GAEA: A Geolocation Aware Conversational Assistant

Supplementary Material

We organize the Supplementary Material as follows: In Section 7, we provide additional details of our dataset, GAEA-1.4M . In Section 8, we summarize baselines we compared against, describe prompts used during training, inference, and evaluation, and provide training details and additional results on GAEA.

7. Addendum to the Dataset

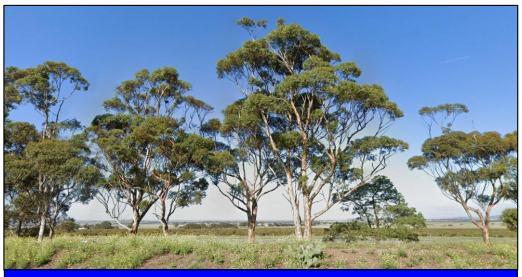
In this section, we present the dataset statistics and challenges encountered in its creation. Additionally, we discuss our plans to address these limitations in future works.

7.1. GeoGuessr Context Clues

To support geographically-grounded reasoning in our dataset, we incorporate country-specific context clues inspired by the popular game GeoGuessr [2]. We web-crawled and extracted 129 high-quality clues from Plonkit [4], a community-driven open-source resource widely used by over 65 million GeoGuessr players. To improve country coverage and address missing entries (e.g., France, New Zealand), we augmented the set with 58 additional GPT-4o-generated clues, carefully aligned with Plonkit's descriptive style. This expanded our coverage to 187 countries. These clues highlight distinctive, often visual, geographical features, such as Bangladesh's unique turquoise license plates, Australia's widespread gum trees, or the ice-covered terrain of Grise Fiord in Canada (see Fig. 9). These examples are leveraged to generate reasoning-based multiple-choice and open-ended questions. Furthermore, we enrich each sample with auxiliary metadata: Köppen-Geiger climate classification [13], traffic orientation from WorldStandards [7], land cover data from EarthEnv [1], and scene labels derived from the Places2 database [68].

7.2. Challenges with Open Street Maps (OSM)

OpenStreetMaps (OSM) [43] is a rich data source for geospatial applications. It contains a wide variety of geographic and infrastructure-related information. Using such a vast open-source dataset, we can collect data about stationary objects in the world, including infrastructure, topological information, various types of amenities (e.g., schools, hospitals, restaurants), transportation networks, international country boundaries, historical and cultural sides, and natural features (e.g., forests, rivers, and seas). Each feature from the OSM dataset has several associated features, such as names and physical characteristics. In GAEA-1.4M, we geocode the visual sample with its GPS coordinates and use the location information (longitude and



Country : Australia

Gum trees (eucalyptus) are by far the most common type of tree in Australia. Their bark is typically light-coloured, with puffy puffy clusters of leaves. Their look may vary significantly, depending on your location within the country.

Country : Bangladesh

Bangladeshi licence plates are either white, black or turquoise. NOTE: The turquoise and black licence plate can only be found in Bangladesh.

Country : Canada

Grise Fiord is the northernmost covered town in Canada. It can easily be recognized by these unique mountains and the ice floes on the coast.

Figure 9. **Examples of country-specific geographical clues** used in our dataset. These clues are used to support geographically grounded reasoning in our QA generation.

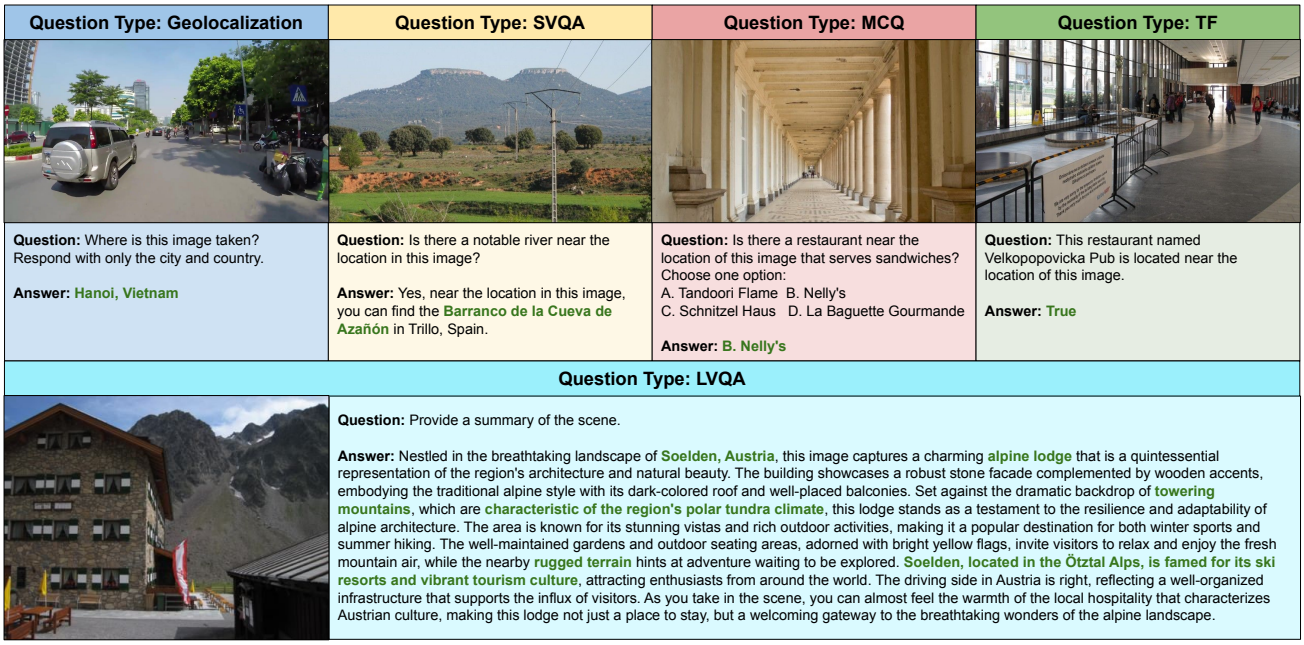


Figure 10. **Examples of the four question types in our dataset:** SVQA, MCQ, TF, and LVQA. Each type targets a distinct reasoning skill grounded in geographical, visual, or contextual understanding. Our dataset has three categories, including Geolocalization, Reasoning(LVQA), and Conversational (SVQA, MCQ, TF) QAs, as shown in the figure.

latitude) as a query to the OSM database to fetch geospatial information within a 1 KM radius and further utilize that information to generate question-answer pairs for the training of GAEA.

Despite being such a rich source of data, OSM faces several challenges. One major issue is the variability in data quality and completeness, as contributions to OSM are made by the open-source community, which may result in inconsistent information across different regions. Urban areas often have much more detailed information than rural areas, leading to less comprehensive annotations for rural regions. Another inconsistency related to human annotations stems from the different representations of the same label in different areas, introducing inherent heterogeneity in the structure of OSM data. For instance, some users might label a path as a "trail," while others might call it a "footway," and distinctions between what counts as a "park" versus a "garden" are not always clear. Moreover, querying and retrieving data from OSM is a compute-intensive task. It often becomes slower as the number of queries increases and struggles to handle dense or redundant information, necessitating efficient filtering and optimization techniques. Lastly, the information is not always up-to-date, as volunteers update different areas at different times. While some locations may have very recent data, others may be outdated, and sometimes different parts of the same area may contain information from varying periods.

Total Images	822,951
Total Cities / Countries	41,481 / 234
Total Questions	1,432,519
Total Geo-Loc QAs	822,951
Total LVQAs	236,935
Total SVQAs	267,668
Total MCQs	48,673
Total True/False QAs	56,292

Table 3. Dataset Statistics

7.3. Training Examples and Data Statistics

GAEA-1.4M covers 234 different countries and territories, and 41,481 cities. Table 3 shows the statistics of GAEA-1.4M in detail. Fig. 10 shows some qualitative examples of various question types in our GAEA-1.4M leveraging OSM, and GeoGuessr geographical clues for constructing conversational QA pairs.

8. Addendum to Baseline and Evaluation

This section covers the models used for comparison with GAEA, the prompts used during training and inference, the prompts used for evaluating GAEA-Bench, and the training hyperparameters.

Geolocalization Prompt

As a geography and tourism expert, analyze the image to determine its exact location. Utilize your extensive knowledge of geography, terrain, landscapes, flora, fauna, infrastructure, and recognizable landmarks to identify the city and country where the image was taken. Question:

LVQA Prompt

Drawing upon your expertise in geography and tourism, examine the image and provide a comprehensive description of the community or lifestyle depicted. Include insights about cultural practices, geographic features, terrain, local flora and fauna, infrastructure, and any natural or man-made elements that characterize the location. Consider how these factors influence the lifestyle and community in the area. Question:

SVQA Prompt

Provide a short answer on notable landmarks, museums, parks, restaurants, or activities that visitors might enjoy in the area. Highlight amenities and services that enhance the tourism experience at this location. Question:

MCQ Prompt

Use your comprehensive knowledge of geography, landmarks, and tourism to analyze the image and determine the correct answer from the options provided. Note, your final answer should be a choice of either A, B, C, or D, including both the letter and the complete text of the option exactly as presented. Question:

TF Prompt

Use your comprehensive knowledge of notable landmarks, museums, parks, restaurants, and related attractions to evaluate the following statement. Provide your final answer as either 'True' or 'False'. Question:

Figure 11. **Task-specific training prompts** used to instruct the model for each of the five question types in GAEA. Each prompt is carefully designed to elicit a targeted form of geographical reasoning. These prompts ensure consistent and interpretable outputs during both the training and evaluation phases while training GAEA.

Evaluation Prompt for SVQA

Evaluate the following predicted answer by comparing it to the provided ground truth. Focus on the accuracy of 1) location prediction, and 2) specificity and relevance.

- **Question:** {question}
- **Ground Truth:** {ground_truth}
- **Model Prediction:** {predicted_answer}

Scoring Guidelines:

- High score: Predicted response closely matches the specific location and provides specific information
- that closely aligns with the ground truth.
- Low score: Predicted response lacks knowledge or is unrelated to the ground truth
- Provide a score out of 10 for each criterion.
- Return only the numeric score, without additional commentary

Evaluation Prompt for LVQA

Evaluate the following predicted answer by comparing it to the provided ground truth. Focus on the accuracy of 1) location prediction, 2) cultural aspect matching, 3) consistency and quality of reasoning, 4) specificity and relevance, 5) and fluency and clarity.

- **Question:** {question}
- **Ground Truth:** {ground_truth}
- **Model Prediction:** {predicted_answer}

Instructions:

- How accurately does the predicted answer identify the specific country, city, or state mentioned in the ground truth?
- Does the predicted answer capture and reflect the cultural aspects present in the ground truth?
- Is the predicted answer logically consistent and demonstrates sound reasoning based on the information provided?
- Does the predicted answer provide specific information that is directly relevant to the question and closely aligns with the ground truth?
- Is the language in the predicted answer fluent, clear, and well-articulated?
- Provide a single overall score out of 10, based on these five criteria, weighing the criteria in the order listed, with location relevance and cultural aspect matching receiving the most weight.
- Return only the numeric score, without additional commentary

Evaluation Prompt for MCQ/TF

Evaluate the following answer based on Accuracy:

- **Question:** {question}
- **Ground Truth:** {ground_truth}
- **Model Prediction:** {predicted_answer}

Instructions:

- Match the meaning of the ground truth with the model prediction.
- If it matches, give a score of 10. Otherwise, give a score of 0.
- Strictly return only the numeric score, without any additional commentary.

Figure 12. **Evaluation prompts used for GPT-based scoring of model predictions across SVQA, LVQA, MCQ, and TF question types in the GAEA-Bench.** Each prompt specifies detailed criteria, such as location accuracy, cultural relevance, specificity, and fluency, used to assign a numeric score for qualitative assessment.

8.1. Baselines

We benchmark 8 top-performing open-source LMMs, including LLaMA 3.2-Vision [10], InternVL2 [17], Qwen2.5-VL [12], Phi3.5-vision-instruct [8], GeoChat [32], LlaVA-OneVision [35], GLM-4V-9B [24], LLaVA-NeXT-Mistral-7B [38], and 3 proprietary models, Open-AI's GPT-4o, GPT-4o-mini [9], and Google's Gemini-2.0-Flash [55] on GAEA-Bench.

Additionally, we compared the performance of GAEA against six state-of-the-art (SoTA) specialized geolocalization models, namely PlaNet [62], CPlaNet [50], ISNs [28], TransLocator [60], GeoDecoder [18], and PIGEON [25] and open-source geolocalization LMMs GeoReasoner [37] and GaGA [20] on three standard geolocalization benchmarks including IM2GPS [26], IM2GPS3k [59], and GWS15k [18]. We also compare our city-country classification performance with other LMMs on 3 benchmarks CityGuessr68k [33], GeoDE [46] and DollarStreet [23]. Preprocessing for these benchmarks is described in Section 5.1 in the main paper.

8.2. Prompts Used During Training and Inference

When training GAEA, we employed the task-specific prompts shown in Fig. 11 to align the model's understanding with target objectives. During inference, these identical prompts were used for all models evaluated on GAEA-Bench to ensure comparability.

8.3. Prompts Used in Evaluation

Fig. 12 presents the task-specific prompts used for evaluating model-generated answers via GPT-based assessment. For SVQA and LVQA, the prompts emphasize multi-faceted criteria such as location accuracy, specificity, cultural relevance, and reasoning quality, encouraging a nuanced evaluation of open-ended responses. In contrast, MCQ and TF tasks are scored based on strict binary accuracy by matching the predicted answer to the ground truth. These structured prompts ensure consistent, interpretable, and criterion-aligned evaluation across different QA types.

8.4. Training Hyperparameters

We perform single-stage training on the baseline, Qwen2.5-VL [12] using GAEA-1.4M. The training is conducted for 1 epoch with a global batch size of 128, accumulating gradients every 4 steps. The initial learning rate is set to 10^{-5}, using a cosine learning rate scheduler to provide a smooth decay in the learning rate. The warmup ratio is configured at 0.03. We performed LoRA[29]-finetuning with a rank, $r = 16$, $\alpha = 32$, and a dropout rate of 0.01. The model operates in `bfloat16` precision. We also use flash attention [19]. We list the training hyperparameters in the Table 4.

Number of epochs	1
Global batch size	128
Gradient accumulation steps	4
Initial learning rate	10^{-5}
Learning rate scheduler	cosine
Warmup ratio	0.03
LoRA rank (r)	16
LoRA scaling factor (α)	32
LoRA dropout	0.01
Precision	`bfloat16`

Table 4. Hyperparameters used for training GAEA.

8.5. Additional Results

In this Section, we discuss additional qualitative results of GAEA and compare them with selected open-source and proprietary models (as mentioned in Fig. 1 in the main paper). Fig. 13 presents a comparison of city-country predictions against other competing models. We also show the qualitative results of GAEA on short questions (SVQA), multiple-choice questions (MCQs), and true or false questions (TF) in Figs. 14, 15, 16. For these figures, we highlight correct predictions with green, while incorrect predictions are marked as red. Quantitative results on GAEA-Bench are summarized in Fig. 17.

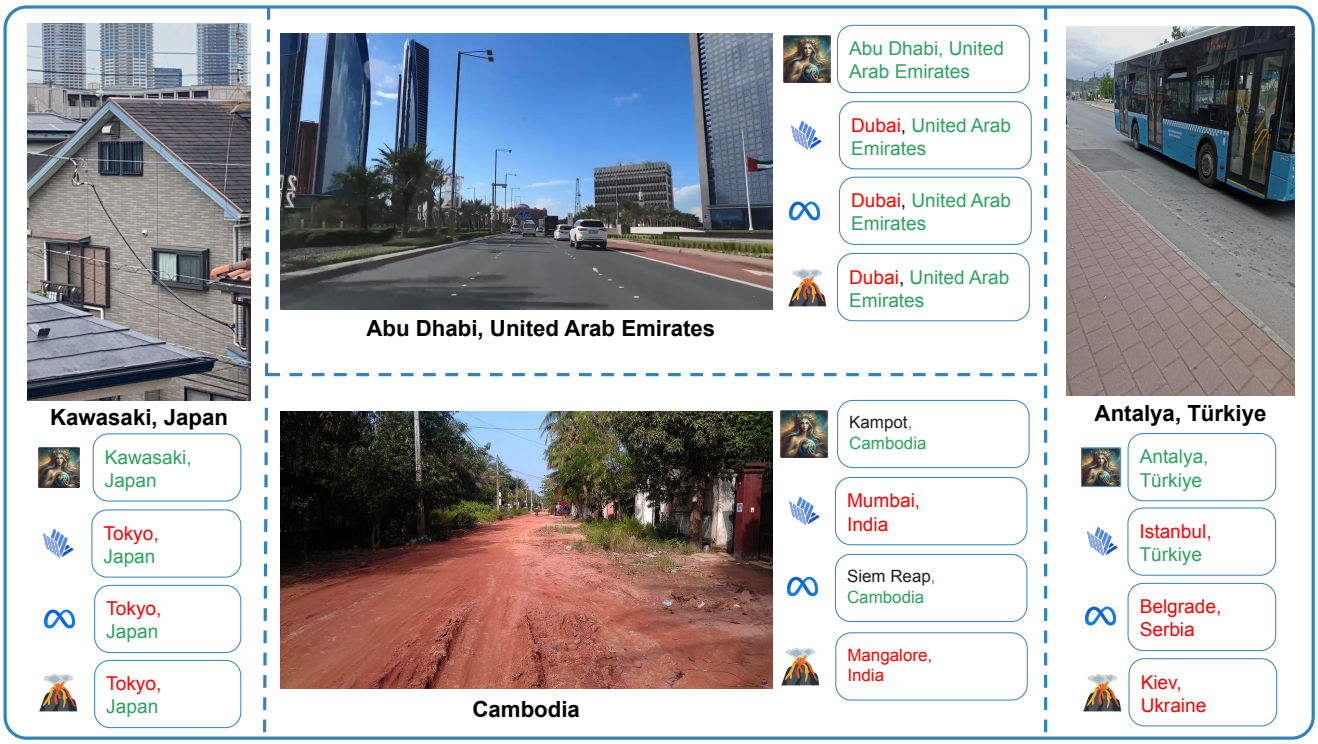


Figure 13. **Qualitative geolocalization comparison** showing GAEA's ability to predict accurate locations across varied types of scenes and landscapes around the world.

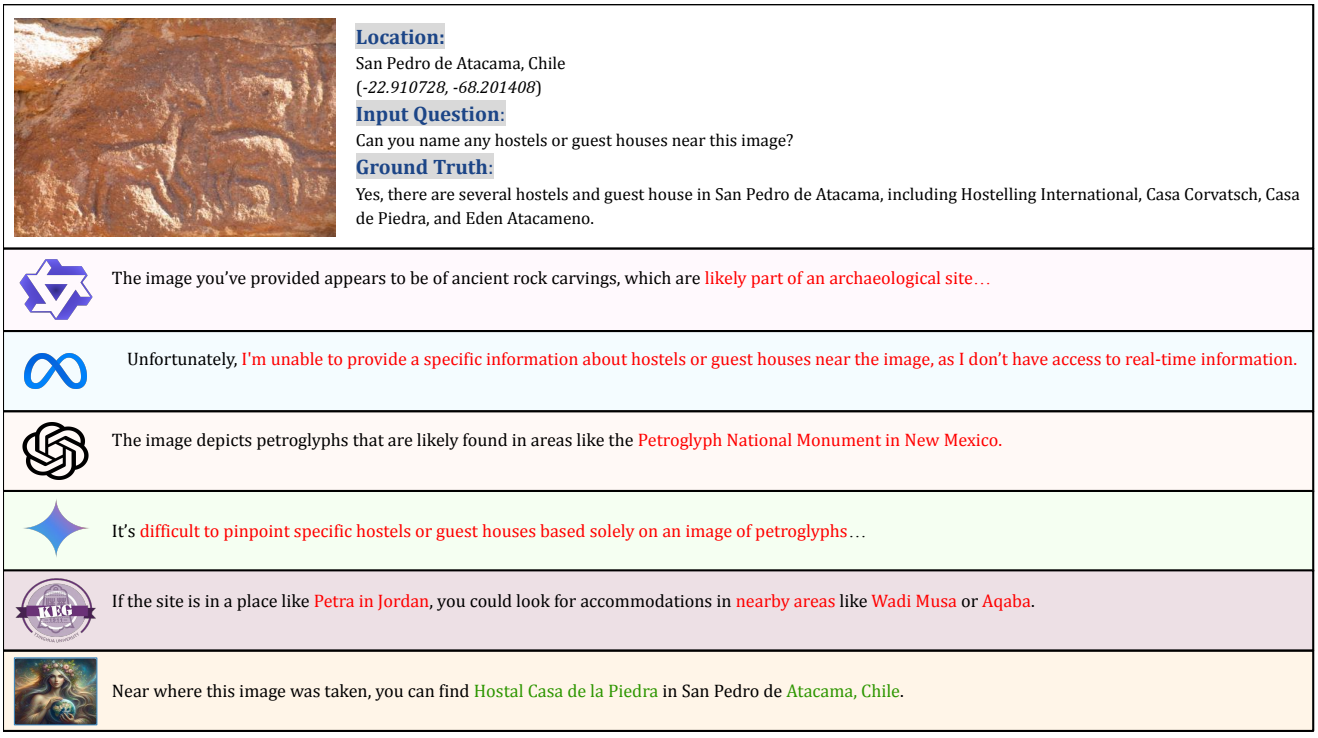


Figure 14. **Qualitative SVQA comparison** showing GAEA's ability to provide accurate, location-specific answers where other LMMs fail.

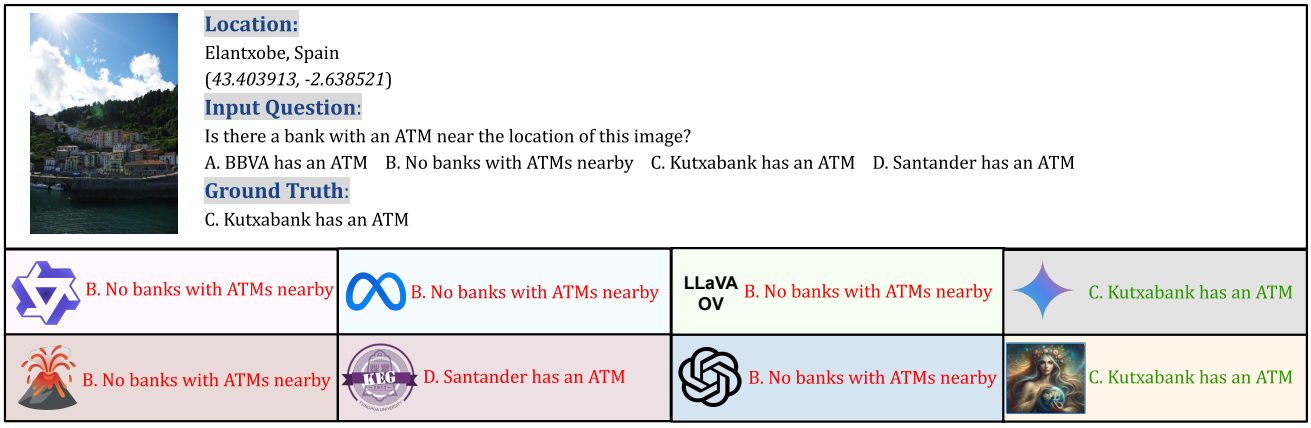


Figure 15. **Qualitative MCQs comparison** showing GAEA's ability to provide accurate answers where other LMMs fail.

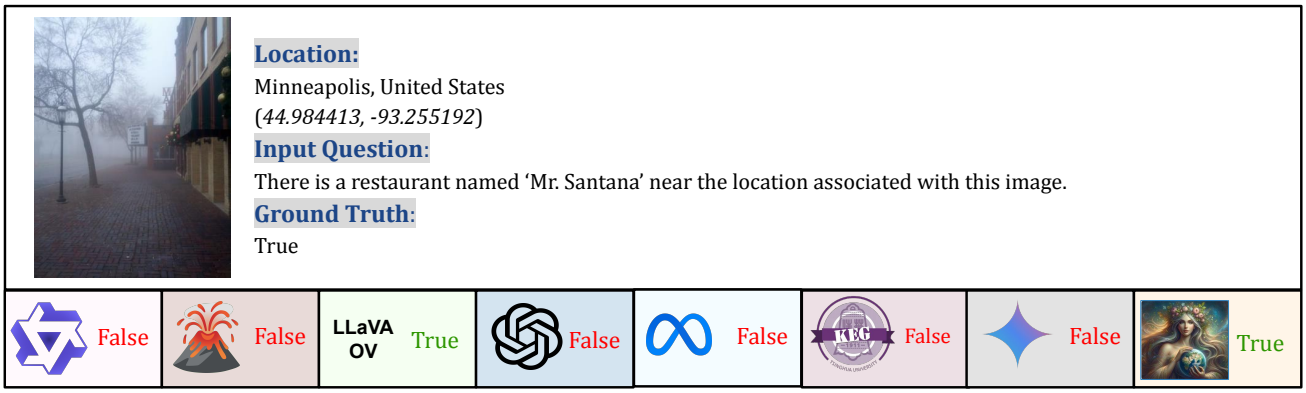


Figure 16. **Qualitative True/False comparison** showing GAEA's ability to provide accurate answers where other LMMs fail.

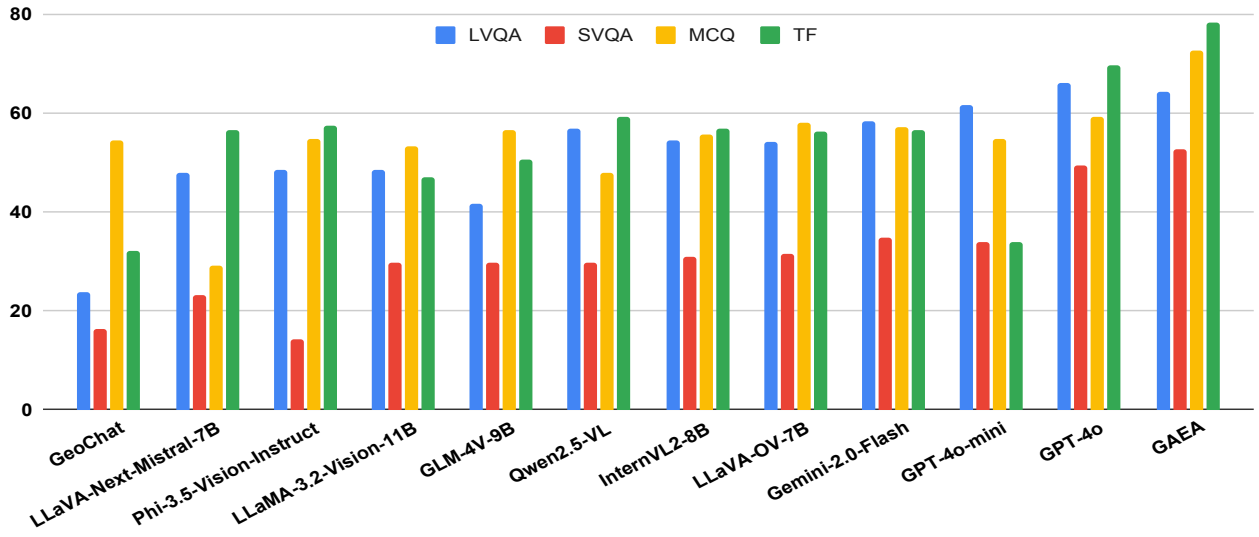


Figure 17. **Performance of various LMMs on four diverse question types.** GAEA outperforms on average across all question forms.