# Can Multimodal Large Language Models Identify Thematic Maps in Scientific Articles?

**Tomas Turek**[1, 2], **Vincent Nguyen**[1], **Dhirendra Singh**[1], **Sarvnaz Karimi**[3], **Xiuzhen Zhang**[2]
[1]CSIRO (Australia), [2]RMIT University (Australia), [3]Monash University (Australia)
**Correspondence:** firstname.lastname@csiro.au, sarvnaz.karimi@monash.edu, xiuzhen.zhang@rmit.edu.au

## Abstract

Information seeking systems have yet to access multimodal information about thematic maps in scientific literature. Thematic maps are images of geographical maps showing themes of information and are crucial for understanding the impact of climate on natural resources. Although Multimodal Large Language Models (MLLMs) are applied in different applications, their effectiveness in identifying thematic maps is not explored. We address this gap by evaluating some of the state-of-the-art MLLMs to (1) classify thematic maps using different visual and textual contexts; (2) explain their predictions and uncover their sources of bias using *attribute discovery*; and (3) assist humans with the annotation of thematic maps in the scientific literature. We release a new public dataset[1] and use it to show that MLLMs that were investigated in our study have room for improvement. We observe that MLLMs can be biased toward visual attributes of thematic maps (graphical indicators). The best classification result is an 80% strict F1 score by the Grok 4.3 model, which can fluctuate between 76% and 84% due to classification bias. Overall, we consistently find that the input of images with captions and examples of the correct answers yields substantial gains in task performance.

## 1 Introduction

Current multimodal tasks lack connection to the early work exploring the information in thematic maps (Tousidou and Manolopoulos, 2000; d'Onofrio and Pourabbas, 2001; Mauro and Ardissono, 2017). Understanding complex geographical context and user demands remains a challenge when searching for maps, even in structured sources of information (Gui et al., 2024). For example, it is difficult to classify themes in images and text metadata to help users find relevant maps from spatial databases (Yang et al., 2020; Wei et al., 2021). Emerging multimodal retrieval applications critically need methods to identify complex visual-textual elements in documents (Qiao et al., 2025).

To illustrate, identifying climate adaptation strategies for a specific location requires climate data and knowledge of the scientific literature in text and figures. One type of such figures is thematic maps, which allow analysing the spatial distribution of thematic information in a specific region. For a system that retrieves relevant literature, such as a question-answering system for climate adaptation (Nguyen et al., 2025), identifying these maps in scientific articles could be a first step.

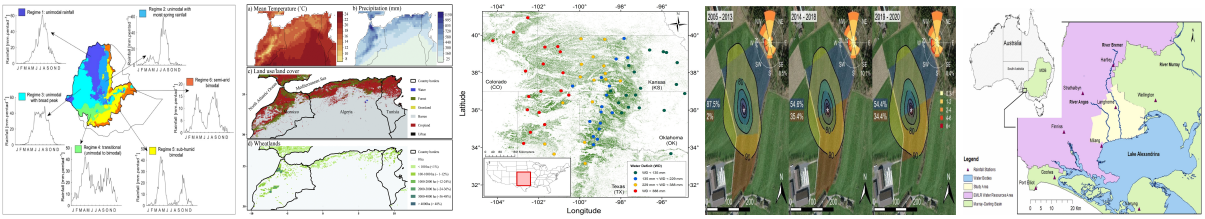
MLLMs are a key candidate for this task, but there is no study to investigate how effective they would be in detecting such figures in scientific articles. To bridge this gap, we investigate the classification of *thematic maps* in agricultural adaptation journals (examples shown in Figure 1). Thematic maps are spatial maps that illustrate particular themes in a landscape, such as precipitation in regional areas, and require specialised knowledge to interpret. We define a thematic map as an image that includes (i) a **base map** showing geographical features and (ii) a **thematic overlay** (Monmonier, 2001) showing measurable climatic parameters, such as land and sea surface temperatures. This is an essential theme of remote sensing information required for climate monitoring (WMO, 2025) that practitioners seek to better assess climate-related risks (Alonso-Robisco et al., 2026).

We evaluate closed- and open-source MLLMs on their ability to classify figures from scientific articles that contain thematic maps, using the image, its caption, and related discourse (text sentences that reference the figure) as input. Varying the input data demonstrates that the length and relevance of the textual context influence the performance of classifiers, e.g., combining shorter captions with longer scientific discourse degrades the output of

[1] https://github.com/tomtx/cp-thematic-maps



*Proceedings of the 3rd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2026)*, pages 231–245
October 28, 2026 ©2026 Association for Computational Linguistics



(a) chorochromatic map (b) choropleth map (c) dot distribution map (d) isopleth map (e) general reference map

Figure 1: Examples of thematic map types in agricultural adaptation showing climatic parameters: (a) rainfall regimes, (b) precipitation and temperature, (c) water deficit, and (d) wind climatology. Example of other maps (e).

models. Analysing the visual attributes of map figures that lead to correct classifications reveals complex and spurious features, showing that models are biased toward directional, numerical, and physical representations on maps (graphical indicators) and struggle with colours and layouts of figures (gradients and subplots).

In this context, we advance the MLLM-assisted processing of multimodal research evidence in climate science with three contributions: (i) a novel classification task of thematic map identification in scientific articles with user instructions; (ii) a novel attribute discovery methodology to explain predictions of MLLMs in the classification of thematic maps; and (iii) a new data source of human-annotated figures with heterogeneous thematic maps of various types and locations in public, peer-reviewed scientific articles (CC BY-NC-SA 4.0). This assists downstream tasks in scientific domains that require multimodal evidence in scientific figures.

## 2 Related Work

**Classification of Scientific Figures** The closest work to our research is the ACL-FIG pipeline (Karishma et al., 2023) for a multi-class classification of images into standard types of figures, such as bar charts and maps, with 83% accuracy (vision transformer). They extract images from scientific articles, cluster and annotate image vectors into figure types, and link the figure metadata. Previous methods only label whether the image is a figure or table, such as DEEPFIGURES (Siegel et al., 2018). Our efforts advance ACL-FIG with a *multimodal classification* of scientific figures by (i) a classification of multimodal context of figures (images with texts instead of images only) and (ii) a fine-grained classification of map figures (thematic maps with more complex information instead of maps only).

**Datasets with Map Figures** Other works on benchmark datasets with map figures do not capture specialised data representations of different maps in scientific domains to enable multimodal classification. Benchmarks either contain unannotated or unimodal figures as the images are not labelled nor accompanied by captions or other relevant texts from their source. For example, SCIFIBENCH (Roberts et al., 2024) contains captioned figures, some of which are maps, but lacks annotated labels to understand whether the image is a box plot or a map and which article it originates from to contextualise the data. MAPQA (Chang et al., 2022) is the only dedicated dataset for thematic maps showing health indicators in choropleths. However, this is not a source of scientific figures, given the lack of captions. In addition, it is overly homogeneous and does not represent common types and spatial context of thematic maps, as the data are distributed over the same geographic region map.

**Geospatial Tasks** Images of thematic maps, which are published in scientific articles, are generally constructed by processing various remote sensing imagery (Foody, 2004; Schowengerdt, 2007). For example, geo-referenced aerial and satellite imagery are the geospatial data that are often used as base maps of thematic maps. MLLMs currently underperform in fine-grained object classification of multimodal geospatial data in the computer vision literature, e.g., below a 60% accuracy on GEOBENCH (Danish et al., 2025). This is due to their sensitivity to distribution shifts and prompt variations and their weak numerical, temporal, and spatial reasoning capabilities in complex geospatial scenes. However, such complex images of base maps do not contain thematic overlays captured by remote sensing technologies (such as climatic parameters) to understand the performance of MLLMs on thematic maps in geospatial tasks.

**Visual Bias of Classifiers** Previous efforts in classification tasks do not analyse visual attributes of scientific figures to understand spurious correlations between their labels and attributes. It is known that classifiers trained on biased images underperform on images that conflict with bias, e.g., misclassifying waterbirds (labels) on land backgrounds (attributes) as landbirds (Yenamandra et al., 2023). Most of the work is done on natural images with vision classifiers, some of which use language models to generate descriptions of images in order to better understand attributes in which classifiers fail (Ghosh et al., 2025).

Although MLLMs are generally not used as classifiers, other tasks report on their bias toward particular attributes in images. MLLMs over-rely on learnt spurious correlations in training data and then poorly generalise on out-of-domain data, e.g., developing a class-dependent spurious bias in object detection tasks (Hosseini et al., 2025). MLLMs are also prone to visual hallucinations, which is a misalignment between the responses generated and the visual evidence provided, in question answering tasks (Suri et al., 2025).

## 3 Methodology

**Multimodal Classification** We propose a *multimodal classification* task: given an image and text of a map figure, the task is to correctly identify a thematic map. We formulate the task as a multi-label classification of themes since a map figure can contain multiple plots of thematic maps, each with multiple themes. This is a problem of instance classification, where each instance may have more than one correct class. We expect that models already learnt during their training a mapping function between instances in a multimodal dataset and their classes to assign the correct class labels to a given instance. In other words, $f : ({x_v}^{(i)} \times {x_t}^{(i)}) \rightarrow y^{(i)}$, where $y \in \bigcup_{j=1}^{n} y_j$ represents the classes (themes), $n$ is the number of classes, and $x$ is the feature input vector of textual and visual modalities, $t$ and $v$ respectively, for the instance $i$ (map figure).

We leverage MLLMs to classify the map figure into eight classes that pertain to thematic maps. MLLMs need to generate a multi-response answer to a multi-choice question from users. The question is defined by our domain experts as a command: *"Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none."*.

**Attribute Discovery** We aim to explain the predictions of MLLMs on our classification task by systematically analysing the visual attributes of map figures. In this context, a *visual attribute* refers to a distinct visual feature of the image, such as legends, grid lines, and displayed values. For this purpose, we define a *target attribute* as any visual feature under investigation to determine its impact on classifier performance. Since a single figure can simultaneously contain multiple visual attributes, these target attributes are not mutually exclusive and can overlap both within a single figure and across multiple figures. To isolate the specific effects of these target attributes, we categorise them into *favourable* and *unfavourable* attributes based on whether they correlate with an increase or decrease in classifier performance, respectively. Ultimately, we define a *biased* attribute as a favourable attribute that introduces the risk of shortcut learning due to spurious correlations between task-relevant labels and task-irrelevant attributes (Yenamandra et al., 2023).

With the above definitions, we propose an *attribute discovery* methodology leveraging MLLMs to explain the general classification performance of classifiers. To identify target attributes within figures, MLLMs serve as hypothesis generators and self-report on their bias through a three-step process by generating (i) detailed textual descriptions of images; (ii) hypotheses about target attributes (which influence classifier performance due to inherent data and model biases) using randomly selected samples of captions and generated descriptions; and (iii) labels to denote whether a given figure caption or image contains a target attribute. Subsequently, these target attributes are filtered and validated through a four-step process by selecting (i) attributes present in a significant sample size of figures; (ii) favourable and unfavourable attributes that are measured by changes in classifier performance when these attributes are present versus absent in the figures; (iii) biased attributes from the subset of favourable attributes, which introduce disproportionate increases in classifier performance when present, indicating the exploitation of a spurious correlation (shortcut learning); and (iv) a small subset of figures for manual inspection to verify the presence of bias in the ground truth.

We discover such a bias of proprietary MLLMs in visual attributes by modifying the LADDER methodology (Ghosh et al., 2025) to eliminate their specialised training of classifiers. Our training-free approach evaluates classifiers at inference time via MLLM-generated: (i) descriptions of images to provide context on visual attributes, which is common when contextualising figures (Medisetti et al., 2024); (ii) hypotheses about biased attributes from captions and generated descriptions of images in both correct and incorrect classifications (instead of LLM-generated hypotheses from captions only); and (iii) labels for each hypothesis in a figure.

## 4 Data Annotation

We curate our own multimodal dataset through human-model collaboration to extract, select, and annotate the figures. The curated dataset contains 1,208 annotated maps from 195 publicly available agricultural adaptation journals (CC BY 4.0) published between 1998 and 2024[2], each with an image, caption, discourse, DOI, annotated themes (ground truth answers to classification questions), generated description, and a list of generated attributes; see Table 1.

**Extraction of Multimodal Figure Data** Our domain experts review the literature for representative quality journals on agricultural adaptation in climate science and create a set of domain keywords to include or exclude articles based on titles and texts. Then, an automatic processing (i) identifies peer-reviewed adaptation journals; (ii) selects articles that are publicly available, contain figures, and match adaptation measures and strategies; and (iii) extracts image, caption, and discourse of each figure in the article.

**Selection of Map Figures** MLLM and vision models classify images of extracted scientific figures into map figures. The predictions are then aggregated via ensemble learning (hard voting) and verified on a small annotated sample by a domain expert. We use OpenAI o3 (our choice) and GPT-4o (the best classifier of aerial image objects in GEOBENCH) models by answering a binary classification question from users, which our domain experts define as: *"Does the figure show a map (a representation of the land or water surface area with outlined area boundaries)?"*. We train our own classifier by performing supervised fine-tuning of a vision transformer (ViT Base Patch16 224 IN21k) on map images from the ACL-FIG dataset. Performance in a random sample of 100 annotated figures is the F1 score of 93% by OpenAI o3 (90% precision, 97% recall), 92% by OpenAI GPT-4o (89% precision, 96% recall), and 71% by our fine-tuned vision transformer (75% precision, 68% recall).

| **Figure Type** | Thematic Map | Other Map |
|---|---|---|
| **Count** | 336 | 872 |
| **Size** | 861k bytes [37k, 4M] | 733k bytes [17k, 6M] |
| **Resolution** | 6.2M pixels [110k, 14M] | 5.7M pixels [114k, 16M] |
| **Caption** | 55 words [4, 271] | 37 words [4, 199] |
| | 350 characters [25, 2k] | 236 characters [19, 1k] |
| **Discourse** | 86 words [0, 589] | 55 words [0, 311] |
| | 505 characters [0, 4k] | 331 characters [0, 2k] |

Table 1: Summary of our curated dataset (averages in size and ranges of individual figure characteristics).

**Annotation of Thematic Maps** Our annotation guidelines direct two climate scientist annotators and one computational scientist facilitator to identify thematic maps from the machine-selected map figures by (i) discussing the task and modifying its setup to ensure the annotation instructions are sound [30 minutes]; (ii) individually annotating a randomly selected 30 samples, evaluating the annotations to check for substantial agreement, and discussing the challenges together and adjusting the task [2 hours]; and (iii) individually annotating the full dataset [4 weeks at own pace]. Given the image with caption of a map figure and two examples of correct answers, annotators determine whether it is a thematic map by answering the following question formulated by our domain experts as: *"Does the figure show a climatic parameter, such as atmospheric pressure, evaporation, humidity, precipitation, radiation, temperature, or wind?"*. The facilitator resolves the ties and skipped annotations and labels the individual themes (climatic parameters) of each annotated thematic map. The result is a near-perfect inter-rater agreement of domain experts (91.87% Cohen's kappa coefficient).

The curated dataset can be generalised in other scientific domains across its annotated themes of map figures at various locations. Our pipeline is reproducible with other scientific articles given the reduced workload for human annotation by the semi-automated extraction and selection of figures. We use models to automatically select map figures since manual selection is not feasible, even though models may introduce a selection bias when select-

[2]XML articles with JPEG figures from Elsevier. https://dev.elsevier.com

ing a subset of figures with maps for human annotation. However, models rarely miss maps when aggregating their predictions, omitting only 46 from 16,025 figures. Although the scale of data is relatively small, this is currently the largest collection of map figures compared to other benchmarks with scientific figures (only 53 and 100 maps are in SCIFIBENCH and ACL-FIG datasets, respectively).

## 5 Experiments

The experimental setup for the classification task evaluates MLLMs (configured as in Table 5) in zero- and few-shot learning settings, both with and without Chain-of-Thought (CoT) prompting. To determine which context in the scientific article yields the best classification performance, we investigate two input combinations: the image paired with its caption, and the image paired with both its caption and discourse. Furthermore, to understand the suitability of MLLMs for assisting humans with the identification of research evidence, we align one of the prompts with the annotation instructions given to our domain experts. Specifically, we supply the models with an image, its corresponding caption, and two exemplar answers without CoT.

For few-shot learning, prompts include a positive and negative exemplar. We evaluate in-distribution capabilities by sourcing them directly from our curated dataset to provide models with high-fidelity, domain-specific demonstrations of map figures. The exemplars are selected via stratified random sampling from a pool of 10 expert-curated exemplars with an equal class split, excluding the evaluation instance to prevent data leakage.

For CoT prompting, we experiment with both implicit and explicit commands. The implicit command appends the *"Let's think step by step."* instruction proposed by Kojima et al., 2022 (see Figures 4 and 7). Conversely, the explicit command replicates the reasoning process of our domain experts via a step-by-step procedural instruction. We handcrafted a sequence of reasoning steps required to complete this task, focusing on both the textual and visual modalities of the figures (see Figure 5). For few-shot learning, our exemplars are augmented with the reasoning steps taken following the approach of Wei et al., 2022 (see Figure 8).

We measure the effectiveness of models across a set of classes using F1 scores for fine-grained type problems (Ling and Weld, 2021). The F1 score is computed from precision and recall metrics, which we calculate using strict, loose macro, and loose micro variations. The strict version defines both metrics as the unweighted mean of the exact match between the predicted and labelled set of classes for each instance. The loose macro version calculates the unweighted mean of each metric across classes for each instance. Finally, the loose micro version sums the dividends and divisors across classes for each instance before computing the final quotient.

For the attribute discovery experiments, we explain the predictions made under the best-performing configuration to better understand classifier performance. We prompt each MLLM in a zero-shot setting to generate concise, three-sentence descriptions of the figure images, as well as hypotheses about the target attributes (including attribute names and examples within map figures). The inputs for hypothesis generation consist of randomly selected captions paired with the generated descriptions of figures from both correct and incorrect classifications (100 samples per category). Following established protocols (Yenamandra et al., 2023; Hosseini et al., 2025; Ghosh et al., 2025), we then analyse spurious features in the images by comparing a set of unmodified images containing the target attributes against a distinct set of unmodified images lacking those attributes.

To filter these target attributes for analysis, we establish thresholds for both model performance and sample size; this ensures the discovered attributes represent distinct groups of common visual attributes on maps that demonstrably influence classifier performance. To set the model performance threshold, we select the five attributes that yield the largest absolute changes in classifier effectiveness (measured by the strict F1 score). To set the sample size threshold, we select attributes present in more than 10% of the figures in the classification dataset. Relaxing these filtering criteria would increase the likelihood of sampling errors and surface less representative attributes (see Table 6). For example, the 'blue-red contour' attribute appears in only 4% of the maps and represents a minor subset of the broader 'blue-red gradient' attribute, which captures 21% of the maps.

## 6 Results

The results shown in Table 2 indicate that the highest performance in thematic map identification is an 80% strict F1 score by the Grok 4.3 model. This result was achieved using an image with caption input

| Model | Source | CoT | Zero-Shot Learning | | | | | | Few-Shot Learning | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Image + Caption | | | Image + Caption + Discourse | | | Image + Caption | | | Image + Caption + Discourse | | |
| | | | *Strict* | *Macro* | *Micro* | *Strict* | *Macro* | *Micro* | *Strict* | *Macro* | *Micro* | *Strict* | *Macro* | *Micro* |
| Claude Opus 4.6 | closed | | 60.0 | 66.7 | 61.7 | 35.7 | 42.6 | 39.9 | 69.5 | 75.2 | 71.8 | 62.5 | 69.1 | 65.3 |
| Gemini 3.5 | closed | | 72.4 | 78.0 | 73.3 | 60.5 | 67.1 | 61.5 | 76.3 | 81.1 | 76.0 | 71.0 | 76.6 | 71.7 |
| GPT-5.5 | closed | | 74.5 | 79.3 | 74.6 | 66.8 | 73.1 | 68.1 | 76.3 | 80.8 | 75.9 | 70.0 | 75.4 | 70.0 |
| Grok 4.3 | closed | | 70.1 | 75.2 | 65.3 | 64.4 | 70.8 | 61.2 | 79.9 | 82.3 | 75.7 | 75.5 | 79.3 | 72.6 |
| Kimi K2.5 | open | $\mathcal{X}$ | 39.0 | 46.5 | 42.0 | 28.2 | 36.4 | 32.8 | n/a | n/a | n/a | n/a | n/a | n/a |
| Llama 4 Maverick | open | | 44.8 | 51.3 | 50.4 | 32.9 | 39.3 | 41.0 | 71.0 | 76.9 | 72.3 | 54.1 | 60.6 | 56.3 |
| Gemma 4 | open | | 74.4 | 80.2 | 74.8 | 70.0 | 75.7 | 70.6 | 76.3 | 81.6 | 76.4 | 68.0 | 72.6 | 67.8 |
| DeepSeek V4 | open | | 67.3 | 74.1 | 67.8 | 59.0 | 66.6 | 60.8 | 74.1 | 80.9 | 75.2 | 76.0 | 82.2 | 75.5 |
| Mistral Large 3 | open | | 9.7 | 15.5 | 16.5 | 6.0 | 12.7 | 14.1 | n/a | n/a | n/a | n/a | n/a | n/a |
| Qwen3-VL | open | | 63.8 | 70.0 | 64.7 | 39.6 | 46.5 | 46.5 | n/a | n/a | n/a | n/a | n/a | n/a |
| Claude Opus 4.6 | closed | | 56.3 | 64.2 | 56.8 | 45.9 | 54.1 | 48.9 | 69.9 | 76.0 | 72.1 | 63.7 | 70.4 | 66.3 |
| Gemini 3.5 | closed | | 72.3 | 78.0 | 73.0 | 60.5 | 67.3 | 61.6 | 77.2 | 82.1 | **77.2** | 72.1 | 77.7 | 72.4 |
| GPT-5.5 | closed | | 73.3 | 78.4 | 73.9 | 67.5 | 73.2 | 68.2 | 76.1 | 80.5 | 75.6 | 70.4 | 75.6 | 70.4 |
| Grok 4.3 | closed | | 69.7 | 75.2 | 65.3 | 64.0 | 70.1 | 59.8 | **80.0** | **82.4** | 75.6 | 75.1 | 79.0 | 72.3 |
| Kimi K2.5 | open | $\mathcal{I}$ | 34.2 | 41.4 | 35.4 | 30.0 | 38.6 | 34.0 | n/a | n/a | n/a | n/a | n/a | n/a |
| Llama 4 Maverick | open | | 40.5 | 47.2 | 47.4 | 29.8 | 36.1 | 37.9 | 69.9 | 76.0 | 71.6 | 49.8 | 56.1 | 52.4 |
| Gemma 4 | open | | 72.3 | 78.1 | 73.0 | 68.7 | 74.6 | 69.8 | 75.1 | 80.5 | 76.1 | 67.2 | 72.0 | 67.3 |
| DeepSeek V4 | open | | 66.8 | 74.0 | 67.5 | 58.8 | 66.2 | 58.7 | 71.5 | 79.0 | 73.0 | 74.4 | 81.1 | 74.2 |
| Mistral Large 3 | open | | 8.8 | 15.0 | 16.2 | 6.9 | 13.2 | 14.6 | n/a | n/a | n/a | n/a | n/a | n/a |
| Qwen3-VL | open | | 67.1 | 72.9 | 66.5 | 47.1 | 54.0 | 51.6 | n/a | n/a | n/a | n/a | n/a | n/a |
| Claude Opus 4.6 | closed | | 31.5 | 33.4 | 28.2 | 21.5 | 22.8 | 18.0 | 70.9 | 76.6 | 72.2 | 66.3 | 72.4 | 67.0 |
| Gemini 3.5 | closed | | 70.9 | 76.5 | 70.9 | 61.9 | 68.8 | 63.3 | 76.7 | 81.5 | 76.6 | 70.3 | 75.9 | 70.8 |
| GPT-5.5 | closed | | 71.4 | 76.2 | 70.6 | 64.0 | 70.2 | 65.2 | 74.2 | 78.6 | 73.4 | 68.0 | 73.7 | 68.3 |
| Grok 4.3 | closed | | 71.2 | 76.0 | 67.8 | 61.8 | 68.4 | 59.0 | 78.1 | 80.9 | 74.4 | 72.0 | 76.2 | 69.2 |
| Kimi K2.5 | open | $\mathcal{E}$ | 38.8 | 45.5 | 39.8 | 42.1 | 49.6 | 42.7 | n/a | n/a | n/a | n/a | n/a | n/a |
| Llama 4 Maverick | open | | 29.1 | 35.5 | 38.6 | 18.9 | 25.6 | 30.0 | 62.3 | 68.5 | 65.7 | 28.1 | 34.9 | 34.5 |
| Gemma 4 | open | | 67.7 | 74.0 | 68.8 | 63.0 | 69.5 | 65.1 | 56.6 | 61.2 | 59.4 | 10.7 | 12.3 | 13.4 |
| DeepSeek V4 | open | | 49.3 | 56.9 | 49.3 | 43.0 | 51.3 | 44.6 | 65.1 | 72.6 | 66.1 | 64.2 | 72.0 | 64.8 |
| Mistral Large 3 | open | | 17.9 | 23.3 | 22.9 | 21.7 | 27.0 | 25.2 | n/a | n/a | n/a | n/a | n/a | n/a |
| Qwen3-VL | open | | 25.2 | 32.0 | 41.4 | 26.9 | 34.4 | 40.7 | n/a | n/a | n/a | n/a | n/a | n/a |

Table 2: Classification of themes in map figures using input context of image with caption or image with caption and discourse via zero- and few-shot learning without ($\mathcal{X}$) or with Implicit ($\mathcal{I}$) or Explicit ($\mathcal{E}$) Chain-of-Thought (CoT) prompting. Evaluation metrics are strict, loose macro, and loose micro F1 scores (n/a due to model limits).

under few-shot learning with implicit CoT prompting, which we found to be the best configuration for multi-label classification across MLLMs supporting multimodal few-shot exemplars. However, the performance does not align well with human judgement as classifiers are generally unable to follow human-prescribed reasoning steps (explicit CoT) and only self-generated reasoning pathways (implicit CoT) improve their performance.

MLLMs struggle to classify map figures with blue-red colours (56% by DeepSeek V4) and multiple plots (64% by Llama 4 Maverick) as per Table 3. A further analysis of the classifications reveals that the performance of models also drops when removing the figures with visual attributes on which the models are biased. Such attributes are scale bars, north arrows, and river outlines that represent numerical trends, spatial directions, and physical waterway flows on maps, respectively.

**Q1: Does relevant text in scientific articles improve the identification of thematic map figures by MLLMs?** Classifiers perform the thematic map identification task best without a discourse in their input, relying solely on images paired with their captions. This text context generally does not improve performance and only a few MLLMs benefit from its inclusion. Specifically, Kimi K2.5, Mistral Large 3, and Qwen3-VL yield minor gains of up to 3 points of strict F1 scores in zero-shot learning with explicit CoT prompting as per Table 2. This discourse-related degradation is not due to context window constraints, as it persists under the reduced context of zero-shot learning, where classifiers processed complete inputs.

Our hypothesis is that scientific knowledge in discourse becomes too complex for MLLMs to interpret accurately when linked to the already dense context within the image of a map figure. For example, only domain experts may understand whether the scientific discussion about adaptation strategies in discourse is relevant to particular themes of climatic parameters in a map figure, since such themes are sometimes not explicitly mentioned in any input of image, caption, or discourse. MLLMs may lack this implicit domain knowledge to bridge such modal gaps. In addition, an image with a caption generally contains richer and more closely related information about a thematic map figure than any other relevant text in an article (discourse).

Our work builds upon ablation studies examining the input context of figures provided to multimodal systems, demonstrating that combining images with captions outperforms using images alone for question answering over scientific articles (Pra-

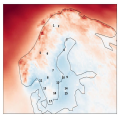
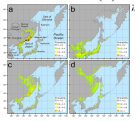
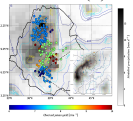
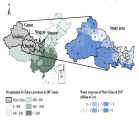
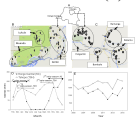
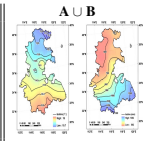
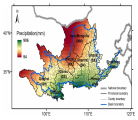
| Model | Source | Attribute | | | | | | | | | | Attribute Union | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Blue-Red Gradient (A) | | Multi-Panel (B) | | Scale Bar (C) | | North Arrow (D) | | River Outline (E) | | A ∪ B | | C ∪ D ∪ E | |
| | |  | |  | |  | |  | |  | |  | |  | |
| | | 137 thematic maps<br>118 other maps | | 297 thematic maps<br>524 other maps | | 149 thematic maps<br>541 other maps | | 146 thematic maps<br>496 other maps | | 39 thematic maps<br>192 other maps | | 310 thematic maps<br>547 other maps | | 178 thematic maps<br>624 other maps | |
| | | *Present* | *Absent* | *Present* | *Absent* | *Present* | *Absent* | *Present* | *Absent* | *Present* | *Absent* | *Present* | *Absent* | *Present* | *Absent* |
| Claude Opus 4.6 | closed | 57.6 | 73.1 | 64.4 | 81.4 | 75.2 | 62.7 | 74.6 | 64.5 | 73.2 | 69.1 | 64.6 | 82.6 | 74.1 | 61.6 |
| Gemini 3.5 | closed | 67.5 | 79.9 | 72.8 | 86.6 | 80.3 | 73.2 | 81.3 | 72.6 | 80.5 | 76.5 | 72.8 | 88.0 | 80.0 | 71.1 |
| GPT-5.5 | closed | 68.6 | 78.1 | 72.2 | 84.2 | 78.4 | 73.0 | 78.7 | 73.1 | 78.8 | 75.4 | 72.2 | 85.5 | 78.2 | 71.9 |
| Grok 4.3 | closed | 70.6 | 82.5 | 75.5 | 89.4 | 81.9 | 76.8 | 82.9 | 76.7 | 84.0 | 79.0 | 75.5 | 90.9 | 82.2 | 75.6 |
| Llama 4 Maverick | open | 56.2 | 73.6 | 63.7 | 82.9 | 75.5 | 62.4 | 76.5 | 62.4 | 76.2 | 68.4 | 63.9 | 84.3 | 74.8 | 60.1 |
| Gemma 4 | open | 63.9 | 78.1 | 70.5 | 84.8 | 79.9 | 68.7 | 80.4 | 69.1 | 81.0 | 73.7 | 70.6 | 86.0 | 79.8 | 65.8 |
| DeepSeek V4 | open | 56.1 | 75.7 | 64.8 | 85.8 | 76.7 | 64.7 | 77.3 | 65.0 | 74.9 | 70.7 | 65.3 | 86.6 | 76.1 | 62.6 |
| *influence on performance* | | unfavourable | | unfavourable | | favourable | | favourable | | favourable | | unfavourable | | favourable | |

Table 3: Explanation of classification performance showing discovered attributes when present or absent in map figures, including maps with multiple sets of attributes (union of maps) and their influence on performance. Evaluation of best performing configuration via strict F1 score (few-shot learning with implicit chain-of-thought prompting on input context of images with captions).

manick et al., 2024). Collecting a caption and reference text about a figure in an article is standard practice in processing scientific figures (Sundar et al., 2024), but existing studies analysing fine-grained citation contexts overlook figure citations (Jantsch et al., 2025). Addressing this gap is essential to better contextualise the associated scientific discourse within articles. To avoid irrelevant context in our collected discourse, we follow the established protocol of extracting only the citing sentences rather than the broader surrounding text, which may include unrelated details about the figures.

**Q2: Can examples of map figures with correct answers improve the classification performance of MLLMs?** Performance of MLLMs improves when given the same instructions as annotators, i.e., few-shot examples of images with captions and answers to classification questions (compared to zero-shot). This is consistent with human requirements, as our domain experts need examples of answers to correctly perform the annotation task. Few-shot learning provides a maximum increase of 45 strict F1 points, achieved by Claude Opus 4.6 under explicit CoT prompting using images with captions and discourse (Table 2).

The ability of MLLMs to process multiple of multimodal examples and with high resolution and large size of images is limited. Models struggle with an increased volume of complex information in few-shot learning and are generally unable to process more than two examples due to their input context limits for both textual and visual modalities. Some models can process only one map image in their input, provided that it is small in size and low in resolution. For example, Kimi K2.5, Mistral Large 3, and Qwen3-VL cannot process inputs for our few-shot learning experiments and Claude Opus 4.6 limits the image size to 5MB. However, we did not reduce the size of the images as this reduced the classification performance of MLLMs.

Interestingly, the performance of MLLMs in interpreting figures may decrease when few-shot examples are included in the prompts (Roberts et al., 2024). We have only noticed such behaviour for the Gemma 4 model and mainly under its input of image with caption and discourse when going from zero-shot to few-shot learning. Fine-tuning and CoT prompting can mitigate some of the challenges of MLLMs in processing multimodal data for in-context learning (Zeng et al., 2024).

**Q3: Are MLLM classifiers biased toward visual attributes of map figures?** The goal is to understand visual attributes of map figures that lead to reliable classifications in practice. Therefore, we analysed attributes that influence results of MLLMs and validated whether these classifiers are biased toward such attributes. Table 3 shows an analysis of discovered attributes that are shared across MLLMs (self-reported by each model) for the best performing configuration of few-shot learning with implicit CoT prompting on images with captions.

We observed that the best classifier (Grok 4.3) correctly identified thematic maps when given map figures that show scale bars, north arrows, and river outlines. These attributes appear in about half of the thematic maps (178 out of 336). However, classification performance drops by nearly 7 points on the strict F1 score for figures without such visual attributes (representing the remaining half of the thematic maps). Our domain experts verified

| Model | Source | Theme | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Atmospheric Pressure** 10 map figures [1 with 1 theme, 4 with 2 themes, 4 with 3 themes, 1 with 4 themes] | | | **Evaporation** 118 map figures [13 with 1 theme, 22 with 2 themes, 37 with 3 themes, 46 with 4 themes] | | | **Humidity** 54 map figures [0 with 1 theme, 3 with 2 themes, 6 with 3 themes, 45 with 4 themes] | | | **Precipitation** 251 map figures [67 with 1 theme, 95 with 2 themes, 39 with 3 themes, 50 with 4 themes] | | | **Radiation** 36 map figures [8 with 1 theme, 8 with 2 themes, 14 with 3 themes, 6 with 4 themes] | | | **Temperature** 205 map figures [32 with 1 theme, 75 with 2 themes, 48 with 3 themes, 50 with 4 themes] | | | **Wind** 16 map figures [11 with 1 theme, 1 with 2 themes, 2 with 3 themes, 2 with 4 themes] | | |
| | | *P* | *R* | *F1* | *P* | *R* | *F1* | *P* | *R* | *F1* | *P* | *R* | *F1* | *P* | *R* | *F1* | *P* | *R* | *F1* | *P* | *R* | *F1* |
| Claude Opus 4.6 | closed | **83.3** | **50.0** | **62.5** | 54.9 | 52.5 | 53.7 | 21.4 | 16.7 | 18.8 | 60.1 | **86.5** | **70.9** | 41.4 | **66.7** | 51.1 | 57.3 | 72.7 | **64.1** | 36.6 | **93.8** | 52.6 |
| Gemini 3.5 | closed | **83.3** | **50.0** | **62.5** | 73.1 | 41.5 | 53.0 | 31.0 | **16.7** | **21.7** | 68.7 | 73.3 | **70.9** | **55.9** | 52.8 | 54.3 | **66.8** | 61.0 | 63.8 | 41.7 | **93.8** | 57.7 |
| GPT-5.5 | closed | **83.3** | **50.0** | **62.5** | 65.4 | 44.9 | 53.3 | 18.2 | 11.1 | 13.8 | 66.8 | 73.7 | 70.1 | 55.3 | 58.3 | **56.8** | 66.5 | 56.1 | 60.8 | 41.7 | **93.8** | 57.7 |
| Grok 4.3 | closed | 71.4 | 50.0 | 58.8 | **69.4** | 36.4 | 47.8 | **42.9** | 11.1 | 17.6 | **69.3** | 56.6 | 62.3 | 54.3 | 52.8 | 53.5 | 65.9 | 54.6 | 59.7 | **50.0** | **93.8** | **65.2** |
| Llama 4 Maverick | open | **83.3** | **50.0** | **62.5** | 63.2 | 36.4 | 46.2 | 27.8 | 9.3 | 13.9 | 61.7 | 75.7 | 68.0 | 48.6 | 50.0 | 49.3 | 51.3 | 75.6 | 61.1 | 42.9 | **93.8** | 58.8 |
| Gemma 4 | open | **83.3** | **50.0** | **62.5** | 59.2 | **60.2** | **59.7** | 38.9 | 13.0 | 19.4 | 63.4 | 80.1 | 70.8 | 55.6 | 55.6 | 55.6 | 63.4 | 62.4 | 62.9 | 48.4 | **93.8** | 63.8 |
| DeepSeek V4 | open | 66.7 | 20.0 | 30.8 | 68.6 | 29.7 | 41.4 | 13.6 | 5.6 | 7.9 | 59.0 | 83.7 | 69.2 | 40.5 | 47.2 | 43.6 | 48.7 | **88.8** | 62.9 | 40.0 | 75.0 | 52.2 |

Table 4: Analysis of binary classification errors per individual themes in map figures (each map can have multiple themes). Evaluation of best performing configuration via Precision (P), Recall (R), and F1 score (few-shot learning with implicit chain-of-thought prompting on input context of images with captions).

that these attributes are irrelevant to determining whether a map is a thematic map, since they do not relate to climatic parameter themes, and the model is biased toward such attributes of map figures.

Our analysis identifies a distinct model bias induced by a visual shortcut, in which MLLMs exploit spurious, non-causal associations between the class of thematic maps and the broader figure category of maps. Selection bias in classification datasets can amplify this observation when models omit difficult samples and retain only easy examples during data selection. Our curated dataset overcomes this issue by using a high-accuracy model ensemble to select a comprehensive collection of maps from the initial set of extracted figures. Although we verified a subset of the MLLM-generated descriptions for quality, we did not fully assess the faithfulness of these descriptions to the visual content of maps, nor did we measure how accurately MLLMs discover visual attributes or their robustness to prompt variance.

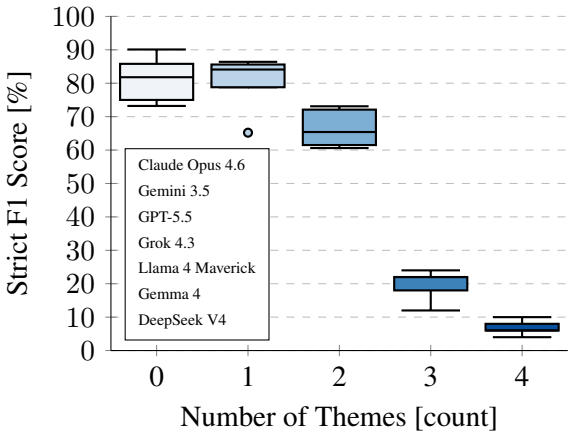



Figure 2: Analysis of multi-label classification errors per number of themes in map figures (0 themes in 872 maps, 1 in 132, 2 in 104, 3 in 50, and 4 in 50). Evaluation of best performing configuration via strict F1 score (few-shot learning with implicit chain-of-thought prompting on input context of images with captions).

## 7 Error Analysis

The error analysis reveals challenges for MLLMs in understanding labels of climatic parameters and classifying multiple labels. We examined the results of the best performing configuration, which is few-shot learning with implicit CoT on the input context of images with captions. The classification performance per individual themes is shown in Table 4 and per number of themes in Figure 2.

The analysis of binary classification errors shows that the models cannot classify certain climatic parameters. The best F1 score is only 22% on thematic maps showing a single climatic parameter of humidity (see Table 4). For example, MLLMs misclassify 'precipitable water' data in thematic maps as 'precipitation', which domain experts classify as 'humidity' and 'atmospheric pressure' as these are the climatic parameters used to calculate the liquid water depth from condensed water vapour in the atmosphere. Our observation is that MLLMs do not correctly link climate science terminology explicitly stated in a caption or image of a map figure to climatic parameters. Other example is the 'RCP' term, which denotes Representative Concentration Pathways by amount of radiative forcing in near future, but models do not link it to (implicitly stated) 'radiation' climatic parameter and misclassify this as a 'none' theme. This demonstrates that MLLMs do not understand climatic parameters well, most likely due to the lack of data and instructions focused on numerical reasoning with climatic parameters during their training.

The analysis of multi-label classification errors reveals a performance decline with an increasing number of themes in the map figures. As illustrated in Figure 2, the evaluated MLLMs demonstrate a

strong baseline capability for single-label classification (zero or one theme), achieving median strict F1 scores over 80% for individual labels. However, performance decreases as task complexity increases, dropping to a median of approximately 65% for two labels, 20% for three labels, and under 10% for four labels. This 45 percentage point drop between two and three labels indicates that multi-label classification of multimodal data is a challenging task for MLLMs.

## 8 Conclusion

We evaluated whether MLLMs can identify thematic maps in scientific literature. This is a crucial step towards designing an information seeking system that can access different sources of multimodal data, such as climate data and text and figures from scientific articles, in answering complex questions on climate adaptation. We show that off-the-shelf MLLMs do not perform well on the classification of thematic maps found in scientific figures of maps. This poses risks as MLLMs are increasingly adopted as multi-purpose models in high-stakes applications, such as AI systems advising on climate adaptation. We find that classification performance is best when MLLMs are supplied with figures with related captions and human-verified examples of good classifications. Surprisingly, any additional context, such as a sentence referring to the respective figure, leads to worse performance. Further research is required to improve MLLM classification of multimodal scientific figures, investigate their ability to understand complex visual-textual context, such as one presented in thematic maps, and control and mitigate their classification bias.

## Limitations

The task formulation is limited to classification to establish the baseline performance of MLLMs in identifying thematic maps. Another practical evaluation for a downstream use of MLLMs on such a dataset of thematic maps should follow. For example, question-answering for a deeper understanding of content in map figures.

The scope of our work is limited to maps in the practice of climate adaptation. Although map figures in the curated dataset contain other themes, such as land use and agricultural commodities, only climatic parameters are annotated. Collecting articles from various domains and annotating their map figures with themes would clarify the generalisability of thematic maps across different themes. For example, future work can investigate whether MLLMs perform similarly when classifying climatic parameters versus healthcare indicators.

## Ethical Considerations

Only publicly available data were used in this research, and no additional data were collected. The authors obtained ethics approval from CSIRO for user studies with human experts, including their annotations (Climate Services for Agriculture, Phase 2, 131/24).

## Acknowledgements

The authors gratefully acknowledge Willow Hallgren, Ashley Harkin, Vihan Weeraratne, and Mahesh Prakash for their agricultural and climate science expertise and their valuable contributions to the think-aloud studies. Tomas Turek is supported by the CSIRO PhD scholarship.

## References

Andres Alonso-Robisco, Jose Manuel Carbo, Emily Kormanyos, and Elena Triebskorn. 2026. Houston, we have a problem: Can satellite information bridge the climate-related data gap? *Latin American Journal of Central Banking (LAJCB)*, 7(1):100173.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. MapQA: A dataset for question answering on choropleth maps. In *NeurIPS 2022 First Table Representation (TRL) Workshop*, New Orleans, Louisiana, United States.

Muhammad Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. 2025. GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7132–7142, Honolulu, Hawaii, United States.

Alberto d'Onofrio and Elaheh Pourabbas. 2001. Formalization of temporal thematic map contents. In *Proceedings of the 9th ACM International Symposium on Advances in Geographic Information Systems*, GIS '01, page 15–20, New York City, New York, United States. Association for Computing Machinery (ACM).

Giles M. Foody. 2004. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering & Remote Sensing*, 70(5):627–633.

Shantanu Ghosh, Rayan Syed, Chenyu Wang, Vaibhav Choudhary, Binxu Li, Clare B Poynton, Shyam

Visweswaran, and Kayhan Batmanghelich. 2025. LADDER: Language-driven slice discovery and error rectification in vision classifiers. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22935–22970, Vienna, Austria. Association for Computational Linguistics (ACL).

Zhipeng Gui, Xinjie Liu, Anqi Zhao, Yuhan Jiang, Zhipeng Ling, Xiaohui Hu, Fa Li, Zelong Yang, Huayi Wu, and Shuangming Zhao. 2024. Map retrieval intention recognition based on relevance feedback and geographic semantic guidance: For better understanding user retrieval demands. *Information Processing & Management*, 61(4):103767.

Parsa Hosseini, Sumit Nawathe, Mazda Moayeri, Sriram Balasubramanian, and Soheil Feizi. 2025. SpurLens: Spurious correlations in multimodal LLMs. In *Workshop on Spurious Correlation and Shortcut Learning (SCSL): Foundations and Solutions*, Singapore. The Thirteenth International Conference on Learning Representations (ICLR).

Lasse M. Jantsch, Dong-Jae Koh, Seonghwan Yoon, Jisu Lee, Anne Lauscher, and Young-Kyoon Suh. 2025. FineCite: A novel approach for fine-grained citation context analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24525–24542, Vienna, Austria. Association for Computational Linguistics (ACL).

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. 2023. ACL-Fig: A dataset for scientific figure classification. In *Proceedings of the Workshop on Scientific Document Understanding (SDU) co-located with 37th Conference on Artificial Inteligence (AAAI)*, volume 3656, Washington, DC, United States. CEUR-WS.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 22199–22213. Curran Associates, Inc.

Xiao Ling and Daniel Weld. 2021. Fine-grained entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI Technical Track: AI and the Web)*, volume 26, page 94–100, Toronto, Ontario, Canada. The Association for the Advancement of Artificial Intelligence (AAAI) Press.

Noemi Mauro and Liliana Ardissono. 2017. Thematic maps for geographical information search. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, UMAP '17, page 337–342, New York City, New York, United States. Association for Computing Machinery (ACM).

Gowtham Medisetti, Zacchaeus Compson, Heng Fan, Huaxiao Yang, and Yunhe Feng. 2024. LitAI: Enhancing multimodal literature understanding and mining with generative AI. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 471–476, San Jose, California, United States. Institute for Electrical and Electronics Engineers (IEEE).

Mark Monmonier. 2001. Thematic maps in geography. In Neil J. Smelser and Paul B. Baltes, editors, *International Encyclopedia of the Social Behavioral Sciences*, pages 15636–15641. Pergamon, Oxford, United Kingdom.

Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, and Mahesh Prakash. 2025. Question answering in climate adaptation for agriculture: Model development and evaluation with expert feedback. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7045–7075, Vienna, Austria. Association for Computational Linguistics (ACL).

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. SPIQA: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 118807–118833, Vancouver, Canada. Curran Associates, Inc.

Jingfen Qiao, Jia-Huei Ju, Xinyu Ma, Evangelos Kanoulas, and Andrew Yates. 2025. Reproducibility, replicability, and insights into visual document retrieval with late interaction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3335–3345, New York City, New York, United States. Association for Computing Machinery (ACM).

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. SciFIBench: Benchmarking large multimodal models for scientific figure interpretation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 18695–18728, Vancouver, Canada. Curran Associates, Inc.

Robert A. Schowengerdt. 2007. Chapter 9 - thematic classification. In *Remote Sensing (Third Edition)*, third edition edition, pages 387–XXXIII. Academic Press, Burlington, Massachusetts, United States.

Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. 2018. Extracting scientific figures with distantly supervised neural networks. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 223–232, New York City, New York, United States. Association for Computing Machinery (ACM).

Anirudh Sundar, Jin Xu, William Gay, Christopher Richardson, and Larry Heck. 2024. cPAPERS: A dataset of situated and multimodal interactive conversations in scientific papers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 66283–66304, Vancouver, Canada. Curran Associates, Inc.

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi, and Dinesh Manocha. 2025.

ChartLens: Fine-grained visual attribution in charts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22447–22462, Vienna, Austria. Association for Computational Linguistics (ACL).

Eleni Tousidou and Yannis Manolopoulos. 2000. A performance comparison of quadtree-based access methods for thematic maps. In *Proceedings of the 2000 ACM Symposium on Applied Computing - Volume 1*, SAC '00, page 381–388, New York City, New York, United States. Association for Computing Machinery (ACM).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837. Curran Associates, Inc.

Zhengyang Wei, Zhipeng Gui, Min Zhang, Zelong Yang, Yuao Mei, Huayi Wu, Hongbo Liu, and Jing Yu. 2021. Text GCN-SW-KNN: a novel collaborative training multi-label classification method for WMS application themes by considering geographic semantics. *Big Earth Data*, 5(1):66–89.

WMO. 2025. The 2022 GCOS ECVs Requirements. Technical Report Global Climate Observing System (GCOS) Report-245, World Meteorological Organization (WMO), Geneva, Switzerland.

Zelong Yang, Zhipeng Gui, Huayi Wu, and Wenwen Li. 2020. A latent feature-based multimodality fusion method for theme classification on web map service. *Institute for Electrical and Electronics Engineers (IEEE) Access*, 8:25299–25309.

Sriram Yenamandra, Pratik Ramesh, Viraj Prabhu, and Judy Hoffman. 2023. FACTS: First amplify correlations and then slice to discover bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4794–4804, Paris, France. Institute for Electrical and Electronics Engineers (IEEE).

Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. 2024. Can MLLMs perform text-to-image in-context learning? In *First Conference on Language Modeling (COLM)*, Philadelphia, Pennsylvania, United States.

## A Appendix

This section details the model configurations, prompts, and additional evaluation results. The aim is to support the reproducibility of our empirical findings on thematic map identification within map figures, which are multimodal scientific figures sourced from scientific articles.

| Model | Source | Provider | Release | Size | Context Window | Configuration |
|---|---|---|---|---|---|---|
| Claude Opus 4.6 V1 | closed | Anthropic (United States) | 5 February 2026 | undisclosed | 1M tokens | reasoning effort: medium<br>retry limit: 3<br>temperature: 0 |
| Gemini 3.5 Flash | closed | Google DeepMind (United Kingdom) | 19 May 2026 | undisclosed | 1M tokens | reasoning effort: medium<br>retry limit: 3<br>temperature: 0 |
| GPT-5.5 | closed | OpenAI (United States) | 23 April 2026 | undisclosed | 1M tokens | reasoning effort: medium<br>retry limit: 3 |
| Grok 4.3 | closed | SpaceXAI (United States) | 17 April 2026 | undisclosed | 1M tokens | reasoning effort: medium<br>retry limit: 3<br>temperature: 0 |
| Kimi K2.5 | open | Moonshot AI (China) | 27 January 2026 | 1T parameters (32b active) | 256k tokens | retry limit: 3 |
| Llama 4 Maverick 17b Instruct V1 | open | Meta AI (United States) | 25 May 2025 | 400b parameters (17b active) | 1M tokens | retry limit: 3<br>temperature: 0 |
| Gemma 4 31B IT | open | Google DeepMind (United Kingdom) | 2 April 2026 | 31b parameters (31b active) | 256k tokens | reasoning effort: medium<br>retry limit: 3<br>temperature: 0 |
| DeepSeek V4 Flash | open | DeepSeek (China) | 31 June 2026 | 284b parameters (13b active) | 1M tokens | reasoning effort: medium<br>retry limit: 3<br>temperature: 0 |
| Mistral Large 3 675b Instruct | open | Mistal AI (France) | 2 December 2025 | 675b parameters (41b active) | 256k tokens | retry limit: 3<br>temperature: 0 |
| Qwen3 VL 235B A22B | open | Alibaba Cloud (China) | 23 September 2025 | 235b parameters (22b active) | 256k tokens | retry limit: 3<br>temperature: 0 |

Table 5: Overview of benchmarked models on our multimodal classification dataset, including their configurations.

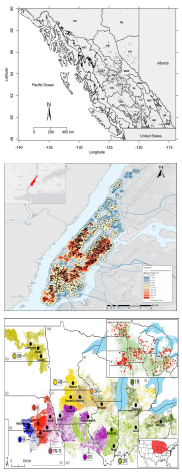
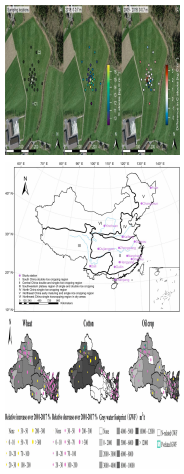
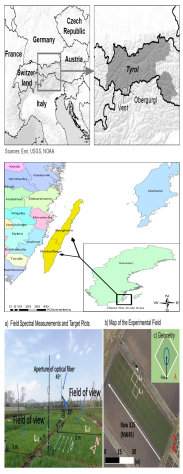
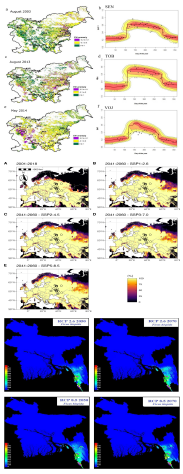
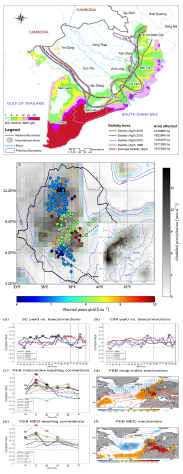
| Model | Source | Attribute | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Black Dot | | Pink Dot | | Flow Arrow | | Year Label | | Blue-Red Contour | |
| | |  | |  | |  | |  | |  | |
| | | 69 thematic maps<br>150 other maps | | 5 thematic maps<br>29 other maps | | 20 thematic maps<br>104 other maps | | 59 thematic maps<br>85 other maps | | 16 thematic maps<br>33 other maps | |
| | | *Present* | *Absent* | *Present* | *Absent* | *Present* | *Absent* | *Present* | *Absent* | *Present* | *Absent* |
| Claude Opus 4.6 | closed | 72.6 | 69.3 | 67.7 | 69.9 | 70.2 | 69.8 | 68.8 | 70.0 | 71.4 | 69.8 |
| Gemini 3.5 | closed | 76.7 | 77.4 | 79.4 | 77.2 | 79.0 | 77.0 | 74.3 | 77.6 | 77.6 | 77.2 |
| GPT-5.5 | closed | 77.2 | 75.8 | 70.6 | 76.2 | 78.2 | 75.8 | 73.6 | 76.4 | 77.6 | 76.0 |
| Grok 4.3 | closed | 76.7 | 80.7 | 82.4 | 79.9 | 82.3 | 79.7 | 77.8 | 80.3 | 81.6 | 79.9 |
| Llama 4 Maverick | open | 73.1 | 69.2 | 73.5 | 69.8 | 72.6 | 69.6 | 66.7 | 70.3 | 73.5 | 69.7 |
| Gemma 4 | open | 76.3 | 74.8 | 73.5 | 75.1 | 77.4 | 74.8 | 72.9 | 75.4 | 78.0 | 75.0 |
| DeepSeek V4 | open | 75.8 | 70.6 | 73.5 | 71.5 | 73.4 | 71.3 | 68.1 | 72.0 | 69.4 | 71.6 |

Table 6: Explanation of classification performance showing discovered attributes when present or absent in map figures, specifically attributes that yield the lowest absolute changes in classifier effectiveness. Evaluation of best performing configuration via strict F1 score (few-shot learning with implicit chain-of-thought prompting on input context of images with captions).

**Classification Prompt: Zero-Shot In-Context Learning**

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response.

{Scientific Figure Image}
{Scientific Figure Caption}

RESPONSE:

Figure 3: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for zero-shot in-context learning without chain-of-thought prompting.

**Classification Prompt: Zero-Shot In-Context Learning with Implicit Chain-of-Thought Prompting**

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response. Let's think step by step.

{Scientific Figure Image}
{Scientific Figure Caption}

RESPONSE:

Figure 4: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for zero-shot in-context learning with implicit chain-of-thought prompting.

**Classification Prompt: Zero-Shot In-Context Learning with Explicit Chain-of-Thought Prompting**

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Perform the below reasoning steps in the same order to correctly respond.

1. Read the provided caption of a scientific figure and locate the possible options for the answer, including both explicitly stated options and options implied in the caption.
2. Look at the provided image of a scientific figures and locate the possible options for the answer, including both explicitly stated options and options implied in the image.
3. Aggregate your located answer options from the caption and image of a scientific figure.
4. Select unique options for the answer from your aggregated options and remove the 'h) none' option if other unique options are present.
5. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response.

{Scientific Figure Image}
{Scientific Figure Caption}

RESPONSE:

Figure 5: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for zero-shot in-context learning with explicit chain-of-thought prompting.

Classification Prompt: Few-Shot In-Context Learning

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Consider the provided examples of correct responses. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response.

{Scientific Figure Image}
{Scientific Figure Caption}

The below are examples of correct responses.

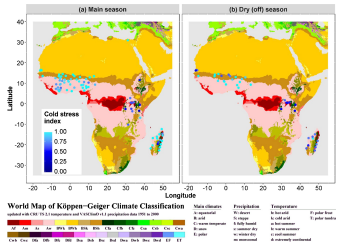



Fig. 5. Simulated cold stress: (a) main season and (b) dry (off) season. Dots are sites for which cold stress was simulated using 5 years (irrigated rice) or 9 years (rainfed rice) of weather data for each dot. The cold stress index is SFCOLD, i.e. the cold-induced fertility: a low value means severe loss due to cold. Köppen–Geiger climate zones are shown in the background, legend and colour scheme from Kottek et al. (2006).

The response is: [‘precipitation’, ‘temperature’]

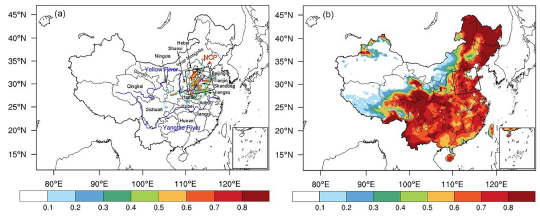



Fig. 1. (a) Model domain and irrigation fraction in the North China Plain (NCP) region, and (b) prescribed green vegetation fraction (GVF) averaged over growing season in China. The black empty box indicates the area of NCP. Several related provinces are marked in Fig. 1a.

The response is: [‘none’]

RESPONSE:

Figure 6: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for few-shot in-context learning without chain-of-thought prompting.

Classification Prompt: Few-Shot In-Context Learning with Implicit Chain-of-Thought Prompting

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Consider the provided examples of correct responses. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response. Let’s think step by step.

{Scientific Figure Image}
{Scientific Figure Caption}

The below are examples of correct responses.

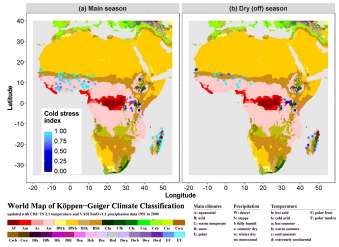



Fig. 5. Simulated cold stress: (a) main season and (b) dry (off) season. Dots are sites for which cold stress was simulated using 5 years (irrigated rice) or 9 years (rainfed rice) of weather data for each dot. The cold stress index is SFCOLD, i.e. the cold-induced fertility: a low value means severe loss due to cold. Köppen–Geiger climate zones are shown in the background, legend and colour scheme from Kottek et al. (2006).

The response is: [‘precipitation’, ‘temperature’]

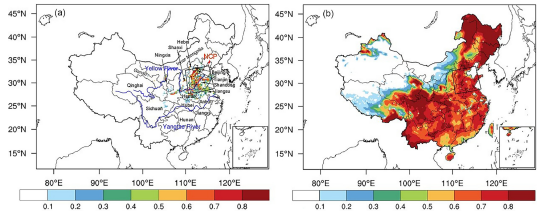



Fig. 1. (a) Model domain and irrigation fraction in the North China Plain (NCP) region, and (b) prescribed green vegetation fraction (GVF) averaged over growing season in China. The black empty box indicates the area of NCP. Several related provinces are marked in Fig. 1a.

The response is: [‘none’]

RESPONSE:

Figure 7: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for few-shot in-context learning with implicit chain-of-thought prompting.

Classification Prompt: Few-Shot In-Context Learning with Explicit Chain-of-Thought Prompting

INSTRUCTION: Given the image and caption of a scientific figure, select all relevant themes of climatic parameters that apply: a) atmospheric pressure, b) evaporation, c) humidity, d) precipitation, e) radiation, f) temperature, and g) wind or h) none. Consider the provided examples of correct responses. Perform the below reasoning steps in the same order to correctly respond.

1. Read the provided caption of a scientific figure and locate the possible options for the answer, including both explicitly stated options and options implied in the caption.
2. Look at the provided image of a scientific figures and locate the possible options for the answer, including both explicitly stated options and options implied in the image.
3. Aggregate your located answer options from the caption and image of a scientific figure.
4. Select unique options for the answer from your aggregated options and remove the 'h) none' option if other unique options are present.
5. Respond with a list of items, which are the full names of your selected options for the answer, in a Python format and do not include anything else in your response.

{Scientific Figure Image}
{Scientific Figure Caption}

The below are examples of correct responses.

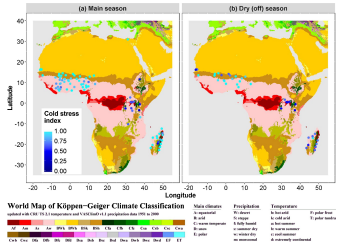



Fig. 5. Simulated cold stress: (a) main season and (b) dry (off) season. Dots are sites for which cold stress was simulated using 5 years (irrigated rice) or 9 years (rainfed rice) of weather data for each dot. The cold stress index is SFCOLD, i.e. the cold-induced fertility: a low value means severe loss due to cold. Köppen–Geiger climate zones are shown in the background, legend and colour scheme from Kottek et al. (2006).

The performed reasoning steps are:

1. The possible options for the answers are 'precipitation' and 'temperature'. Although there are no explicitly stated options when reading the caption, the 'temperature' is implied in the mention of 'cold stress' climate science term that represents strain caused by prolonged exposure to low temperatures and temperature drops. Also, both 'temperature' and 'precipitation' is implied in the 'Köppen–Geiger' climate science term that classifies Earth's climates based on temperature and precipitation.
2. The possible options for the answers are 'precipitation' and 'temperature', which are both explicitly stated options when looking at the image.
3. The aggregated answer options are 'precipitation' and 'temperature' from the caption and 'precipitation' and 'temperature' from the image of a scientific figure.
4. The selected unique options for the answer are 'precipitation' and 'temperature' after removing the duplicated options from the aggregated answer options.
5. The response in Python format for a list of items is ['precipitation', 'temperature'].

The response is: ['precipitation', 'temperature']

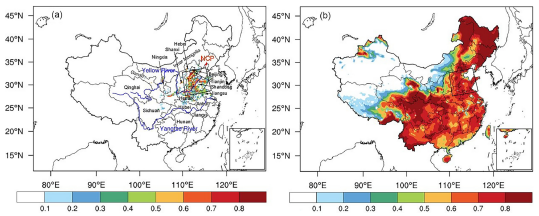



Fig. 1. (a) Model domain and irrigation fraction in the North China Plain (NCP) region, and (b) prescribed green vegetation fraction (GVF) averaged over growing season in China. The black empty box indicates the area of NCP. Several related provinces are marked in Fig. 1a.

The performed reasoning steps are:

1. The possible options for the answers are 'none' as there are no explicitly or implicitly stated options when reading the caption.
2. The possible options for the answers are 'none' as there are no explicitly or implicitly stated options when looking at the image.
3. The aggregated answer options are 'none' from the caption and 'none' from the image of a scientific figure.
4. The selected unique options for the answer are 'none' after removing the duplicated options from the aggregated answer options.
5. The response in Python format for a list of items is ['none'].

The response is: ['none']

RESPONSE:

Figure 8: A prompt to classify scientific figures of maps into themes of climatic parameters when given images and captions of figures and instructions for few-shot in-context learning with explicit chain-of-thought prompting.