. Deep neural networks for automated image captioning to improve accessibility for visually impaired users. *International Journal of Intelligent Systems and Applications in Engineering*, 12(2s):267–281.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of models.

Passant Elchafei and Amany Fashwan. 2025. Multimodal arabic captioning with interpretable visual concept integration. (arXiv:2510.03295). ArXiv:2510.03295 [cs].

Obeida ElJundi, Mohamad Dhaybi, Kotaiba Mokadam, Hazem Hajj, and Daniel Asmar. 2020. Resources and end-to-end neural network models for arabic image captioning:. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, page 233–241, Valletta, Malta. SCITEPRESS - Science and Technology Publications.

Jonathan Emami, Pierre Nugues, Ashraf Elnagar, and Imad Afyouni. 2022. Arabic image captioning using pre-training of deep bidirectional transformers. In *Proceedings of the 15th International Conference on Natural Language Generation*, page 40–51, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

FutureBee AI. 2025. Arabic image caption dataset. https://www.futurebeeai.com/dataset/multi-modal-dataset/arabic-image-caption-dataset. Accessed: February 7, 2026.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517.

Abdelhamid Haouhat, Slimane Bellaouar, Attia Nehar, Hadda Cherroun, and Ahmed Abdelali. 2025. Arabic multimodal machine learning: Datasets, applications, approaches, and challenges. *arXiv e-prints*, pages arXiv–2508.

Mohammad Nehal Hasnine, Brendan Flanagan, Gokhan Akcapinar, Hiroaki Ogata, Kousuke Mouri, and Noriko Uosaki. 2019. *Vocabulary Learning Support System Based on Automatic Image Captioning Technology*, volume 11587 of *Lecture Notes in Computer Science*, page 346–358. Springer International Publishing, Cham.

Hani Hejazi and Khaled Shaalan. 2021. Deep learning for arabic image captioning: A comparative study of main factors and preprocessing recommendations. *International Journal of Advanced Computer Science and Applications*, 12(11).

Khalil Hennara, Muhammad Hreden, Mohamed Motaism Hamed, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Mutarjim: Advancing bidirectional arabic-english translation with a small language model.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. (arXiv:2104.08718). ArXiv:2104.08718 [cs].

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. Tiger: Text-to-image grounding for image caption evaluation. (arXiv:1909.02050). ArXiv:1909.02050 [cs].

Romrawin JP. 2025. Multilingual coco dataset. https://huggingface.co/datasets/romrawinjp/multilingual-coco. Accessed: February 7, 2026.

Karima Kadaoui, Hanin Atwany, Hamdan Al-Ali, Abdelrahman Mohamed, Ali Mekky, Sergei Tilga, Natalia Fedorova, Ekaterina Artemova, Hanan Aldarmaki, and Yova Kementchedjhieva. 2025. Jeem: Vision-language understanding in four arabic dialects. (arXiv:2503.21910). ArXiv:2503.21910 [cs].

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. (arXiv:2406.06004). ArXiv:2406.06004 [cs].

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. 2024. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. 2024. Aloha: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, page 342–357, Mexico City, Mexico. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. (arXiv:1809.02156). ArXiv:1809.02156 [cs].

Sara Sarto, Marcella Cornia, Rita Cucchiara, et al. 2025. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025a. Fanar: An arabic-centric multimodal generative ai platform.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025b. Gemma 3 Technical Report.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. (arXiv:1904.09675). ArXiv:1904.09675 [cs].

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Amir Zur, Elisa Kreiss, Karel D'Oosterlinck, Christopher Potts, and Atticus Geiger. 2024. Updating clip to prefer descriptions over captions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20178–20187.