International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Comparative Evaluation of GPT-4o, Gemini, Llama, And Grok On Remote Sensing Imagery

**Abdulmumini Imam Ibrahim[1], Abdullahi Muhammad Auwal[2], Jidda Harun Abba[3]**

[1]Department of Computer Science, Ramat Polytechnic, Maiduguri
[2]Department of Computer Science, Ramat Polytechnic, Maiduguri
[3]Department of Computer Science and Engineering, Integral University, Lucknow

**Abstract**

This study presents an in-depth comparative evaluation of four Multimodal Large Language Models (MLLMs) GPT-4o, Gemini 2.5 Pro, Llama 4, and Grok 3 on satellite image captioning and classification using the Remote Sensing Image Captioning Dataset (RSICD). Using structured prompts and expert human judgment, we assessed each model across the following qualitative metrics: accuracy, relevance, understanding depth, and classification precision. Our findings show that MLLMs, while not replacements for specialized remote sensing tools, offer substantial support as analytical partners and produce context-aware interpretations and reliable classifications. Distinct performance profiles emerged, and we outlined critical directions for future research in quantitative benchmarking, advanced prompt engineering, and hybrid model architectures.

**Keywords: Multimodal Large Language Models, Satellite Imagery, Remote Sensing, Image Captioning, Image Classification**

## 1 INTRODUCTION

Satellite imagery has emerged as an essential tool across a range of disciplines, providing critical data for understanding and managing our planet[1]. These images offer a unique perspective for observing large-scale phenomena and tracking changes over time. Captured by sensors orbiting the Earth, key applications include land cover classification, which aids in distinguishing among various surface types, such as urban areas, vegetation, water bodies, and bare soil; environmental monitoring, such as tracking deforestation, assessing crop health, and monitoring water resources; disaster management, involving rapid damage assessment following events such as floods, earthquakes, or wildfires, and monitoring potential hazards, such as landslides; and urban planning, which benefits from insights into city growth, infrastructure development, and land use patterns[1].

Due to the complexity and substantial volume of satellite data, advanced analytical techniques are essential. Typically, methods such as image segmentation, feature extraction, and classification using algorithms like Support Vector Machines (SVMs) or early Convolutional Neural Networks (CNNs) are employed. However, these conventional approaches often encounter several limitations. They heavily depend on visual pixel information, making it challenging to capture the entire semantic context or subtle relationships within an image. Additionally, satellite images frequently exhibit significant variability in

resolution, scale, illumination, and atmospheric conditions. Consequently, traditional, purely pixel-based methods struggle to generalize effectively across more diverse datasets and increasingly complex scenes[1].

The domain of Artificial Intelligence (AI) has recently experienced a notable paradigm shift with the emergence of Multimodal AI. In contrast to traditional AI systems, which typically process a singular type of data (e.g., text or images), multimodal systems are engineered to comprehend, process, and integrate information from multiple data types or modalities concurrently. This integration facilitates a more comprehensive and contextually enriched understanding that closely emulates human perception.

Within this domain, Multimodal Large Language Models (MLLMs) have emerged as a particularly powerful class of models[2] [3]. MLLMs leverage the advanced reasoning and language understanding capabilities of Large Language Models (LLMs) like GPT or LLaMA, extending them to interpret visual or other non-textual inputs. By connecting powerful LLMs (acting as the 'brain') with modality encoders (like visual encoders based on ViT or CLIP architectures), MLLMs can perform tasks that require understanding the interplay between different data types. Key capabilities demonstrated by MLLMs include visual instruction following (performing tasks based on visual input and textual commands), sophisticated reasoning about image content, and generating detailed, context-aware descriptions or analyses[2] [4]. Prominent examples like OpenAI's GPT-4V, Google's Gemini, and open-source initiatives like LLaVA [5][6] showcase the rapidly advancing capabilities in this field.

Previous work in medical imaging and hydrology demonstrates the promise of MLLMs for complex reasoning and transferability[6][5]. However, remote sensing-specific applications require systematic comparison, as traditional methods face limitations in multi-modal data integration and semantic abstraction. This research addresses the gap by benchmarking leading MLLMs on standardized tasks to elucidate performance boundaries and application utility.

## 2 LITERATURE REVIEW

### 2.1 Traditional Approaches and their Shortcomings

Classic remote sensing methods rely on vision-based models (SVM, CNN) that extract pixel-level information for land cover classification and object detection but struggle to generalize owing to variability in image quality and contextual ambiguity[1]. These approaches lack multi-modal reasoning and fail to interactively synthesize visual data with external knowledge. MLLMs offer potential solutions to these limitations by integrating visual and textual information for more robust and context-aware analysis. Their ability to understand and reason across multiple modalities could enhance the interpretation of complex remote sensing data. This study aims to systematically evaluate the performance of MLLMs in remote sensing applications, comparing them to traditional methods and assessing their potential to overcome existing challenges in the field.

### 2.2 Rise of Multimodal Large Language Models

MLLMs overcome these limitations by integrating visual and textual encoders, enabling sophisticated visual instruction following, image captioning, and question answering [2] [7]. Models such as GPT-4o and Gemini 2.5 Pro support zero-shot/few-shot learning and exhibit advanced spatial reasoning. The ability to contextually process images with textual grounding shifts the paradigm from shallow feature matching to cognitive-level analysis[3]. Recent advancements in MLLMs have shown promising results in remote sensing applications, particularly in addressing complex tasks that require both visual and contextual understanding. These models can leverage their vast knowledge bases to interpret satellite

imagery, aerial photographs, and other geospatial data with greater accuracy and flexibility than traditional methods. By combining visual analysis with natural language processing capabilities, MLLMs offer the potential to revolutionize land cover classification, change detection, and environmental monitoring in ways that were previously unattainable.

### 2.3 The Domain Gap and Specialized Solutions

The direct application of MLLMs to remote sensing faces challenges owing to the "domain gap" differences in semantics, scale, and context compared to natural images [8]. Recent specialized frameworks, such as EarthGPT, provide domain adaptation through visual-enhanced perception and cross-modal mutual comprehension for multisensor imagery [9]. Semantic-augmented Multi-level Alignment introduces retrieval-based modules to enrich visual features with knowledge-based cues, bridging the gap for remote sensing-specific multimodal reasoning. These specialized solutions aim to overcome the unique challenges posed by remote sensing data, enabling MLLMs to better interpret and analyze complex geospatial information. By incorporating domain-specific knowledge and adapting to the nuances of satellite and aerial imagery, these frameworks enhance the accuracy and reliability of MLLM-based remote sensing applications. As research in this area progresses, we can expect further refinements and innovations that will continue to narrow the domain gap and unlock the full potential of MLLMs in Earth observation and environmental monitoring.

## 3 METHODOLOGY

### 3.1 Dataset Preparation

The RSICD dataset was chosen due to its extensive class diversity and high-quality annotations, encompassing over 10,000 satellite images with human-generated captions across 30 categories, such as airport, bare land, baseball field, beach, and bridge. A stratified subset of 300 images was selected to ensure a balanced representation of urban, rural, and natural land cover types. All images were resized to 512×512 pixels and normalized to align with the model encoder requirements, with low-quality or ambiguous samples excluded based on standardized criteria. Example Class Distribution: Airport (800), Bare Land (700), Baseball Field (600), Beach (650), Bridge (500).

### 3.2 Model Selection

The selected Multimodal Large Language Models (MLLMs) were chosen based on their multimodal architecture and documented performance: GPT-4o, Gemini 2.5 Pro, Llama 4, and Grok 3, developed by OpenAI, Google DeepMind, Meta, and xAI, respectively. GPT-4o is noted for its context-rich and comprehensive analytical capabilities. Gemini 2.5 Pro is recognized for its concise accuracy. LLaMA 4 effectively integrates spatial-semantic cues. Grok 3 is characterized by its creative reasoning, albeit sometimes at the expense of precision. In instances where APIs were inaccessible, public demonstrations and research previews were utilized to simulate the outputs. These models showcase distinct strengths in natural language processing, each catering to different user needs and preferences. While Gemini 2.5 Pro excels in providing succinct and accurate responses, LLaMA 4's spatial-semantic integration offers a unique approach to contextual understanding. Grok 3's creative reasoning capabilities present an interesting trade-off between innovative thinking and factual precision, highlighting the ongoing challenge in balancing creativity and accuracy in AI language models.

### 3.3 Experimental Design and Prompt Strategy

Each evaluation instance involved pairing an RSICD image with a structured prompt compatible with zero-shot learning, aimed at either captioning or classification tasks. Example prompts included: "Describe

this satellite image in detail, mentioning key features and land cover types," and "Based on this image, classify the scene into one of the predefined RSICD categories." No prior context or example pairs were provided, ensuring that the model responses demonstrated genuine zero-shot capabilities. The output metadata included execution time and token count when available.The evaluation process was designed to assess the model's ability to interpret and describe satellite imagery without prior training on the specific dataset. By utilizing structured prompts, the study aimed to elicit detailed and relevant responses from the AI model, testing its capacity for zero-shot learning in the domain of remote sensing. The approach allowed for a comprehensive assessment of the model's performance in both descriptive and classificatory tasks, providing insights into its generalization capabilities across different types of satellite imagery.

### 3.4 Evaluation Metrics and Reviewer Procedure

A mixed-method approach was employed, integrating expert qualitative assessment and scoring across four dimensions. Caption Accuracy was evaluated on a three-point scale (Correct, Partially Correct, Incorrect), measuring correctness against the ground truth and image content. Relevance was assessed based on the focus on dominant features as opposed to generic content. Understanding Depth involved evaluating spatial, semantic, and contextual richness, such as the proximity of buildings or scene layouts. Classification Precision determined the accuracy of category assignment and the clarity of reasoning. Three annotators with expertise in remote sensing independently scored the outputs, resolving discrepancies through consensus to mitigate subjective bias.The researchers evaluated the quality of descriptions by looking at how well they captured important details rather than general information. They also checked if the descriptions showed a good understanding of the scene's layout and context. Additionally, they assessed how accurately the descriptions categorized what was seen. To ensure fairness, three experts in satellite imagery independently reviewed and scored the descriptions, discussing any differences to reach agreement.

**Sample Output Example for 'Bridge' Scene:**

GPT-4o: “The image displays a large suspension bridge crossing a wide river. Vehicles transit the structure, with adjacent roads leading to urban areas. Vegetation patches flank the waterway.”

Gemini 2.5 Pro: “A bridge spanning a river, with roads and surrounding city infrastructure. Vehicle activity visible."

Llama 4: “Bridge over river, with roads, buildings urban land cover observed.”

Grok 3: “Busy bridge stretching across water, vibrant traffic, reflected urban lights.”

### 3.5 Limitations

API rate limitations, ongoing model updates, and the variability of outputs from proprietary platforms pose challenges to reproducibility. While human scoring is inherently subjective, the establishment of consensus and guidelines serves to mitigate bias. The study's emphasis on captioning and classification excludes consideration of other tasks, such as object detection and change analysis.

## 4 RESULTS AND DISCUSSION

### 4.1 Overall Performance and Comparative Analysis

Each model displayed distinct strengths and weaknesses. GPT-4o delivered highly detailed, context-rich captions; Gemini 2.5 Pro prioritized accurate, succinct responses; Llama 4 balanced multimodal integration; Grok 3 produced creative outputs but sometimes lacked precision.

**Table 1. Comparative Qualitative Performance**

| Model | Accuracy | Relevance | Understanding | Classification Precision |
|---|---|---|---|---|
| **GPT-4o** | High | High | Very High | High |
| **Gemini 2.5 Pro** | High | High | Medium | High |
| **Llama 4** | Medium | High | High | Medium |
| **Grok 3** | Medium | Medium | High | Medium |

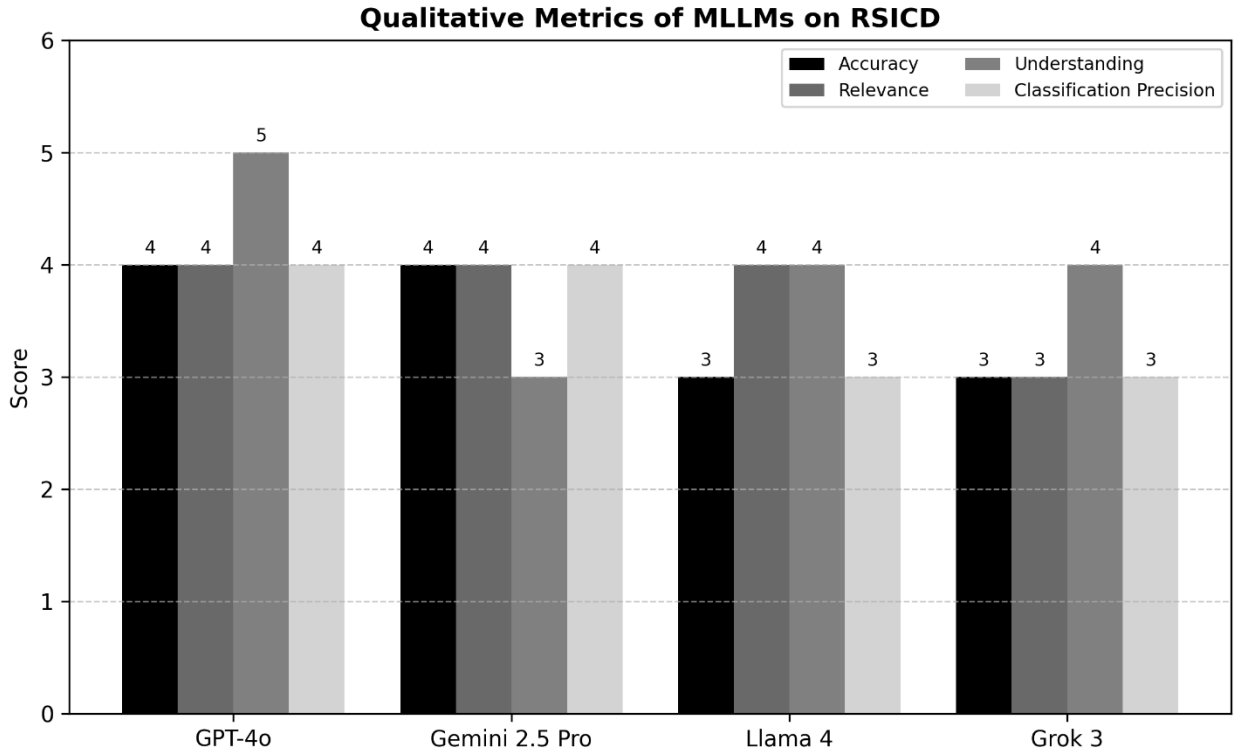


**Fig 1. Comparative qualitative metrics of MLLMs on RSICD**

According to the chart, GPT-4o consistently demonstrated superior performance, particularly in comprehension (score = 5). Gemini 2.5 Pro also exhibited strong performance, notably in Accuracy and Classification Precision (both = 4). Llama 4 presented balanced yet moderate outcomes, whereas Grok 3 was less consistent, although it showed notable comprehension (score = 4). This visualization corroborates the tabular findings, emphasizing that GPT-4o and Gemini 2.5 Pro are the most dependable for captioning and classification, while Llama 4 and Grok 3 display more variable strengths. Performance metrics were depicted through bar charts, confusion matrices, and example-based tables, revealing intra-class confusion (e.g., among 'dense' vs. 'sparse' residential scenes). GPT-4o and Gemini 2.5 Pro were generally the most proficient at making fine distinctions and descriptive syntheses.

### 4.2 Task-Specific Findings

**Image Captioning:** GPT-4o and Gemini 2.5 Pro led in factual correctness and relevance, with captions directly describing key image features. Llama 4 excelled in integrating broad spatial cues, while Grok 3 provided original but sometimes less precise narratives. The results show that GPT-4o and Gemini 2.5 Pro performed best overall in describing and categorizing images. Different types of charts and tables were

used to display the results, which showed that these two models were particularly good at noticing small details and giving accurate descriptions. For image captioning specifically, GPT-4o and Gemini 2.5 Pro were the most accurate, while Llama 4 was good at describing the overall layout of images, and Grok 3 gave creative but sometimes less precise descriptions.

**Image Classification:** All models achieved moderate precision, but GPT-4o and Gemini 2.5 Pro most reliably matched ground truth categories; Llama 4 and Grok 3 occasionally misclassified ambiguous scenes. The performance differences between the models highlight the varying strengths in visual understanding and language generation capabilities. While GPT-4o and Gemini 2.5 Pro demonstrated superior overall performance in image analysis tasks, Llama 4's ability to integrate broad spatial cues suggests a unique strength in comprehending image layouts and spatial relationships. Grok 3's tendency to provide original but less precise narratives indicates a potential for creative applications, albeit with a trade-off in accuracy. These findings underscore the importance of selecting the appropriate model for specific visual AI tasks. The superior performance of GPT-4o and Gemini 2.5 Pro in image captioning and classification suggests their suitability for applications requiring high precision and attention to detail. Conversely, Llama 4's strength in spatial understanding could be valuable in tasks involving complex visual layouts, while Grok 3's creative output might be beneficial in scenarios where novel interpretations are desired.

### 4.3 Richness, Contextualization, and Broader Implications

Multimodal large language models (MLLMs) are most effective when used as tools to complement, rather than replace, human analysts and specialised remote sensing solutions. Their ability to integrate multiple modes enhances contextual understanding and leads to more thorough hypothesis generation. Human validation is essential for tasks with substantial consequences, and the establishment of objective, quantitative benchmarks for evaluating depth and spatial reasoning is required to progress in the field. Hybrid methodologies that combine MLLMs with frameworks like EarthGPT and semantic alignment modules present a promising approach for enhancing precision and scalability[9] [8] .

## 5 CONCLUSION AND FUTURE WORK

Modern multimodal large language models (MLLMs), such as GPT-4o and Gemini 2.5 Pro, are very good at captioning and classifying satellite data, giving us useful and relevant information. Future research should focus on integrating hybrid models and doing further quantitative benchmarking before implementation in mission-critical environments; it is crucial that the outcomes be validated by human specialists. Other research objectives involve the establishment of quantitative benchmarks for multimodal thinking, enhancement of rapid engineering, and investigation of hybrid integration with domain-specific models (e.g., EarthGPT), and expanding evaluations for object detection, change analysis, and multi-sensor fusion. These advancements have the potential to significantly increase the interpretability and scalability of satellite image analysis using MLLMs.

### Conflicts of Interest

The authors declare no conflicts of interest related to this research.

### Ethical Approval

This study used publicly available datasets and did not involve primary data collection or sensitive information.

### References

1. K. Kömürcü and L. Petkevičius, "MiniCPM-V LLaMA Model for Image Recognition: A Case Study on Satellite Datasets," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 18, pp. 7892–7903, 2025, doi: 10.1109/JSTARS.2025.3547144.
2. S. Yin *et al.*, "A Survey on Multimodal Large Language Models," Nov. 2024, doi: 10.1093/nsr/nwae403.
3. J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal Large Language Models: A Survey," in *2023 IEEE International Conference on Big Data (BigData)*, IEEE, Dec. 2023, pp. 2247–2256. doi: 10.1109/BigData59044.2023.10386743.
4. X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, "Vision-Language Models in Remote Sensing: Current Progress and Future Trends," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2305.05726
5. X. Guo, W. Chai, S. Y. Li, and G. Wang, "LLaVA-Ultra: Large Chinese Language and Vision Assistant for Ultrasound," in *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2024, pp. 8845–8854. doi: 10.1145/3664647.3681584.
6. L. A. Kadiyala, O. Mermer, D. J. Samuel, Y. Sermet, and I. Demir, "The Implementation of Multimodal Large Language Models for Hydrological Applications: A Comparative Study of GPT-4 Vision, Gemini, LLaVa, and Multimodal-GPT," *Hydrology*, vol. 11, no. 9, Sep. 2024, doi: 10.3390/hydrology11090148.
7. K. Areerob *et al.*, "Multimodal artificial intelligence approaches using large language models for expert-level landslide image analysis," *Computer-Aided Civil and Infrastructure Engineering*, vol. 40, no. 19, pp. 2900–2921, Aug. 2025, doi: 10.1111/mice.13482.
8. S. Park, Y. Kim, S. Y. Kim, and Y. M. Ro, "Remote Sensing Large Vision-Language Model: Semantic-augmented Multi-level Alignment and Semantic-aware Expert Modeling," Jun. 2025, [Online]. Available: http://arxiv.org/abs/2506.21863
9. W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, "EarthGPT: A Universal Multi-modal Large Language Model for Multi-sensor Image Comprehension in Remote Sensing Domain," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2401.16822