Cloud Publications
International Journal of Advanced Remote Sensing and GIS
2025, Volume 13, Issue 1, pp. 3557-3569
ISSN 2320 – 0243

Research Article

Evaluating Vision-Enabled LLMs a Comparative Study on Cloud Detection using Horizon Camera Imagery

Allan Cerentini[1*], Juliana Marian Arrais[1], Bruno Juncklaus Martins[1], Sylvio Luiz Mantelli Neto[2], Tiago Oliveira da Luz[1], Aldo von Wangenheim[1]

[1]PPGCC - Federal University of Santa Catarina, Florianópolis, Santa Catarina, Brazil.
[2]FOTOVOLTAICA-UFSC, INPE Brazilian National Institute for Space Research, São José dos Campos, São Paulo, Brazil.

Correspondence should be addressed to Allan Cerentini, *allan.c@posgrad.ufsc.br*

Publication Date: 9 April 2025

Abstract The rapid advancement of vision-enabled large language models (LLMs) presents transformative opportunities for specialized domains such as atmospheric science. This study evaluates the efficacy of multimodal LLMs in cloud identification tasks by leveraging a curated subset of the Clouds-1500 dataset, annotated with World Meteorological Organization (WMO) cloud classes. We introduce a novel pipeline that converts segmentation masks into text-based spatial, coverage, and class representations, enabling structured LLM analysis through custom prompts and the BAML library for response standardization. Benchmarking 18 state-of-the-art models revealed significant performance variations, with Anthropic's Claude 3.5 Sonnet (71.67% class accuracy), OpenAI's GPT-4o (68.89%), and xAI's Grok Vision Beta (70.00%) emerging as top performers. However, challenges persist in low-coverage scenarios, where even leading models exhibited accuracy drops of 30–50%. The study demonstrates that while LLMs show promise in interpreting complex meteorological data, their effectiveness depends on task complexity, model architecture, and domain-specific adaptations. These findings provide a framework for integrating LLMs into remote sensing workflows, balancing automation with the precision required for operational meteorology.

Keywords *Remote Sensing; Large Language Models; Sky Clouds; Multimodal LLMs*

1. Introduction

Over the past few years, the rapid evolution of large language models (LLMs) has reshaped remote sensing in fields such as meteorology, environmental studies, and geospatial analysis. Research by (Zhu et al., 2025) and (Lin et al., 2025) shows that incorporating LLMs can streamline task distribution in unmanned aerial vehicle networks and enhance the clarity of image captions used in remote sensing. In parallel, initiatives like Aquila (Lu et al., 2024) combine visual and linguistic cues for more context-sensitive atmospheric analysis, while hybrid systems such as GeoLLM-Squad (Lee et al., 2025) manage multi-agent processes for tackling intricate geospatial challenges. Collectively, these developments signal a significant shift toward solutions that integrate multimodal data processing with specialized adaptability.

Recent breakthroughs in vision-enabled LLMs, including GPT-4o, Claude, and Gemini, have broadened the scope of image interpretation. Although these models are largely applied to general object recognition, their ability to analyze spatial relationships and quantify visual elements hints at promising opportunities in specialized remote sensing tasks. For instance, cloud segmentation stands to gain from this approach; the World Meteorological Organization (WMO) Cloud Atlas offers a standardized framework by classifying clouds according to altitude, shape, and optical properties. Such precise cloud detection is essential not only for aviation safety, for example, ensuring accurate monitoring of low ceilings at airports like Bombay as noted by (Kumar and Patkar, 2022), but also for climate modeling, where accurate albedo measurements are crucial. Conventional techniques based on manual annotation or convolutional neural networks (Veremev, 2021) are often hampered by scalability issues and limited in their ability to extract the positional metadata critical for operational meteorology.

Adding to the challenge, LLM outputs tend to vary, with free-form text complicating direct quantitative analysis. To address this, our approach transforms a dataset of cloud segmentation images; we converted the dataset to capture approximate textual representations of spatial positions, the percentage of the area covered, and the respective cloud classes. Next, we developed a custom prompt that instructs the neural network to provide these details in textual form. Finally, we utilize the BAML library to extract and convert the network's responses into structured format so we can generate metrics. This comprehensive pipeline not only enables rigorous model evaluation but also minimizes potential errors, drawing on strategies inspired by (Anderson et al., 2025).

Integrating computer vision datasets with natural language processing workflows, our method circumvents the high costs of re-annotation while preserving vital spatial details often lost in standard classification systems. This hybrid strategy not only boosts the accuracy of remote sensing insights but also extends their practical relevance. In this work, we apply these principles to atmospheric research, offering a scalable framework for evaluating how effectively LLMs can interpret complex meteorological phenomena using standardized, quantitatively verifiable outputs.

2. Objectives

In this study, our objectives center on evaluating the potential of Large Language Models (LLMs) in cloud identification, using a novel approach that leverages a pre-existing segmented cloud image dataset. Specifically, we aim to:

- Transform a cloud image segmentation dataset (WMO classes: Stratocumuliform, Stratiform, Cirriform, Cumuliform) into class, position, and area representations suitable for LLM analysis.
- Develop prompts that guide LLMs to extract relevant cloud information from images, including cloud class, location, and area coverage.
- Employ the BAML library to structure the LLM-generated textual responses, facilitating quantitative evaluation.
- Compare the performance of various LLMs to identify those with the best cost-benefit ratio for cloud identification tasks.

Through these objectives, we seek to provide insights into the feasibility and efficiency of utilizing LLMs for cloud analysis, contributing to a deeper understanding of their capabilities in interpreting atmospheric phenomena.

3. Methodology

3.1 Dataset

We developed a subset of 45 images from the Clouds-1500 dataset to perform an in-depth analysis of annotated segmentation masks and evaluate their compatibility with outputs from vision-enabled large language models (LLMs). This subset was carefully curated to represent diverse scenarios and challenges. Specifically, we selected images based on all possible combinations of the four cloud classes present in the dataset: Cirriforms, Cumuliforms, Stratiforms, and Stratocumuliforms. For each combination of cloud types (e.g., images containing only Cirriforms, combinations such as Cirriforms with Stratiforms, or images representing all four classes simultaneously), three images were chosen. Two of these images featured significant coverage of the target classes to facilitate classification, while one was selected for its minimal class coverage, providing a more challenging test case. To perform this selection, we developed an algorithm capable of computing the relative areas covered by each class in the segmentation masks. This algorithm identified images with both the largest and smallest coverage for each class combination, ensuring balanced representation of simplicity and complexity in the data. Some examples of the chosen images can be seen on Figure 1.

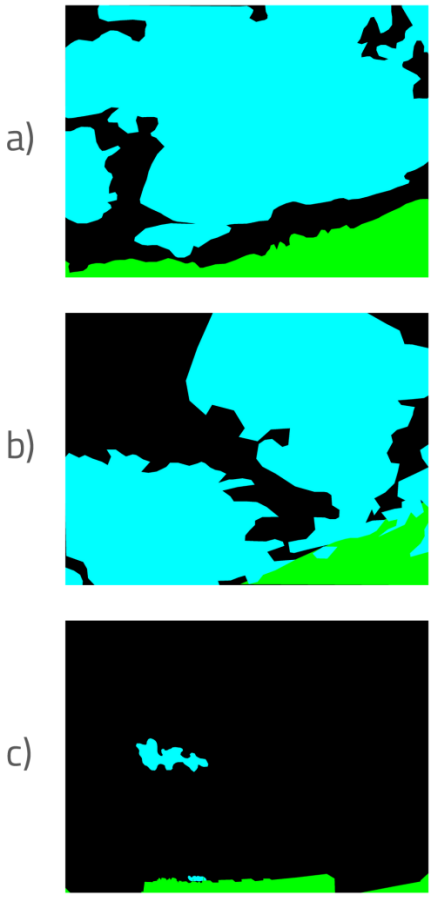

Figure 1: Segmentation mask examples showing the relative areas covered by the Cumuliform class, in cyan and other objects, green. An algorithm selected images with the (a, b) largest and (c) smallest Cumuliform coverage to ensure a balanced representation of data complexity.

The original Clouds-1500 dataset, which this subset was derived from, is an open-source compilation presented by (Arrais, 2023), building upon the Clouds-1000 dataset introduced by (Juncklaus Martins et al., 2022). The dataset comprises 1,500 annotated sky images captured between March 2021 and January 2023 using ground-based cameras installed at the Federal University of Santa Catarina and the Photovoltaic Energy Laboratory in Brazil. Aimed at advancing solar energy forecasting

applications, the dataset classify clouds based on their vertical structure into four main categories, Cirriforms, Cumuliforms, Stratiforms, and Stratocumuliforms, along with a fifth category for non-cloud background objects. These labels are intended to facilitate the study of solar radiation absorption by different cloud types, critical for enhancing solar energy systems' efficiency.

The Clouds-1500 dataset was originally collected using motionEye version 0.41 and Motion version 4.2.2 software, with image capture occurring at a frequency of one frame per minute between 08:00 and 22:00 GMT. The images have a resolution of 2592 x 1944 pixels and were processed locally before being uploaded to Google Drive (Arrais, 2023). The dataset is characterized by a significant variation in class distributions. The Object class, which delineates pixels that do not belong to clouds, is observed most frequently, appearing in 1,376 images and covering 17.02% of the total annotated area. Stratocumuliform clouds are the most prominent cloud type, represented in 1,095 images and accounting for 35.64% of the dataset area. In comparison, Stratiform clouds appear in 453 images (11.01%), Cirriform clouds in 382 images (4.81%), and Cumuliform clouds in 251 images (3.58%). This distribution reflects regional climatic trends, as the humid conditions of the area limit the prevalence of Cumulonimbus clouds, which are more common in drier climates.

By extending this extensive dataset through the creation of a representative subset of 45 images, we aim to explore new methodologies for processing cloud segmentation masks and evaluating the performance of artificial intelligence systems analyzing sky images.

3.2 Benchmark workflow

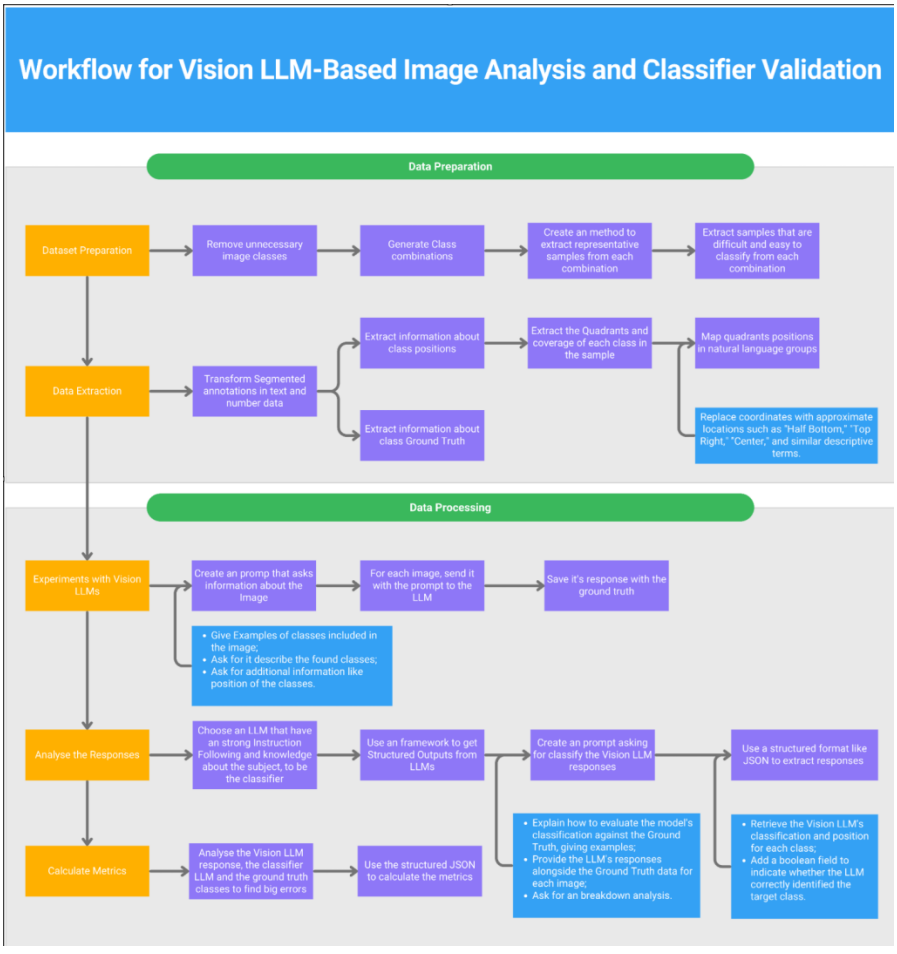

***Figure 2**: This diagram illustrates the workflow for analyzing images and validating classifiers using Vision LLMs. The process includes data preparation, extraction, and processing, followed by response analysis and metric calculation. Key steps are marked in orange, intermediate steps in purple, and helpful tips in blue.*

The diagram in the Figure 2, illustrates the complete methodology used to develop the benchmark presented in this paper. This approach can be applied to various identification tasks to evaluate the performance of LLMs with vision capabilities in solving domain-specific problems. Below, we will discuss each step in detail.

3.2.1 Data Preparation

The first step was to remove any irrelevant classes from the dataset, such as "Object," which was not pertinent to the experiment's focus on cloud identification. Next, combinations of cloud classes within individual images were analyzed to ensure that examples representing various class combinations were present. An algorithm was then designed to identify images with a balanced distribution of cloud types across these combinations. This ensured that the dataset included both straightforward and complex examples for classification. The next phase involved extracting positional and quantitative information from the annotated data. Cloud positions were described using natural language terms such as "Top," "Bottom," "Upper Left," or "Center" to make them more interpretable and comparable with LLMs outputs. A straightforward method splits the image into nine segments. Each segment is then examined for the presence of different cloud types, and a specific label is assigned based on the combination of segments where clouds appear. Additionally, each cloud's coverage area in the image was calculated as a percentage of the total image area.

Finally, this information was structured into a format that LLMs could easily process and understand. JSON was chosen as the primary format due to its simplicity and compatibility with LLMs. For each image, details such as cloud classes, their respective positions (in quadrants), coverage percentages, and the dominant cloud class were included in the JSON structure, like in Figure 3.


```
"stratocumuliform_quadrants": "Upper Left and Left",
"stratocumuliform_coverage": "27.13%",
"stratiform_quadrants": null,
"stratiform_coverage": null,
"cirriform_quadrants": "Top and Right",
"cirriform_coverage": "25.56%",
"cumuliform_quadrants": "Bottom",
"cumuliform_coverage": "17.88%",
"dominant_class": "Stratocumuliform",
"total_pixels": 3556084
```


Figure 3: Example of a JSON item generated by the algorithm.

This structured representation allowed for precise extraction of key features during evaluation while maintaining compatibility with natural language prompts used in querying LLMs. By preparing the data in this way, we ensured that it could effectively bridge the gap between traditional segmentation approaches and LLM-based text-driven analysis methods.

3.2.2. Data Processing

First, state-of-the-art Large Language Models (LLMs) with vision capabilities were explored, leveraging OpenRouter to access a diverse array of such models for experimentation.

A carefully crafted prompt was designed to elicit high-quality outputs. This involved specifying classes, positions, and descriptions for identified objects, alongside a request for a comprehensive image analysis. Specifically, the LLM was instructed to identify clouds, classify their types based on World Meteorological Organization (WMO) standards, provide approximate locations (e.g., top left, bottom right, above the horizon), estimate cloud coverage in percentages, and detail the visual characteristics that informed the cloud type identification.

After processing all images through the LLMs, another network was employed to extract information in a format suitable for comparison across different outputs. The BAML (Basically a Made-up Language) library was used to ensure that the LLM responses were consistently structured, with JSON selected as the output format. The fields within the JSON output were then analyzed and compared against the LLM's original response to determine accuracy (true or false). The "Gemini 2.0 flash thinking exp" LLM was used for this evaluation.

```
{
    "stratocumuliform_quadrants": "Whole Image",
    "stratocumuliform_coverage": "97.00%",
    "stratiform_quadrants": null,
    "stratiform_coverage": null,
    "cirriform_quadrants": null,
    "cirriform_coverage": null,
    "cumuliform_quadrants": null,
    "cumuliform_coverage": null,
    "dominant_class": "Stratocumuliform",
    "total_pixels": 4887860,
    "response": "Here's an analysis of the cloud ty
    "baml_stratocumuliform_class": true,
    "baml_stratocumuliform_position": true,
    "baml_stratocumuliform_coverage": true,
    "baml_stratiform_class": false,
    "baml_stratiform_position": false,
    "baml_stratiform_coverage": false,
    "baml_cirriform_class": false,
    "baml_cirriform_position": false,
    "baml_cirriform_coverage": false,
    "baml_cumuliform_class": true,
    "baml_cumuliform_position": true,
    "baml_cumuliform_coverage": true,
    "baml_found_more_classes_than_gt": true
}
```


Figure 4: Example of JSON generated by BAML, after an LLM analyses the response from the multimodal LLM.

In this example, Figure 4, the LLM correctly identified stratocumuliform clouds covering the whole image with 97% coverage. The LLM also provided a detailed analysis, noting visual characteristics such as thickness, altitude, and structure. BAML was then used to evaluate specific aspects of the response, such as the classification, position, and coverage of different cloud types, marking whether the LLM's assessment aligned with the expected values (true) or not (false), being lenient in cloud coverage and position. For instance, the LLM accurately classified the stratocumuliform cloud type ("baml_stratocumuliform_class": true) and its position ("baml_stratocumuliform_position": true), among other parameters.

To conclude the methodology, the final step involved calculating and analyzing various metrics to evaluate the performance of the Vision LLMs in cloud identification tasks. The JSON outputs generated by BAML were aggregated to compute accuracy scores for each category, including cloud classification, position identification, and coverage estimation. These accuracy metrics provided a quantitative measure of the LLMs' performance across different aspects of cloud analysis. This approach allowed for a comprehensive evaluation of the Vision LLMs' effectiveness in domain-specific image analysis, particularly in the context of cloud identification.

4. Results and Discussions

In our analysis, we have carefully selected some of the most recent and state-of-the-art multimodal large language models, although we must note that certain models were excluded from this study due to technical limitations. Our focus was on identifying systems that push the boundaries of both visual and language processing while delivering innovative solutions for a wide range of applications.

The latest advancements in multimodal large language models (LLMs) bring together cutting-edge vision and language capabilities, with each model series offering unique strengths. Amazon's Nova models, including Nova Lite and Nova Pro, focus on efficient multimodal processing within the Bedrock ecosystem. Anthropic's Claude 3.5 Sonnet excels in agentic tasks, particularly in code generation and

automation, while Claude 3 Haiku emphasizes speed and lightweight performance. Google's Gemini 2.0 series, with Flash and Pro variants, integrates advanced attention mechanisms for chart interpretation and long-context processing. Meta's LLaMA 3.2 Vision models leverage innovative cross-attention techniques to connect image encoders with text-based reasoning. Mistral's Pixtral models debut with a dual encoder-decoder architecture optimized for mathematical and visual reasoning. OpenAI's GPT-4o pushes the boundaries of unified multimodal processing, excelling in real-time text, audio, and image understanding. xAI's Grok Vision Beta introduces temporal convolution layers for long-video analysis, while Qwen VL 72B focuses on spatial reasoning with advanced position embeddings. These models showcase diverse approaches to multimodal AI, setting new benchmarks across various applications.

Table 1: Table displaying experiment results, using accuracy as the metric to show the percentage of correct predictions for each category by different LLMs.

Model Name	Class	Position	Coverage
Amazon Nova Lite v1	59.44	54.44	50.0
Amazon Nova Pro v1	59.44	54.44	51.11
Anthropic Claude 3 Haiku	60.56	50.0	46.67
Anthropic Claude 3.5 Sonnet	71.67	66.11	63.33
Google Gemini 2.0 Flash Lite	45.56	45.0	41.67
Google Gemini 2.0 Pro	62.78	55.0	51.67
Google Gemini Flash 1.5	60.0	60.56	52.78
Google Gemini Pro 1.5	61.11	61.67	55.0
Meta LLaMA 3.2 11B Vision	62.78	53.33	43.89
Meta LLaMA 3.2 90B Vision	57.78	52.78	38.33
Mistral Pixtral 12B	62.22	56.11	43.33
Mistral Pixtral Large	61.11	48.89	48.33
OpenAI GPT-4o	68.89	62.22	61.67
OpenAI GPT-4o Mini	68.89	58.89	48.89
Qwen VL 72B	60.0	55.56	51.11
Qwen VL 7B	50.56	49.44	41.11
Qwen VL Plus	60.0	54.44	47.78
xAI Grok Vision Beta	70.0	62.22	58.89

The Table 1 shows how various multimodal models perform when tasked with identifying WMO cloud types, pinpointing their positions in images, and estimating their area coverage. There are noticeable differences in performance across the board. For example, Anthropic's Claude 3.5 Sonnet leads the pack with scores of 71.67 for cloud type classification, 66.11 for locating clouds, and 63.33 for coverage estimation. Its ability to handle both straightforward and challenging images is likely a result of its strong reasoning and contextual understanding.

OpenAI's GPT-4 Latest comes in a close second, scoring 68.89 for classification, 62.22 for position accuracy, and 61.67 for coverage. However, its smaller GPT-4 Mini variant shows a drop, scoring 58.89 for position and 48.89 for coverage, perhaps because of its reduced architecture.

xAI's Grok Vision Beta also performs impressively, especially in classifying cloud types (70.0) and in position accuracy (62.22), although its coverage score is a bit lower (58.89). This might indicate that while it's good at understanding the overall scene, it struggles slightly with estimating exact cloud areas compared to Claude 3.5 or GPT-4 Latest.

Google's Gemini Pro 1.5 delivers balanced results with a notable strength in position accuracy (61.67) and competitive scores for both classification (61.11) and coverage (55.0). In contrast, its Flash Lite version lags considerably (45.56 for classification, 45.0 for position, and 41.67 for coverage), suggesting it isn't as well-suited for handling more complex visual tasks.

Meta's LLaMA 3.2 Vision models have mixed outcomes. The 11B version outperforms the larger 90B model in both classification (62.78 versus 57.78) and coverage (43.89 versus 38.33). This could imply that the larger model might be overfitting or less efficient at processing the subtle details in challenging images.

Mistral's Pixtral models show reliable performance as well. The Pixtral 12B variant edges out the Pixtral Large version, particularly in position accuracy (56.11 compared to 48.89).

Amazon's Nova models are consistent but not outstanding. Both Nova Lite and Nova Pro register the same scores for classification (59.44) and position (54.44), with Nova Pro just a bit ahead in coverage (51.11 versus 50.0). This consistency suggests they are dependable on simpler tasks, though they may not excel when the image complexity increases.

Lastly, Qwen's VL series presents moderate performance. The VL Plus variant outshines the smaller VL 7B model in all categories, yet it still doesn't reach the heights of the top performers like Claude 3.5 or GPT-4 Latest. With scores of 60.0 for classification, 54.44 for position, and 47.78 for coverage, Qwen VL Plus remains a balanced, though not leading, option.

In short, Anthropic's Claude 3.5 Sonnet clearly emerges as the best overall model for these tasks, while OpenAI's GPT-4 Latest and xAI's Grok Vision Beta also show robust multimodal reasoning. Other models, such as Google's Gemini Pro 1.5 and Mistral's Pixtral 12B, provide solid alternatives but tend to fall short when it comes to managing highly complex visual scenarios.

Table 2: Performance of Large Language Models (LLMs) on Tasks with High Coverage (Easy) and Low Coverage (Hard) Across Different Cloud Types.

Model Name	Strato [H]	Strato [L]	Strati [H]	Strati [L]	Cirri [H]	Cirri [L]	Cumul [H]	Cumul [L]
Amazon Nova Lite v1	81.25	75.0	100.0	87.5	93.75	87.5	81.25	87.5
Amazon Nova Pro v1	100.0	87.5	100.0	87.5	68.75	62.5	75.0	75.0
Anthropic Claude 3 Haiku	93.75	75.0	75.0	62.5	62.5	75.0	87.5	87.5
Anthropic Claude 3.5 Sonnet	81.25	50.0	68.75	12.5	100.0	87.5	87.5	87.5
Google Gemini 2.0 Flash Lite	43.75	0.0	43.75	12.5	37.5	12.5	37.5	12.5
Google Gemini 2.0 Pro	100.0	62.5	75.0	37.5	81.25	87.5	81.25	75.0
Google Gemini Flash 1.5	75.0	25.0	43.75	12.5	12.5	12.5	68.75	50.0
Google Gemini Pro 1.5	31.25	12.5	31.25	0.0	62.5	25.0	81.25	75.0
Meta LLaMA 3.2 11B Vision	93.75	87.5	81.25	62.5	100.0	75.0	93.75	62.5
Meta LLaMA 3.2 90B Vision	87.5	75.0	93.75	87.5	100.0	87.5	93.75	100.0
Mistral Pixtral 12B	100.0	100.0	100.0	100.0	100.0	100.0	87.5	37.5
Mistral Pixtral Large	100.0	75.0	81.25	12.5	100.0	87.5	87.5	87.5
OpenAI GPT-4 Latest	100.0	62.5	75.0	37.5	68.75	87.5	75.0	62.5
OpenAI GPT-4 Mini	93.75	62.5	87.5	62.5	81.25	87.5	87.5	87.5
Qwen VL 72B	100.0	62.5	68.75	12.5	100.0	100.0	87.5	87.5
Qwen VL 7B	93.75	87.5	12.5	25.0	0.0	0.0	18.75	0.0
Qwen VL Plus	87.5	62.5	62.5	37.5	56.25	62.5	62.5	0.0
xAI Grok Vision Beta	93.75	100.0	68.75	12.5	100.0	100.0	100.0	87.5

Table 2 offers a detailed perspective on how various large language models perform when tested on different cloud types with varying degrees of difficulty. The evaluation separates the tasks into two groups: one that is easier referred to as high coverage and another that is more challenging, known as low coverage. The analysis uncovers clear patterns and brings to light both the strengths and weaknesses of each model.

One important observation is that some models deliver consistently strong results in both scenarios while others show significant differences between the two. For example, Mistral's Pixtral 12B achieves exceptional results by scoring 100 in nearly every cloud type. Its only slight weakness is seen in the low coverage category labeled "Cumul," where it scores 37.5. This performance indicates that Pixtral 12B is capable of handling both straightforward and complex tasks, underlining its strength and versatility. In contrast, its smaller sibling, Pixtral Large, performs adequately overall but struggles with more challenging tasks, as seen in the low coverage category "Strati," where it scores only 12.5. This clearly points to its limitations when dealing with harder examples.

Anthropic's Claude 3.5 Sonnet shows a similar tendency. The model excels in tasks requiring high accuracy, scoring a perfect 100 in the "Cirri" high coverage category. However, its performance drops markedly in more demanding low coverage cases, such as the "Strati" category where it reaches only 12.5, and it displays only moderate scores in other challenging areas. Its predecessor, Claude 3 Haiku, offers a more balanced set of scores but generally does not achieve the high marks seen in the newer version.

OpenAI's GPT-4 Latest is another model that stands out, especially in the high coverage tasks "Strato" and "Cumul," where it scores perfectly. Yet, much like several other models, its accuracy diminishes in low coverage settings; for example, it only scores 37.5 in the "Strati" category. The GPT-4 Mini variant, although slightly less capable overall, maintains relatively stable scores. This consistency suggests that its smaller design may limit its ability to perform exceptionally in more complex tasks.

Meta's LLaMA 3.2 Vision models present an interesting contrast between different sizes. The 11 billion parameter version outperforms the larger 90 billion parameter model in certain areas. For instance, in the "Cumul" high coverage category, the smaller version scores 93.75 while the larger one scores 87.5. In other areas, such as the "Cumul" low coverage category, the larger model scores higher, with a 100 compared to 62.5 for the 11 billion model. This unexpected pattern, given the usual advantage of larger models in capturing detail, may point to issues like overfitting or other inefficiencies in the larger architecture.

xAI's Grok Vision Beta achieves impressive results by excelling in both low coverage categories "Strato" and "Strati" with perfect scores, and it also maintains competitive marks in high coverage tasks. This balanced performance demonstrates the model's ability to manage both simple and complex scenarios effectively.

Google's models, on the other hand, exhibit more variability. Gemini Pro 1.5 shows reasonable performance in high coverage situations, scoring 81.25 in the "Cumul" category, but it fails to deliver in low coverage tasks such as "Strati," where it scores 0. The Flash Lite version, with considerably lower scores across all tasks, confirms its limitations in handling both easy and challenging cases. Although Gemini 2.0 Pro performs better overall, it still struggles with more demanding, low coverage scenarios. Amazon's Nova models provide consistent and respectable outcomes without reaching the high peaks of some of the other top performers. Both Nova Lite and Nova Pro score solidly in high coverage tasks, each achieving a perfect score in the "Strato" category. Their performance in more challenging low coverage cases is moderate, suggesting that they encounter difficulties when adapting to harder examples.

Finally, Qwen's VL series shows some promising aspects but generally falls behind the leading models. The VL Plus variant manages a moderate score of 56.25 in the "Cirri" high coverage task but does not perform well in low coverage situations, as seen in the "Cumul" category where it scores 0. The smaller VL 7B model performs poorly across the board, with very low scores in both high and low coverage tasks, indicating that it is not well suited to these challenges.

In conclusion, Mistral's Pixtral 12B emerges as the most dependable model in terms of achieving high accuracy across a range of tasks. Other strong performers include xAI's Grok Vision Beta, OpenAI's GPT-4 Latest, and Anthropic's Claude 3.5 Sonnet, all of which excel in several areas despite varying capacities in more complex cases. Conversely, Meta's larger LLaMA Vision model and most of Google's Gemini variants struggle with nuanced details, while Amazon's Nova models and Qwen VL Plus offer steady but less remarkable performance. This analysis underscores that success in simpler, high coverage tasks does not automatically translate to effective handling of the more challenging low coverage scenarios.

5. Conclusion

Our study presents a scalable method that repurposes legacy segmentation datasets for LLM evaluation without resorting to costly manual re-annotation. By converting pixel-level annotations into text-based descriptors that capture class, position, and coverage and by using the BAML library to organize LLM outputs, we show how decades of domain-specific image archives can be refreshed for modern AI benchmarking. This approach removes the need to label thousands of new images or reconcile unstructured responses because existing segmentation metadata serves as reliable ground

truth. The modular workflow transforms visual data into spatial textual representations, queries LLMs with customized prompts, and automatically extracts structured answers. The resulting template applies to many areas in remote sensing, including land cover classification and disaster monitoring.

This framework also answers an important industry question: when should practitioners use off-the-shelf LLMs and when is custom training necessary? Testing pre-trained models against standardized descriptors allows organizations to quickly determine if current LLMs meet their needs. For example, if a model like Claude 3.5 Sonnet achieves 90 percent accuracy in detecting rare geological formations from translated segmentation data, expensive fine-tuning may be avoided. Conversely, poor performance on specialized tasks such as identifying crop disease patterns indicates the need for domain-specific training. In this way, our method acts as a cost-effective triage system that guides resource allocation in AI deployment.

Future research should include comparisons with the newest LLMs currently available, as the field is growing rapidly. In addition, it is important to test other libraries to generate structured outputs, explore different prompts for data extraction and extend the analysis to a larger number of images.

References

Anderson, M., Cha, M., Freeman, W.T., Perron, J.T., Maidel, N., Cahoy, K., 2025. Measuring and Mitigating Hallucinations in Vision-Language Dataset Generation for Remote Sensing. https://doi.org/10.48550/arXiv.2501.14905

Arrais, J.M., 2023. Clouds-1500. *https://doi.org/10.17632/2KHCHJBGZR.2*

Bradski, G., 2000. The OpenCV Library. Dr. Dobb's Journal of Software Tools.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794.

Chu, L., Liu, Y., Wu, Z., Tang, S., Chen, G., Hao, Y., Peng, J., Yu, Z., Chen, Z., Lai, B., Xiong, H., 2021. PP-HumanSeg: Connectivity-Aware Portrait Segmentation with a Large-Scale Teleconferencing Video Dataset. *https://doi.org/10.48550/ARXIV.2112.07146*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., others, 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: Proceedings of the International Conference on Learning Representations (ICLR).

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning, Adaptive computation and machine learning. MIT Press.

Juncklaus Martins, B., Polli, M., Cerentini, A., Mantelli, S., Chaves, T., Moreira Branco, N., von Wangenheim, A., Arrais, J., 2022. Clouds-1000. *https://doi.org/10.17632/4pw8vfsnpx.1*

Kumar, S., Patkar, P., 2022. Low level wind shear over Bomba airport. MAUSAM.

Lee, C., Paramanayakam, V., Karatzas, A., Jian, Y., Fore, M., Liao, H., Yu, F., Li, R., Anagnostopoulos, I., Stamoulis, D., 2025. Multi-Agent Geospatial Copilots for Remote Sensing Workflows. *https://doi.org/10.48550/arXiv*.2501.16254

Li, H., Wang, S., Zuo, W., Zhang, L., 2020. PPLite: Efficient Convolutional Neural Networks with Dynamic Pointwise Filters. arXiv preprint arXiv:2003.11506.

Lin, H., Hong, D., Ge, S., Luo, C., Jiang, K., Jin, H., Wen, C., 2025. RS-MoE: A Vision-Language Model with Mixture of Experts for Remote Sensing Image Captioning and Visual Question Answering. *https://doi.org/10.48550/arXiv.2411.01595*

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Lu, K., Zhang, R., Huang, X., Xie, Y., 2024. Aquila: A Hierarchically Aligned Visual-Language Model for Enhanced Remote Sensing Image Comprehension. *https://doi.org/10.48550/arXiv.2411.06074*

Luo, W., Zhao, H., Li, L., Wang, C., 2022. PP-LiteSeg: Lightweight Model for Real-Time Semantic Segmentation. arXiv preprint arXiv:2201.00239.

Martins, B.J., Arrais, J.M., Cerentini, A., Wangenheim, A. von, Neto, G.P.R., Mantelli, S., 2023. Segmentation and Classification of Individual Clouds in Images Captured with Horizon-Aimed Cameras for Nowcasting of Solar Irradiance Absorption. American Journal of Climate Change 12, 628–654. *https://doi.org/10.4236/ajcc.2023.124027*

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation, in: 3D Vision (3DV), 2016 Fourth International Conference On. IEEE.

Rahman, S., Wang, Y., 2016. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation, in: International Symposium on Visual Computing. Springer, pp. 234–244.

Rousson, M., Lenglet, C., Deriche, R., 2008. The Dice Metric Considerations. Insight Journal 2008, pp.1–10.

Veremev, N., 2021. The use of artificial neural networks in the problem of classifying cloud types in wide-angle images of the visible hemisphere of the sky.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., others, 2020. Deep High-Resolution Representation Learning for Visual Recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Zhu, E., Zheng, K., Gao, K., Zhang, J., Yang, Z., Wang, Z., 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. arXiv preprint arXiv:2105.15203.

Zhu, F., Huang, F., Yu, Y., Liu, G., Huang, T., 2025. Task Offloading with LLM-Enhanced Multi-Agent Reinforcement Learning in UAV-Assisted Edge Computing. Sensors 25, 175. *https://doi.org/10.3390/s25010175*