



SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY —
INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

**Tracing Locations Through Words and
Images: Investigating Large Language
Model Reasoning in Geo-localization**

Author:	Yujiaxuan Wang
Examiner:	Prof. Dr. Chunyang Chen
Supervisor:	Dr. Mark Huasong Meng
Submission Date:	01.10.2025

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used. This thesis was created under the supervision of Dr. Mark Huasong Meng. Standard writing environments and supporting digital tools were used during the preparation of this thesis, with all content critically reviewed and finalized by myself.

Munich, 01.10.2025

Yujiaxuan Wang

Abstract

Large Language Models (LLMs) have exhibited strong reasoning capability. However, their reasoning process remains largely opaque, especially in multi-modal tasks supporting heterogeneous inputs and outputs. In this study, we investigate the multi-modal reasoning process through the lens of image-based geo-localization around three research questions: the capability of state-of-the-art (SoTA) LLMs in geo-localization, the reasoning interpretability through visualization, and the characterization of reasoning features.

Firstly, we construct a filtered dataset of 86 smartphone images and benchmark three SoTA LLMs using correctness, measured by Granularity Score (GS), and efficiency in terms of tokens and response time. Results show that while all three models achieve city-level correctness, they show differences in efficiency-performance trade-offs. Secondly, we propose GeoMindMap, a visualization framework that decomposes reasoning trajectories into steps and semantic entities, enabling fine-grained interpretability of reasoning dynamics. Finally, we identify and quantify five reasoning features: Self-Reflection, Reasoning Gap, Breadth-First, Depth-First, and Strategy Switch. We analyze their distributions and correlations with GS. Our findings show that these reasoning features are consistently observable and partly associated with performance, exhibiting different behaviors across different models.

Overall, this work contributes methodological tools and empirical insights, inspiring future studies on multi-modal LLMs’ reasoning interpretability and reasoning mechanisms.

Webpage Demo:

<https://wyjxx.github.io/geomindmap/>

Source Code Repository:

<https://github.com/wyjxx/wyjxx.github.io/tree/main/geomindmap>

Note: Repository is intended for reference and may not be fully reproducible.

Contents

Abstract	iii
	iv
1. Introduction	1
2. Literature Review	4
2.1. LLMs Reasoning	4
2.2. Reasoning Visualization	4
2.3. Reasoning in Geo-Localization task	5
3. Methodology	6
3.1. Study Design	6
3.2. RQ1: Performance of SoTA LRMs in Geo-localization Tasks	7
3.2.1. Dataset	7
3.2.2. Reasoning Generation	7
3.2.3. Evaluation Metrics	8
3.3. RQ2: Reasoning Process Visualization	8
3.3.1. GeoMindMap	8
3.3.2. Reasoning Segmentation	10
3.3.3. Entity Extraction	10
3.3.4. Granularity and Parent Assignment	11
3.3.5. Step-Wise Semantic Matching	11
3.3.6. Visualization Framework	12
3.4. RQ3: Reasoning Features	12
3.4.1. Reasoning Features Definition	12
3.4.2. Reasoning Features Identification	14
3.4.3. Statistical Analysis	15
4. Experiment	16
4.1. Experimental Setup	16
4.2. RQ1: Performance of SoTA LLMs in Geo-localization Tasks	17

4.3. RQ2: Reasoning Process Visualization	17
4.3.1. Clue Map	17
4.3.2. Location Map	19
4.3.3. Step-wise Updates and User Interactions	19
4.3.4. Reasoning Characteristics Annotations	19
4.4. RQ3: Reasoning Features	19
4.4.1. Feature Distribution	20
4.4.2. Correlation with the Granularity Score	20
4.4.3. Effect of Strategy Switch	22
5. Discussion	24
5.1. Implications	24
5.1.1. For Users	24
5.1.2. For LLM Developers	24
5.1.3. For Humans	24
5.2. Limitations	25
6. Conclusion	26
List of Figures	27
List of Tables	28
Bibliography	29
A. Appendix	32
A.1. Reasoning Example	32
A.1.1. OpenAI o4-mini Reasoning Trajectory	32
A.2. Prompts	34
A.2.1. Geo-localization Reasoning Prompt	34
A.2.2. Reasoning Segmentation Prompt	35
A.2.3. Entity Extraction Prompt	36
A.2.4. Clue Entities Granularity and Parent Assignment Prompt	37
A.2.5. Location Entities Granularity and Parent Assignment Prompt	39
A.2.6. Semantic Matching Prompt	40
A.2.7. Reasoning Features Identification Prompt	41
A.2.8. Granularity Score Prompt	44

1. Introduction

Large Language Models (LLMs) have shown significant progress in recent years, becoming crucial in artificial intelligence research and applications. More recently, the emergence of multi-modal LLMs has extended this capability further, enabling heterogeneous inputs and outputs, such as images, videos, and audio. Advances in LLMs have also enabled Chain-of-Thought (CoT) style [23] reasoning along with the generation process. By utilizing these advancements, LLMs can solve increasingly complex tasks, such as medical image diagnosis, robotics, and geographic reasoning.

While these developments demonstrate the potential of LLMs, their reasoning process remains largely opaque. In particular, closed-source models such as the GPT-series models are often regarded as a "black box". Beyond this, multi-modal reasoning is even more mysterious, where the integration of heterogeneous inputs increases complexity. The opaque nature of reasoning limits the deployment of LLMs in safety-critical applications. Consequently, interpretability of LLMs has become a crucial research direction, as it can help reveal reasoning mechanism and enhance our understanding of how LLMs operate.

In response, several lines of research have been developed. However, they are limited in the following aspects: (a) Scope of existing analyses. Most reasoning analyses still concentrate on textual reasoning in domains such as mathematics or logical inference; little work has investigated the mechanism underlying multi-modal reasoning [16]. (b) Task-specific studies. In multi-modal tasks such as geo-localization, prior work has mainly focused on evaluating accuracy or overall reasoning quality, with limited attention to analyzing how the reasoning process develops [20, 4]. (c) Methodological limitations. Existing reasoning visualization frameworks have focused on demonstrating semantic-level trajectories, but lack the capability to perform fine-grained and entity-level analysis of reasoning structures [12, 18, 26]. Together, these gaps highlight the need for an approach to investigating the reasoning mechanism of multi-modal LLMs through dynamic visualization.

In this work, we investigate the multi-modal reasoning process through the lens of image-based geo-localization, a task aiming to infer the location of a given image. Due to its complexity, it requires multi-step reasoning. The reasoning naturally follows a structured trajectory with stages such as observation, assumption, and verification [3], containing entities such as clues and hypotheses. This enables a structured analysis for

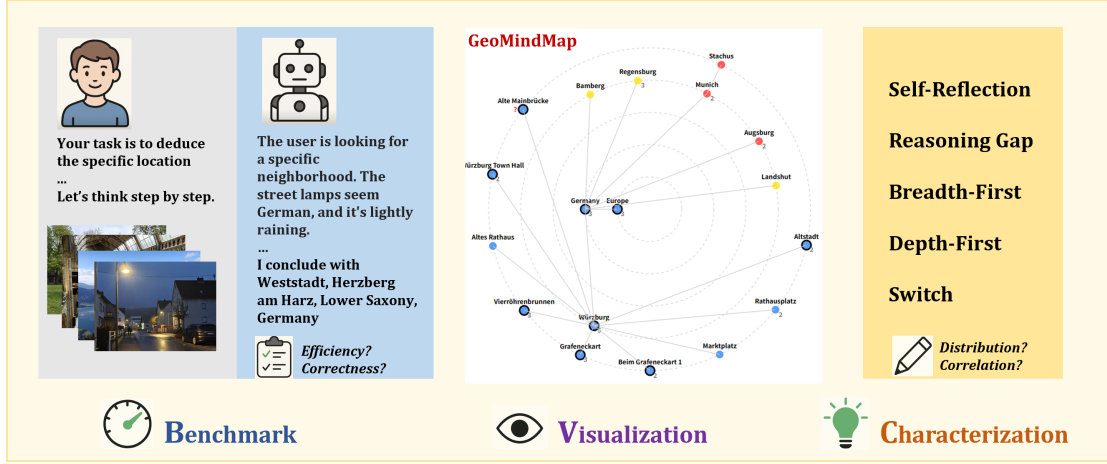


Figure 1.1.: Study design around three research questions

visualization. Moreover, reasoning features of LLMs in geo-localization resemble human problem-solving strategies in a multi-modal reasoning task. This makes geo-localization a suitable lens for investigating both the interpretability and the mechanisms of LLMs' reasoning.

Specifically, we design our study around three research questions (shown in Figure 1.1):

- RQ1: How capable are SoTA LLMs in solving geo-localization tasks?
- RQ2: Based on the output of LLMs, can we visualize the reasoning process in an interpretable way?
- RQ3: Can we recognize reasoning features in the outputs of LLMs?

To answer these questions, we first benchmark three SoTA LLMs on a filtered dataset with 86 smartphone photos, evaluating their correctness and efficiency. Secondly, we propose a visualization framework called GeoMindMap, decomposing the reasoning into steps and entities, and organizing these into two dynamic evolving maps. Finally, we define five reasoning features and conduct a statistical analysis.

Our experiments show that SoTA LLMs generally achieve city-level correctness in geo-localization tasks, though with different efficiency. The proposed GeoMindMap enables step-wise and fine-grained visualization of reasoning trajectories, revealing reasoning characteristics such as Self-Reflection and Reasoning Gap. Finally, we find that reasoning patterns (Breadth-First, Depth-First, and Switch) are measurable and partially correlated with performance.

These findings highlight the potential of reasoning visualization and feature analysis as tools for assessing LLMs’ reasoning mechanisms. Furthermore, our study emphasizes the importance of interpreting intermediate reasoning steps, offering additional insight into model behaviors. Our work fills the gap of systematically investigating the reasoning process of multi-modal LLMs at a fine-grained and dynamic level. By doing so, we contribute both methodological tools and empirical evidence that can inspire subsequent studies on enhancing LLMs’ reasoning interpretability.

We summarize our contributions as follows:

- **Benchmarking:** We construct a filtered images dataset and evaluate the capability of SoTA LRMs in geo-localization tasks.
- **Visualization:** We propose GeoMindMap, a visualization framework that decomposes reasoning trajectories into steps and entities, demonstrating how reasoning evolves dynamically at a fine-grained level.
- **Characterization:** We identify five reasoning features, quantify their distributions, and analyze their correlation with reasoning performance.

2. Literature Review

2.1. LLMs Reasoning

Recent progress in Large Language Models (LLMs) such as Deepseek-R1 [7] demonstrates their ability to perform multi-step reasoning by using Chain of Thought (CoT) [23] and tackle challenging problems [15]. The emergence of multi-modal LLM enables reasoning based on heterogeneous inputs, which is crucial for multi-modal tasks such as image-based geo-localization. At the same time, recent studies have revealed various phenomena of the reasoning process, including self-reflection [19, 27] and overthinking problem [6, 11, 2]. Self-reflection refers to the ability of the model to evaluate its own solutions and rethink the problem, whereas overthinking is observed when the model expends excessive effort on internal reasoning for a simple task. This highlighted both the potential and the challenges of LLMs. As the internal mechanisms behind LLM reasoning remain largely opaque, interpretability is essential to deepen our understanding of their reasoning process.

2.2. Reasoning Visualization

Existing studies to visualize the reasoning process can be divided into two main approaches: model behavior analysis and reasoning process visualization. The former visualizes low-level model behavior like attention mechanism [11, 21], but often fails to capture high-level behavior such as logical connection. Reasoning process visualization, such as ReasonGraph [12], Interactive Reasoning [18], and graph-based analysis [26] transform the reasoning text into a semantic graph of paragraph nodes, while Landscape of Thoughts [29] represents reasoning states as feature vectors into a 2D map. Although these frameworks provide valuable insights, they are primarily designed for general textual tasks, e.g., math or logical tasks, and present an overview of the reasoning process. They rarely perform fine-grained analysis by decomposing the paragraphs into semantic entities, such as clues and hypotheses, and present the dynamic development of these entities. Moreover, these frameworks have not yet been extended to multi-modal reasoning tasks, including geo-localization.

2.3. Reasoning in Geo-Localization task

Geo-localization aims to determine the real-world location of a given image. Existing studies on image-based geo-localization include (a) classifying an image into predefined divided geographic cells [17, 24, 5], (b) retrieving the most similar image from a geo-tagged database [13, 25, 28], and more recently, (c) generating location prediction via LLM using multimodal reasoning [8, 14, 20, 4, 9, 22, 10]. Recent LLM-based frameworks [20, 4, 9, 22, 10] emphasize the necessity of integrating reasoning process in geo-localization. Instead of focusing solely on prediction accuracy, these studies evaluated the quality of reasoning through (a) similarity to reference reasoning trajectories or knowledge bases [20, 4, 22], and (b) quality scores based on human or LLM judgments [20, 4]. However, these studies tend to analyze the reasoning as a monolithic block of text, which often serves as a post-hoc explanation for a prediction rather than a human-like dynamic process. This limitation highlights a gap, motivating our work to investigate multi-modal reasoning as a structured and evolving process through the lens of the geo-localization task.

3. Methodology

3.1. Study Design

To systematically investigate the reasoning process of state-of-the-art (SoTA) Large Language Models (LLMs) in the geo-localization task, we design our study around three research questions (cf. Figure 1.1). These questions address the problem from three perspectives: reasoning capability, reasoning interpretability, and reasoning characterization.

- **RQ1:** How capable are SoTA LLMs in performing geo-localization tasks? This question evaluates the fundamental reliability of LLMs in geo-localization tasks. We assess their performance in terms of accuracy, time efficiency, and token consumption. The result provides a benchmark of the capability and efficiency of SoTA LLMs.
- **RQ2:** Based on the output of LLMs, can we visualize the reasoning process in an interpretable way? Since interpretability is crucial for understanding LLM reasoning, RQ2 investigates whether the reasoning traces can be transformed into structured and interpretable visualizations. To address this, we propose GeoMindMap, a visualization framework that organizes the reasoning trajectory in two interactive maps and demonstrates how it evolves dynamically.
- **RQ3:** Can we recognize reasoning features in the outputs of LLMs? Beyond performance and visualization, this question aims to discover reasoning features that reveal the deeper mechanisms behind LLM reasoning. We analyze their distributions and how these features are associated with model performance. We further discuss whether the reasoning features can inspire humans to resolve reasoning tasks.

Together, these three research questions construct a comprehensive approach: RQ1 benchmarks model capability, RQ2 develops a visualization framework for interpretability, and RQ3 characterizes reasoning features.

3.2. RQ1: Performance of SoTA LRMs in Geo-localization Tasks

3.2.1. Dataset

We construct a dataset of photos captured by our team using smartphones. This avoids reliance on street view images such as Google Street View, which are often included in training data and thus pose a risk of overfitting. The collected photos are manually filtered by considering the following criteria:

- Locatability balance: We adopt the locatability metric [9], quantifying the ease of geo-localization from visual cues. An image with a score above 0.4 is considered highly locatable, while those below are low locatable. Our dataset covers both groups to ensure fairness.
- Geographic diversity: To ensure broad regional coverage, images are selected from multiple countries, covering both urban and rural areas.
- Scene diversity: To extend beyond typical street view images, the dataset includes a proportion of images taken at night, indoors, and in natural environments.

After filtering, all images are resized to a maximum side length of 1024 pixels and compressed to approximately 2 MB. Metadata such as GPS coordinates and timestamps are removed to prevent models from inferring location from this information. The final dataset contains 86 high-quality images used to demonstrate and evaluate our approach.

3.2.2. Reasoning Generation

In this step, LLMs are prompted to predict the location by giving an image from the dataset.

Since our goal is to analyze reasoning as a process rather than merely the final answer, we are inspired by CoT prompting in prior work [1] [20] [22] and design a prompt to guide the model to reason step by step: (a) First, the model identifies all clues, including man-made objects and natural elements. (b) Based on these clues, the model iteratively updates candidate locations by progressively narrowing down possibilities. (c) Finally, it produces a location prediction with street-level granularity.

The model’s response consists of two parts: reasoning and conclusion. Before generating the conclusion, LLMs first perform intermediate reasoning. Some models could use a web search to retrieve external information. Closed-source models such as GPT-series models only provide a summary of reasoning. In our analysis, we treat this as a valid reasoning text. An example trajectory is provided in Appendix A.1.

3.2.3. Evaluation Metrics

We evaluate the capability of LLMs in geo-localization by considering two aspects: correctness and efficiency.

To assess the correctness of the location conclusion, we introduce the Granularity Score (GS). At each reasoning step, the model continuously proposes and updates a location hypothesis; the last-mentioned location hypothesis is regarded as an intermediate conclusion. We prompt LLM to give a score by comparing the intermediate conclusion of each step to the ground truth. GS is defined in four levels as follows:

- 0: No clear conclusion, or completely incorrect.
- 1: Correct at the country level.
- 2: Correct at the city level.
- 3: Correct at the street/neighborhood level.

The GS of the final step is considered as the GS of the entire reasoning trajectory, representing the correctness of the conclusion. It should be noted that the GS of intermediate hypotheses does not strictly represent the correctness of the current step. A lower score may indicate either a coarser but valid hypothesis or an incorrect hypothesis, depending on the granularity level at which the model’s reasoning reaches.

Efficiency is assessed by measuring three aspects: the response generation time, reasoning token usage, and response token usage.

Together, these two metrics enable the evaluation of both accuracy and efficiency of LLMs in performing geo-localization tasks.

3.3. RQ2: Reasoning Process Visualization

3.3.1. GeoMindMap

After benchmarking the model capability (RQ1), we now turn to whether the reasoning process can be visualized (RQ2). We propose GeoMindMap (shown in Figure 3.3), an interpretable visualization framework for investigating the reasoning process in LLM-based geo-localization tasks. Rather than treating the reasoning as a monolithic text, GeoMindMap decomposes it into semantic units and organizes them into two complementary categories: the reasoning support side and the reasoning decision side. The support side includes visual elements directly observed from the image and inferential elements introduced by the model during reasoning. The support

1: Segmentation

Step 1: { "I'm looking at a captivating photo of a European old town square with a fountain... The fountain looks like it could be a Marian column" }

Step 2: { "I'm considering whether this could be the \"Plague Column\" at Stachus in Munich, but..." }

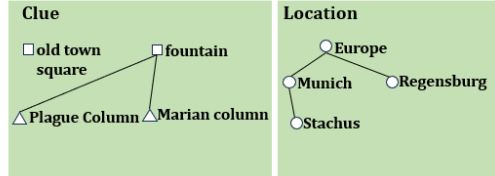
Step 3: { "... This reminds me of Regensburg..." }

GeoMindMap

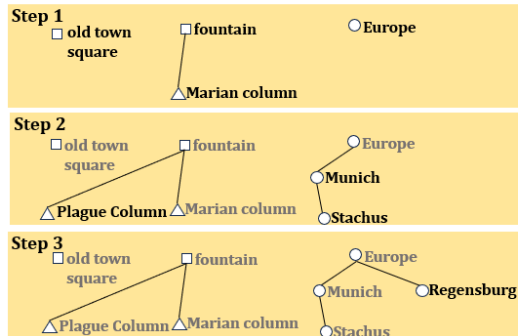
2: Entity Extraction

Clue	Location
□ old town square	○ Europe
□ fountain	○ Stachus
△ Plague Column	○ Munich
△ Marian column	○ Regensburg

3: Parent & Granularity Assignment



4: Step-wise Matching



5: Visualization

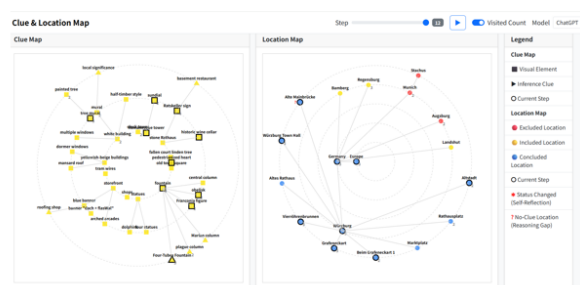


Figure 3.1.: GeoMindMap Pipeline

side contains all candidate locations across five geographic levels, which dynamically evolve during the reasoning process between three statuses: excluded, included, and concluded.

The overall pipeline of GeoMindMap is illustrated in Figure 3.1 and consists of five modules: reasoning segmentation, entity extraction, granularity and parent node assignment, step-wise semantic matching, and visualization. Given an input image, the framework generates a reasoning trace and processes this sequentially by the subsequent modules. The final output consists of two interactive maps covering all reasoning steps: a Clue Map, which groups visual elements and inferential elements from the reasoning support side, and a Location Map, which organizes candidate locations from the reasoning decision side. Together, these maps provide a dynamic and interpretable view of how LLMs combine evidence and inference to reach geo-localization decisions.

The pipeline is partly implemented through LLM prompting, which is carefully designed to constrain the output into structured JSON formats. This enables systematic analysis of the unstructured reasoning text. In the main text, we only describe and explain the role of prompts at each step, while providing the full prompts in the Appendix A.2 to ensure reproducibility.

3.3.2. Reasoning Segmentation

The framework begins by dividing LLM’s reasoning into paragraphs to facilitate step-wise explainability. This is finished by prompting LLM carefully, ensuring that the text is divided into a suitable length as a semantically unified part. The LLMs are guided to generate output in JSON format, allowing subsequent analysis.

3.3.3. Entity Extraction

From the reasoning text of the previous step, we prompt LLM to extract entities, representing the fundamental building blocks of reasoning by merging semantically similar entities (e.g., “high building” and “multi-story building”) into a single one. This process leverages LLM’s ability to understand semantics. In the next step, each entity is categorized to one of the following categories: a) Visual elements, referring to all elements directly observed from the image, such as “costal hills”, or “rectangle windows” b) Inferential elements, including inference and knowledge, which are indirectly introduced by the model during reasoning, associated with visual element, for example “suburban district”, “Baroque style” c) Location, containing mentioned specific locations, such as “Technical university of Munich”, “Spain”.

Specific entities like “European license plate” could cover multiple categories. The

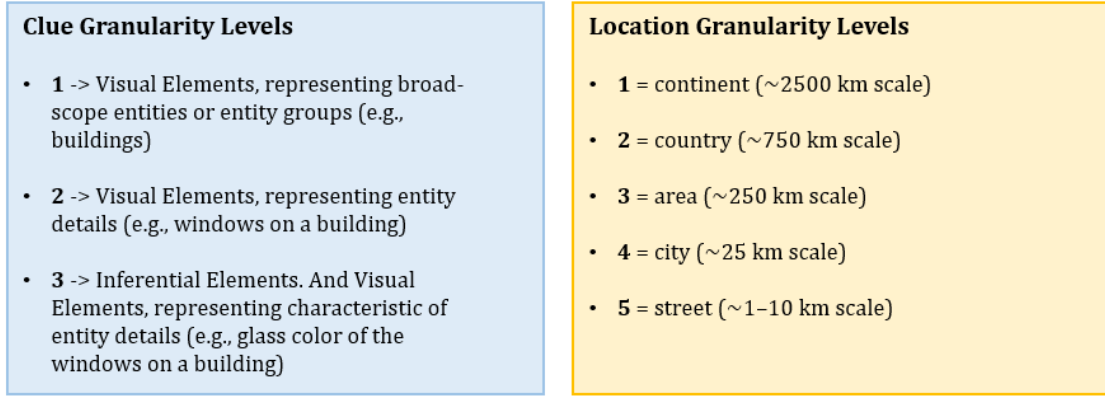


Figure 3.2.: Granularity Levels Definition

location part will be explicitly extracted, ensuring the completeness of reasoning analysis. In the end, an entity list in JSON format is generated.

3.3.4. Granularity and Parent Assignment

After entity extraction, we carefully prompt LLM to assign a granularity level (shown in Figure 3.2) and a parent node to each entity. For the reasoning support side, including visual elements and inferential elements, we define three levels of granularity. Broad-scope visual elements are assigned to granularity 1. More fine-grained visual element details are assigned to granularity 2. Finally, granularity 3 contains all inferential elements or characteristic attributes of details. For entities of granularity 2 and 3, the closest broad-scope entity is chosen as their parent. This hierarchy enables grouping entities from the reasoning support side in an ordered structure.

For the reasoning decision side, including all location entities, we construct a geographic hierarchy with five granularity levels: continent, country, area (state/province/region), city, and street/landmark. Each location is assigned to one granularity and connected to the most closed location parent node. This orders locations hierarchically and enables the connection of nodes in the Location Map.

In the end, this process generates the layouts for both the Location Map and the Clue Map.

3.3.5. Step-Wise Semantic Matching

The purpose of this process is to generate structured information for each reasoning paragraph, realized by carefully designing the prompt. Firstly, each reasoning paragraph is compared with the entity list, and entities identified through fuzzy semantic

matching are listed. Secondly, according to its mentioned context, each location is assigned to one of the three decision statuses: (a) excluded if explicitly ruled out, (b) included if considered as a candidate, (c) and concluded if narrowed down or mentioned as highly likely. Finally, for each location, all clues supporting this status are listed. The final result contains for each step the matched entity list, the location status, and its supported clues.

3.3.6. Visualization Framework

In the final stage, we construct two complementary graphs to visualize the reasoning process: the Clue Map, representing the reasoning support side, and the Location Map, representing the reasoning decision side. Each graph consists of multiple concentric circles, where each layer represents a granularity level. Nodes at higher levels of abstraction are placed closer to the center. Nodes are connected with edges to their parent nodes to build sub-trees.

We implement an interactive web interface allowing users to explore the graphs. The graphs update as the reasoning progresses step by step. Newly activated entities are highlighted or updated according to their status. The user can hover over locations to reveal the associated supporting clues. In addition, we mark several cases, including Self-Reflection, where the location statuses are changed, and Reasoning Gap, where a location is mentioned without any supporting clues.

These two maps together provide a fine-grained and dynamic view, allowing us to analyze LLMs' reasoning process interpretably.

3.4. RQ3: Reasoning Features

In RQ2, we construct GeoMindMap as a visualization framework to enable fine-grained analysis of reasoning trajectories. Based on this, for RQ3, we qualitatively analyzed the reasoning trajectories. We identified five reasoning features: two reasoning characteristics (Self-Reflection and Reasoning Gap) and three reasoning patterns (Breadth-First, Depth-First, and Switch). While reasoning characteristics capture the dynamic change in location decision, reasoning patterns reflect step-level strategies of how models collect information and propose hypotheses.

3.4.1. Reasoning Features Definition

Figure 3.4 shows the overview definition and example of each reasoning feature. We define them in detail as follows:

3. Methodology

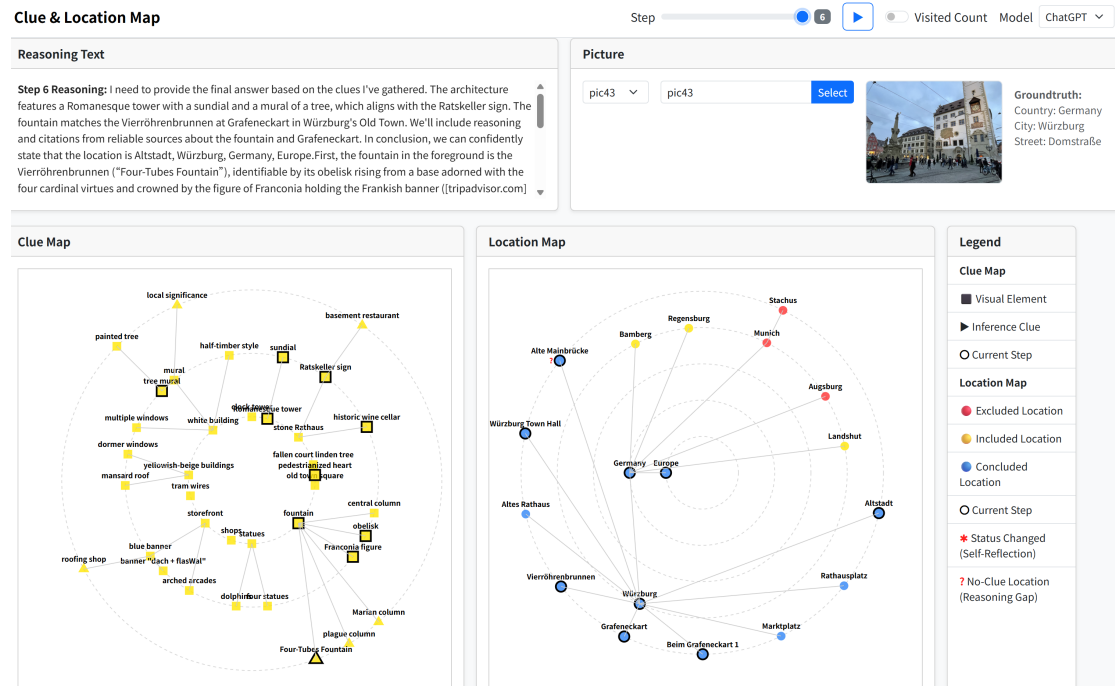


Figure 3.3.: GeoMindMap Webpage

Reasoning Gap	Self-Reflection	Breath-First	Depth-First	Switch
Introducing a location without supporting clues.	Reconsidering a previously mentioned location with a different status.	Divergent exploration, listing clues and multiple candidate locations.	Convergent verification, focusing on a specific hypothesis with detailed checks.	Transition between BF and DF when shifting exploration or verification focus
Given the user is asking for a neighborhood, could this be Augsburg or its village Inningen ? The license plate seems to show 'L', which might indicate Landshut.	Step 5: The blue recycling bins suggest a local standard, likely confirming it's in Munich . Step 7: It's possible this is a different area, maybe in Bavaria, but not in Munich .	Describing Image & Listing Clues: The image features European-style houses with red-brown tiled roofs and white plaster walls, typical of southern Germany or Austria ... Listing Multiple Candidates: I'm considering options like Eichenau or Puchheim near Munich, but Puchheim has more modern buildings. Germering seems too large, so I'm thinking smaller suburbs like Gröbenzell or Karlsfeld. Could it be in southwestern Germany, such as towns in Baden-Württemberg? ...	Analyzing Target in Details: I'm analyzing the area of Dorfstrasse in Obermenzing, particularly noting traditional buildings like the Gasthof zum Alten Wirt. The photo shows small houses ... Verifying/Eliminating Candidates: The wiki mentions that typical gable-facing buildings line Eversbuschstraße, but our photo seems to show a narrow side street instead... It's possible this is a different area, maybe in Bavaria, but not in Munich.	BF -> DF: ...My guess could point towards Untermerzing or Obermenzing, particularly along smaller streets. -> I'm analyzing the area of Dorfstrasse in Obermenzing... DF -> BF: I'm starting to think it's not Munich ... but things are hard to confirm without clearer details. -> I'm trying to identify the brand of the bins... I'm wondering if the ornamental shrub...

Figure 3.4.: Definition and example of five reasoning features

- **Self-Reflection:** Each location entity is categorized into one of the three statuses: excluded, included, or concluded. When a location reappears in a later step with a status different from the previous mention. This is regarded as Self-Reflection, indicating that the model reevaluates and changes its decision during reasoning.
- **Reasoning Gap:** A reasoning gap occurs when a location is mentioned without any supported clues. This indicates that the model introduces a location without providing any grounds, which can be considered as a form of “jumping” reasoning.
- **Breadth-First (BF):** A divergent exploration strategy when there is no specific hypothesis. The model tends to gather information broadly through observing the image and listing clues. After mentioning multiple hypothetical locations, the model briefly recalls their features and performs shallow comparisons for verification.
- **Depth-First (DF):** A convergent verification strategy, appearing often when a specific hypothesis is derived from clues. The model attempts to verify the hypothesis by observing the image purposefully, searching for particular detail clues. The model carefully recalls location features or searches online for more information. After comparison, the location hypothesis is confirmed or eliminated.
- **Strategy Switch:** A transition between BF and DF. When the model converges on local key clues or proposes a specific hypothesis, the strategy is switched from exploration to verification of the hypothesis (BF to DF). Conversely, after confirming the current granularity hypothesis or encountering contradictions, the model turns to divergent exploration to gather more clues (DF to BF).

3.4.2. Reasoning Features Identification

The identification of reasoning characteristics is directly enabled through the structured JSON output of GeoMindMap (cf. Section 3.3.5). Self-Reflection is detected by comparing the status of the same location across different steps. Reasoning Gaps are detected by checking whether a mentioned location has no supporting clues.

For the three reasoning patterns, we employed a prompt-based annotation approach to enable large-scale identification. The prompt provides a detailed definition of patterns and few-shot examples. Each reasoning step is classified into BF or DF. Additionally, transitions between two different strategies are marked as switches. Each pattern is appended with a brief explanation. The full prompt can be found in the Appendix A.2.7.

3.4.3. Statistical Analysis

We quantitatively characterize the distribution and effects of reasoning features:

a) Reasoning Characteristics

We compute the average frequency of Self-Reflection and Reasoning Gap across all reasoning steps of the dataset, and examine their association with the final GS using a correlation test.

b) Reasoning Patterns

We measure the average frequency of BF, DF, and Switch within each reasoning trajectory, and examine how these patterns relate to both step-level GS and final GS. Additionally, we investigate how Strategy Switch influences the subsequent step GS.

4. Experiment

In this section, we conduct a series of comprehensive experiments to address the three research questions. Section 4.1 provides the experimental setup. Section 4.2 investigates the performance of SoTA LLMs in solving geo-localization tasks. Section 4.3 demonstrates how GeoMindMap visualizes reasoning trajectories. Section 4.4 evaluates the distributions and effects of reasoning features.

4.1. Experimental Setup

We evaluate our approach on a dataset of 86 filtered smartphone photos (cf. Section 3.2.1). Three multi-modal LLMs are tested: OpenAI o4-mini (medium), Gemini 2.5 Pro, and Claude Sonnet 4. The models are run with both reasoning and the web search tool enabled. For consistency, we choose OpenAI o4-mini as a fixed auxiliary annotator (“tool LLM”) to perform all structured processing, including reasoning segmentation, entity extraction, granularity and parent assignment, semantic matching, granularity scoring, and pattern identification. We provide the full prompts in the Appendix A.2. For each image, the models are prompted to generate a reasoning trajectory and a final location conclusion (cf. Section 3.2.2). The reasoning text is processed through GeoMindMap, which produces step-wise visualizations in form of a Clue Map and a Location Map (cf. Section 3.3.1). Finally, reasoning features are identified from the reasoning text (cf. Section 3.4.2).

The evaluation is structured according to the three research questions as follows:

- RQ1: We evaluate the performance through two metrics (cf. Section 3.2.3): correctness, which is measured in terms of Granularity Score (GS) at three geographic levels, and efficiency, which is assessed by measuring response tokens usage, reasoning tokens usage, and response time.
- RQ2: We demonstrate that reasoning traces can be transformed into structured visualizations, producing two dynamic updated maps with annotation of reasoning characteristics.
- RQ3: We perform statistical analysis (cf. Section 3.4.3) of reasoning features, including their average frequency, correlation with the final GS, correlations

between reasoning strategies and step-level GS, and the effect of Strategy Switch on step-level GS.

4.2. RQ1: Performance of SoTA LLMs in Geo-localization Tasks

Model	Granularity Score (0-3)	Response Tokens (n)	Reasoning Tokens (n)	Response Time (s)
OpenAI o4-mini	2.08	25,970	1,288	63.3
Gemini 2.5 Pro	2.25	4,143	2,056	27.2
Claude Sonnet 4	1.89	1,352	—	45.5

Table 4.1.: Performance of SoTA LLMs measured by correctness (Granularity Score, cf. Section 3.2.3) and efficiency (tokens usage and response time)

Table 4.1 shows the performance evaluation results of three SoTA LLMs. The reasoning tokens data of Claude Sonnet 4 is unavailable due to API limitations. The results show that all models achieve approximately city-level GS (i.e., 2.0). Among the three models, Gemini 2.5 Pro demonstrates the best performance in both GS and time efficiency, while OpenAI o4-mini achieves comparable GS with significantly higher token and time cost. Claude Sonnet 4 is lightweight in token usage, but underperforms in conclusion correctness.

4.3. RQ2: Reasoning Process Visualization

To address RQ2, we implement GeoMindMap (shown in Figure 3.3), a framework to visualize the reasoning trajectory into two interpretable maps: the Clue Map (reasoning support side) and the Location Map (reasoning decision side).

4.3.1. Clue Map

Entities of reasoning support side, including visual and inferential elements, are organized into concentric layers according to their granularity. Broad visual elements

4. Experiment



(a) Ground Truth: Domstraße, Würzburg, Germany

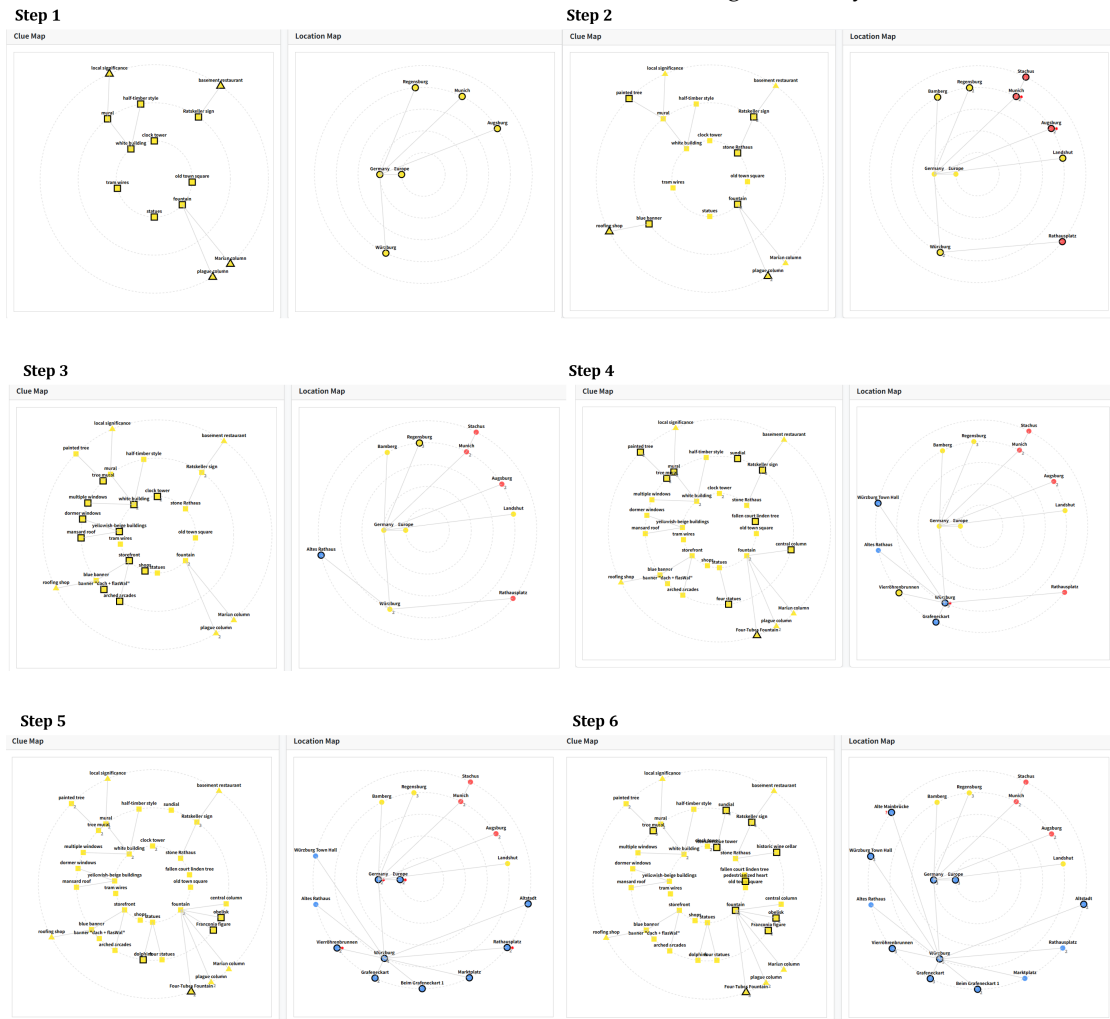


Figure 4.1.: Step-wise dynamic evolution of GeoMindMap, Model: OpenAI o4-mini, Reasoning trajectory: cf. Appendix A.1, Legend: cf. Figure 3.3

(e.g., mountain range, coastal hills) are placed near the center. In contrast, fine-grained visual details and inferential elements (e.g., rectangular windows, Baroque style) are placed in outer layers. Visual elements are shaped in a rectangle, while inferential elements are shaped in a triangle. Each entity is connected to its most closely related parent entity. This hierarchical structure reveals the connection between global observations and local details during reasoning.

4.3.2. Location Map

Entities of reasoning support side, including candidate locations, are presented hierarchically based on geographic granularity level (continent, country, region, city, street). Each node is color-coded according to its status (included, excluded, concluded) and connected to the most closed location parent node. This enables users to track how the candidates list evolves as reasoning proceeds.

4.3.3. Step-wise Updates and User Interactions

GeoMindMap supports dynamic updates synchronized with reasoning steps. Figure 4.1 presents the complete steps of an example case, with its whole reasoning trajectory provided in the Appendix A.2. At each step, newly appeared entities are highlighted. The color of location nodes is updated according to their status. The connection between decision and evidence is visualized by allowing users to hover over a location to display the supporting clues.

4.3.4. Reasoning Characteristics Annotations

The visualization reveals and annotates two dynamic behaviors: Self-Reflection, where the location’s status changes, and Reasoning Gap, where the location is mentioned without supporting clues. These phenomena serve as reasoning characteristics and motivate the analysis in RQ3.

Overall, GeoMindMap serves as a dynamic visualization framework to interpret LLMs’ reasoning process, enabling further identification of reasoning features for addressing RQ3.

4.4. RQ3: Reasoning Features

We quantitatively characterize the reasoning features, including reasoning characteristics and reasoning patterns, as defined in Section 3.4.1. The experiment consists of three parts: Feature Distribution, Correlation with Granularity Score, and Effect of Switches.

4.4.1. Feature Distribution

Model	BF proportion	DF proportion	Switch proportion	Self-Reflection (n per step)	Reasoning Gap (n per step)
OpenAI o4-mini	25.0%	75.0%	21.9%	0.726	1.792
Gemini 2.5 Pro	19.1%	80.9%	14.7%	0.232	0.531
Claude Sonnet 4	67.1%	32.9%	19.5%	0.264	0.243

Table 4.2.: Proportions of Breadth-First (BF), Depth-First (DF) and Switch across steps of all trajectories, and average frequency of Self-Reflection and Reasoning Gap per step (definitions and examples cf. Figure 3.4).

We compute the distribution of reasoning characteristics and patterns across all trajectories. The result is shown in Table 4.2. We observed that the five reasoning features occur consistently in all three models, but their distributions behave differently:

OpenAI o4-mini shows a high proportion of DF reasoning but also the largest frequency of Self-Reflection and Reasoning Gap, suggesting instability in reasoning decision-making.

Gemini 2.5 Pro shows the highest proportion of DF reasoning and the lowest proportion of Strategy Switch. It has fewer Reasoning Gaps and the fewest Self-Reflection, indicating more stable reasoning trajectories.

Claude Sonnet 4 differs from the other two models and relies mainly on BF exploration. It has a relatively low frequency of Self-Reflection with the fewest Reasoning Gaps.

4.4.2. Correlation with the Granularity Score

Table 4.3 shows the result of the average step-wise GS in case of BF and DF. Table 4.4 summarizes the relationship between reasoning features and the final GS. In the following, we introduce several key observations:

4. Experiment

	OpenAI o4-mini	Gemini 2.5 Pro	Claude Sonnet 4
BF avg step GS (0-3)	1.09	0.62	0.20
DF avg step GS (0-3)	1.70	1.78	1.46
Mann-Whitney U test	U=21672, p<0.001	U=20577, p<0.001	U=15183, p<0.001

Table 4.3.: Average step-level Granularity Score (GS) for Breadth-First (BF) and Depth-First (DF), with Mann-Whitney U test assessing the significance of differences between the two patterns.

	OpenAI o4-mini	Gemini 2.5 Pro	Claude Sonnet 4
BF prop. vs final GS	-0.364 (p=0.001)	0.273 (p=0.011)	-0.278 (p=0.010)
DF prop. vs final GS	0.364 (p=0.001)	-0.295(p=0.006)	0.292 (p=0.007)
Switch prop. vs final GS	-0.373 (p<0.001)	0.298 (p=0.005)	0.118 (p=0.282)
Self-Reflection vs final GS	-0.032 (p=0.767)	-0.090 (p=0.407)	0.187 (p=0.085)
Reasoning Gap vs final GS	0.182 (p=0.093)	0.047 (p=0.670)	0.130 (p=0.232)

Table 4.4.: Correlations between reasoning features frequencies and final Granularity Score (GS) using Spearman’s rank correlation. Each cell reports the correlation coefficient and significance level p in parentheses .

Depth-First outperforms Breadth-First at the step-level granularity score:

Across all three models, DF steps achieve significantly higher GS than BF steps. For example, OpenAI o4-mini achieves 1.70 in DF vs. 1.09 in BF, and Gemini 2.5 Pro reaches 1.78 in DF vs. 0.62 in BF. This indicates that convergent verification tends to be more effective for formulating intermediate hypotheses.

Different models exploit patterns differently:

- OpenAI o4-mini: DF proportion is positively correlated with final granularity score ($\rho=0.359$, $p=0.001$), while BF proportion and Switch proportion are both negatively correlated ($\rho=-0.364$ and $\rho=-0.373$). This suggests OpenAI o4-mini performs best when it follows stable DF reasoning with minimal switching.
- Gemini 2.5 Pro: In contrast, Gemini 2.5 Pro shows positive correlations for BF proportion and Switch proportion ($\rho=0.273$ and $\rho=0.298$), but a negative correlation for DF proportion ($\rho=-0.295$). This indicates Gemini benefits from BF reasoning and frequent Strategy Switch.
- Claude Sonnet 4: Similar to OpenAI o4-mini, Claude Sonnet 4 benefits from DF ($\rho=0.292$) and is negatively affected by BF ($\rho=-0.278$), while Strategy Switch shows no significant effect.

Reasoning characteristics are frequent but not predictive:

Self-Reflections and Reasoning Gaps are broadly observed across all models, but their correlations with the final GS are weak and statistically non-significant. This suggests they appear commonly in reasoning but do not directly affect the model’s outcome.

4.4.3. Effect of Strategy Switch

Table 4.5 summarizes the average change of step-level GS before and after Strategy Switch by using t-test to test the statistical significance. Switching to Depth-First has different impacts across models. While the small gain of OpenAI o4-mini is not statistically significant, both Gemini 2.5 Pro and Claude Sonnet 4 show large and highly significant improvement. In contrast, switching to BF exhibits either negligible or negative effects. For OpenAI o4-mini and Claude Sonnet 4, the small reductions are not significant. However, Gemini 2.5 Pro shows a significant decrease in step-level GS. The results indicate an asymmetric effect of Strategy Switch: while switching to DF is mostly beneficial, especially for Gemini 2.5 Pro and Claude Sonnet 4, moving back to BF does not improve the step-level GS.

4. Experiment

Model	Change in Step GS (BF→DF)	Change in Step GS (DF→BF)
OpenAI o4-mini	+0.188 (p=0.112)	-0.025 (p=0.850)
Gemini 2.5 Pro	+1.101 (p<0.001)	-0.741 (p=0.005)
Claude Sonnet 4	+1.21 (p<0.001)	-0.286 (p=0.208)

Table 4.5.: Paired *t*-test results for changes in step-level Granularity Score (GS) after switching reasoning strategies. Values indicate the average change in step GS with corresponding p-values.

5. Discussion

Building on these experimental findings, we now discuss the implications and limitations of our study.

5.1. Implications

Our findings contribute both methodological tools and empirical evidence, which have implications in three aspects as follows:

5.1.1. For Users

The visualization tools enable users to examine intermediate reasoning steps beyond the final answer. This enhances interpretability and transparency, which can facilitate trust in LLMs’ decision-making and support applications, such as education and decision support. By revealing reasoning patterns associated with lower outcome performance, the tools encourage a more critical use of AI rather than simple reliance on it.

5.1.2. For LLM Developers

The findings of feature analysis highlight both the potential and limitations of LLMs’ reasoning: some features are associated with better performance, while others may be negatively correlated with the outcome. Developers can use these findings to identify problems, such as excessive Breadth-First exploration. These features can be used as metrics for evaluating reasoning quality and improving model design. In addition, methodological contributions such as visualization tools can support meeting emerging requirements in AI governance.

5.1.3. For Humans

The identified reasoning features exhibit similarity to human reasoning strategies. Complex tasks are rarely solved in a linear way: some are well-founded, while others are speculative, resembling the characteristic of the Reasoning Gap. As reasoning progresses, humans often shift into a Depth-First mode by collecting more detailed

information to verify specific hypotheses. This may involve Self-Reflection: previous assumptions are reevaluated. Importantly, all these reasoning steps, including exploration, speculation, verification, and re-evaluation, are necessary for shaping the final conclusion. Although the conclusion is incorrect, the reasoning trajectory itself can provide valuable insights, such as the range of hypotheses and the interpretation of evidence. Moreover, analyzing model reasoning can further reveal reasoning strategies, which may inspire future human approaches to complex reasoning tasks.

5.2. Limitations

This study has several limitations. First, the dataset consists of only 86 manually filtered smartphone images. The limited scale restricts the diversity of scenarios and the generality of the results. A larger and more diverse dataset should be constructed. Second, reasoning trajectories are segmented into reasoning steps, which serve as basic units in reasoning visualization and analysis. This provides a coarser view and may overlook finer reasoning behaviors and dynamics. Segmentation at a finer granularity level, such as sentence-level, may enable more precise analysis. Third, the entire pipeline relies heavily on LLM-based natural language processing. Although we carefully designed prompts to mitigate noise, this may introduce potential biases and inconsistencies. Incorporating human annotators would help reduce biases and ensure higher reliability. Finally, the proposed analysis framework is restricted to the geo-localization task. In the future, this can be extended to a broader range of tasks to examine the generality of reasoning features across domains. These limitations provide directions for future work.

6. Conclusion

In this study, we investigate the reasoning process of SoTA LLMs in the geo-localization tasks around three research questions on the reasoning capability, reasoning visualization, and reasoning characterization. Firstly, we constructed a dataset with 86 filtered smartphone images and benchmarked three SoTA LLMs by evaluating both correctness and efficiency. We found that LLMs generally achieve city-level correctness in geo-localization, but they behave differently in efficiency and performance trade-offs. Secondly, we introduce GeoMindMap as a visualization framework that decomposes the reasoning trajectory into reasoning steps and semantic entities, demonstrating how reasoning evolves dynamically at a fine-grained level. This framework enables fine-grained reasoning interpretability and reveals reasoning characteristics such as Self-Reflection and Reasoning Gap. Finally, we identify five reasoning features: Self-Reflection, Reasoning Gap, Breadth-First, Depth-First, and Strategy Switch. We showed that reasoning features are consistently observable and measurable behaviors. These features exhibit distinct distributions and behaviors across different models. Reasoning patterns such as BF, DF, and Strategy Switch serve as indicators of reasoning performance. Overall, our study provides meaningful insights into how visualization framework and reasoning feature analysis contribute to a deeper understanding of AI reasoning mechanisms.

List of Figures

1.1. Study design around three research questions	2
3.1. GeoMindMap Pipeline	9
3.2. Granularity Levels Definition	11
3.3. GeoMindMap Webpage	13
3.4. Definition and example of five reasoning features	13
4.1. Step-wise dynamic evolution of GeoMindMap, Model: OpenAI o4-mini, Reasoning trajectory: cf. Appendix A.1, Legend: cf. Figure 3.3	18
A.1. Ground Truth: Domstraße, Würzburg, Germany	33

List of Tables

4.1. Performance of SoTA LLMs measured by correctness (Granularity Score, cf. Section 3.2.3) and efficiency (tokens usage and response time)	17
4.2. Proportions of Breadth-First (BF), Depth-First (DF) and Switch across steps of all trajectories, and average frequency of Self-Reflection and Reasoning Gap per step (definitions and examples cf. Figure 3.4).	20
4.3. Average step-level Granularity Score (GS) for Breadth-First (BF) and Depth-First (DF), with Mann-Whitney U test assessing the significance of differences between the two patterns.	21
4.4. Correlations between reasoning features frequencies and final Granularity Score (GS) using Spearman’s rank correlation. Each cell reports the correlation coefficient and significance level p in parentheses	21
4.5. Paired t -test results for changes in step-level Granularity Score (GS) after switching reasoning strategies. Values indicate the average change in step GS with corresponding p -values.	23

Bibliography

- [1] S. Alexander. *Testing AI's GeoGuessr Genius*. May 2025. URL: <https://www.astralcodexten.com/p/testing-ais-geoguessr-genius> (visited on 09/21/2025).
- [2] X. Chen, J. Xu, T. Liang, Z. He, J. Pang, D. Yu, L. Song, Q. Liu, M. Zhou, Z. Zhang, R. Wang, Z. Tu, H. Mi, and D. Yu. *Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs*. 2025. arXiv: 2412.21187 [cs.CL].
- [3] Z. Chen, Q. Zhou, Y. Shen, Y. Hong, H. Zhang, and C. Gan. *See, Think, Confirm: Interactive Prompting Between Vision and Language Models for Knowledge-based Visual Reasoning*. 2023. arXiv: 2301.05226 [cs.CV].
- [4] F. Cheng, J. Wang, S. Wang, Z. Huang, and X. Li. *GeoGuess: Multimodal Reasoning based on Hierarchy of Visual Information in Street View*. 2025. arXiv: 2506.16633 [cs.CL].
- [5] B. Clark, A. Kerrigan, P. P. Kulkarni, V. V. Cepeda, and M. Shah. *Where We Are and What We're Looking At: Query Based Worldwide Image Geo-localization Using Hierarchies and Scenes*. 2023. arXiv: 2303.04249 [cs.CV].
- [6] A. Cuadron, D. Li, W. Ma, X. Wang, Y. Wang, S. Zhuang, S. Liu, L. G. Schroeder, T. Xia, H. Mao, N. Thumiger, A. Desai, I. Stoica, A. Klimovic, G. Neubig, and J. E. Gonzalez. *The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks*. 2025. arXiv: 2502.08235 [cs.AI].
- [7] DeepSeek-AI, D. Guo, D. Yang, et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948 [cs.CL].
- [8] J. Huang, J.-t. Huang, Z. Liu, X. Liu, W. Wang, and J. Zhao. *AI Sees Your Location, But With A Bias Toward The Wealthy World*. 2025. arXiv: 2502.11163 [cs.CV].
- [9] L. Li, Y. Ye, B. Jiang, and W. Zeng. *GeoReasoner: Geo-localization with Reasoning in Street Views using a Large Vision-Language Model*. 2024. arXiv: 2406.18572 [cs.CV].
- [10] L. Li, Y. Zhou, Y. Liang, F. Tsung, and J. Wei. *Recognition through Reasoning: Reinforcing Image Geo-localization with Large Vision-Language Models*. 2025. arXiv: 2506.14674 [cs.CV].

- [11] X. Li, Z. Yu, Z. Zhang, X. Chen, Z. Zhang, Y. Zhuang, N. Sadagopan, and A. Beniwal. *When Thinking Fails: The Pitfalls of Reasoning for Instruction-Following in LLMs*. 2025. arXiv: 2505.11423 [cs.CL].
- [12] Z. Li, E. Shareghi, and N. Collier. *ReasonGraph: Visualisation of Reasoning Paths*. 2025. arXiv: 2503.03979 [cs.CL].
- [13] L. Liu and H. Li. *Lending Orientation to Neural Networks for Cross-view Geolocalization*. 2019. arXiv: 1903.12351 [cs.CV].
- [14] W. Luo, T. Lu, Q. Zhang, X. Liu, B. Hu, Y. Zhao, J. Zhao, S. Gao, P. McDaniel, Z. Xiang, and C. Xiao. *Doxing via the Lens: Revealing Location-related Privacy Leakage on Multi-modal Large Reasoning Models*. 2025. arXiv: 2504.19373 [cs.CR].
- [15] S. V. Marjanović, A. Patel, V. Adlakha, M. Aghajohari, P. BehnamGhader, M. Bhatia, A. Khandelwal, A. Kraft, B. Krojer, X. H. Lù, N. Meade, D. Shin, A. Kazemnejad, G. Kamath, M. Mosbach, K. Stańczak, and S. Reddy. *DeepSeek-R1 Thoughtology: Let’s think about LLM Reasoning*. 2025. arXiv: 2504.07128 [cs.CL].
- [16] P. Mondorf and B. Plank. *Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models – A Survey*. 2024. arXiv: 2404.01869 [cs.CL].
- [17] E. Müller-Budack, M. Künnel, and R. Ewerth. “Geolocation Estimation of Photos using a Hierarchical Model and Scene Classification.” In: *Computer Vision – ECCV 2018 Workshops*. Vol. 11129. Lecture Notes in Computer Science. Springer, 2019, pp. 563–579. doi: 10.1007/978-3-030-11024-6_34.
- [18] R. Y. Pang, K. J. K. Feng, S. Feng, C. Li, W. Shi, Y. Tsvetkov, J. Heer, and K. Reinecke. *Interactive Reasoning: Visualizing and Controlling Chain-of-Thought Reasoning in Large Language Models*. 2025. arXiv: 2506.23678 [cs.HC].
- [19] M. Renze and E. Guven. “The Benefits of a Concise Chain of Thought on Problem-Solving in Large Language Models.” In: *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE, Nov. 2024, pp. 476–483. doi: 10.1109/fllm63129.2024.10852493.
- [20] Z. Song, J. Yang, Y. Huang, J. Tonglet, Z. Zhang, T. Cheng, M. Fang, I. Gurevych, and X. Chen. *Geolocation with Real Human Gameplay Data: A Large-Scale Dataset and Human-Like Reasoning Framework*. 2025. arXiv: 2502.13759 [cs.CV].
- [21] J. Vig. *Visualizing Attention in Transformer-Based Language Representation Models*. 2019. arXiv: 1904.02679 [cs.HC].
- [22] C. Wang, X. Pan, Z. Pan, H. Wang, and Y. Song. *GRE Suite: Geo-localization Inference via Fine-Tuned Vision-Language Models and Enhanced Reasoning Chains*. 2025. arXiv: 2505.18700 [cs.CV].

- [23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL].
- [24] T. Weyand, I. Kostrikov, and J. Philbin. “PlaNet - Photo Geolocation with Convolutional Neural Networks.” In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 37–55. ISBN: 9783319464848. DOI: 10.1007/978-3-319-46484-8_3.
- [25] S. Workman, R. Souvenir, and N. Jacobs. *Wide-Area Image Geolocalization with Aerial Reference Imagery*. 2015. arXiv: 1510.03743 [cs.CV].
- [26] Z. Xiong, Y. Cai, Z. Li, and Y. Wang. *Mapping the Minds of LLMs: A Graph-Based Analysis of Reasoning LLM*. 2025. arXiv: 2505.13890 [cs.CL].
- [27] S. Yang, J. Wu, X. Chen, Y. Xiao, X. Yang, D. F. Wong, and D. Wang. *Understanding Aha Moments: from External Observations to Internal Mechanisms*. 2025. arXiv: 2504.02956 [cs.CL].
- [28] X. Zhang, X. Li, W. Sultani, Y. Zhou, and S. Wshah. *Cross-view Geo-localization via Learning Disentangled Geometric Layout Correspondence*. 2023. arXiv: 2212.04074 [cs.CV].
- [29] Z. Zhou, Z. Zhu, X. Li, M. Galkin, X. Feng, S. Koyejo, J. Tang, and B. Han. *Landscape of Thoughts: Visualizing the Reasoning Process of Large Language Models*. 2025. arXiv: 2503.22165 [cs.LG].

A. Appendix

A.1. Reasoning Example

A.1.1. OpenAI o4-mini Reasoning Trajectory

I'm looking at a captivating photo of a European old town square with a fountain, statues, and a large white building that has a clock tower, likely a Town Hall. The "Ratskeller" sign suggests a basement restaurant, common in German towns. The mural hints at local significance, potentially linking it to Augsburg or Regensburg. The overall architecture with half-timber style and tram wires overhead feels reminiscent of cities like Regensburg, Munich, or Wuerzburg. The fountain looks like it could be a Marian or plague column.

I'm considering whether this could be the "Plague Column" at Stachus in Munich, but that fountain's style is different. I'm also thinking about Wuerzburg since it has a town hall with a "Ratskeller." Perhaps it's Regensburg, noted for its stone Rathaus and similar features. Could it be Augsburg's town hall in Rathausplatz? No tree there, though. The painted tree reminds me of several towns, including Bamberg and Landshut. Finally, I notice the blue banner that says "Dach + Ziegel," indicating a roofing shop nearby.

I'm thinking this might be a storefront with yellowish-beige buildings featuring arched arcades and multiple windows. The building to the left has a mansard roof with dormer windows and shops underneath. The banner looks like "dach + flasWal," but it's hard to read. This reminds me of Regensburg, where buildings across from the town hall share similar features. The main building, which is white and has a clock tower, seems to resemble the "Altes Rathaus" in Regensburg. I'll search for "Regensburg Altes Rathaus tree mural" for more details.

So, it seems the tower of the Wurzburg town hall is called Grafeneckart. They have a sundial and a painted tree mural that connects to a fallen court linden tree. There's also a Ratskeller entrance nearby with the appropriate signage. I want to look deeper into this and see if they have a Vierroehren fountain, which appears to have four statues and a central column. I'll search more about the Vierroehrenbrunnen in Wuerzburg next!

I see the Four Tubes Fountain in front of the Grafeneckart. It features an obelisk with the Franconia figure holding a flag and dolphins spouting



Figure A.1.: Ground Truth: Domstraße, Würzburg, Germany

water! This confirms it's the Vierrhrenbrunnen, with the Grafeneckart building behind it and the Ratskeller entrance to the right. So, we're in the Old Town (Altstadt) of Wuerzburg, Germany, specifically at Marktplatz / Rathausplatz. The precise address is Beim Grafeneckart 1, 97070 Wuerzburg, Germany. Hence, the location is: Altstadt, Wuerzburg, Germany, Europe.

I need to provide the final answer based on the clues I've gathered. The architecture features a Romanesque tower with a sundial and a mural of a tree, which aligns with the Ratskeller sign. The fountain matches the Vierrhrenbrunnen at Grafeneckart in Wuerzburg's Old Town. We'll include reasoning and citations from reliable sources about the fountain and Grafeneckart.

In conclusion, we can confidently state that the location is Altstadt, Wuerzburg, Germany, Europe.

First, the fountain in the foreground is the Vierrohrenbrunnen (Four-Tubes Fountain), identifiable by its obelisk rising from a base adorned with the four cardinal virtues and crowned by the figure of Franconia holding the Frankish banner ([tripadvisor.com](https://www.tripadvisor.com/Attraction_Review-g187321-d7934881-Reviews-Vierrohrenbrunnen-Wuerzburg_Lower_Franconia_Franconia_Bavaria.html?utm_source=openai), [tripadvisor.co.nz](https://www.tripadvisor.co.nz/ShowUserReviews-g187321-d7934881-r824989212-Vierrohrenbrunnen-Wuerzburg_Lower_Franconia_Franconia_Bavaria.html?utm_source=openai)). This fountain stands directly in front of the Grafeneckart, the oldest part of Wuerzburgs Town Hall.

Second, the white-cream tower to the right is the Grafeneckart itself a 55 m-high Romanesque tower with both a clock and a sundial on its faade, and on its side wing a painted court-linden mural (Gruenbaum) originally executed in 1597 and restored in 1914 ([extensionembed.wanderlog.com](https://extensionembed.wanderlog.com/list/geoCategory/105007/?utm_source=openai)). The doorway beneath, marked

```
Ratskeller, leads to the historic wine cellar in the town hall complex.\n\nBoth features lie at Beim Grafeneckart 1 in the pedestrianized heart of\nWuerzburg's Altstadt (Old Town), directly across from the Alte Mainbruecke.\n\n\n**Location Conclusion** Altstadt, Wuerzburg, Germany, Europe"
```

A.2. Prompts

A.2.1. Geo-localization Reasoning Prompt

Your task is to deduce the specific location where a photo was taken, with the result refined down to a neighborhood within a city.

A single image contains many types of detailed features, which we call clues. Clues can be categorized into the following main groups:

1. Man-made objects: including but not limited to building types and styles, window designs, brick/tile materials and colors, contractor markings, railing types, utility pole styles, wiring layouts, road paving styles, road signs and markings, curb details, vehicle and license plate designs, signage text, business names, clothing, and skin tones of people.
2. Natural elements: including but not limited to landforms, elevation changes, rock colors, soil types, ground vegetation, tree or shrub species, distant mountain ridges or coastlines, shadow length and angle, shadow intensity, cloud height and shape, position of sunlight source.

After identifying one or more clues, you must analyze what possible location(s) those clues suggest. For example:

- Man-made objects may help infer the country or region.
- Natural elements may suggest climate zone, elevation, or type of natural environment.
- Sky and lighting features may help infer northern/southern hemisphere, latitude, or climate.

Some clues may support a candidate location, while others may contradict it. You must stay open-minded and avoid prematurely ruling out possibilities. When clues lean in favor, you can add candidate locations. When too many clues contradict a region, you may eliminate it. When clear and specific clues appear, you can narrow down to finer levels. This stage is crucial and where you are most likely to make mistakes. Constantly ask yourself:

"Wait! Did I exclude other regions too early? Are there nearby areas with the same clues?"

List possible options. Actively look for evidence supporting them. Compare them directly with your leading guess without bias.

How compatible is each clue with different locations? How strong is the evidence?

Through iterative hypothesis and elimination, you must continuously narrow down the estimated location until it is refined to a neighborhood within a city.

Make your final decision only when you are fully confident in your conclusion.

Your response output contains a single final location as answer, it must strictly follow the following example string format:

"Final Conclusion: Kurfuerstenstrasse, Berlin, Germany, Europe"

Cautions:

- The image may not come from Google Street View; many are personally taken photos.
- Metadata such as image names or EXIF info must not be treated as valid clues.
- Do not jump to a final conclusion when multiple regions are still possible.
- Important: Response output contains only a single conclusion!!!

Let's think step by step.

A.2.2. Reasoning Segmentation Prompt

Keep the full original text unchanged. Split the text into semantic paragraphs. If the text already has titles/headings, use them for paragraph division. If not, split according to meaning. Control the length of each paragraph so it is neither too short nor too long: merge overly short ones, and split overly long ones if necessary.

Output the result in JSON format like this:

```
[
{
  "title": "title1",
  "content": "content1"
},
{
  "title": "title2",
  "content": "content2"
}
]
```

Important:

- Strictly follow the example output format
- Keep the full original text unchanged!
- Please output only raw JSON. Do not use any Markdown syntax

Your task:

A.2.3. Entity Extraction Prompt

Based on the image content, please extract the following three types of entity keywords from the reasoning text:

1. l: location includes all specific locations (continents, countries, administrative regions, cities, towns, villages, streets, and geographic names such as mountains, rivers, lakes, landforms). Do not include vague location terms like city, village, neighbourhood, urban.

Note that a location can appear as both an adjective and a noun do not miss any location mentions.

2. v: visual element directly observable elements from the image, including man-made objects, natural features, and global/overall scene information.

3. i: inference and knowledge element includes information inferred from visual elements, background knowledge, or search-based clues.

Do not include entities unrelated to human or geographic features.

Guidelines:

- Each noun should be a separate entity. For example:
a tall building with a rectangle window should be split into a tall building and a rectangle window.
- Keep expressions concise, with entity names preferably no longer than 3 words.
- Split entities that include regional elements. For example:
"European road sign" extract both "Europe" and "European road sign"
"Munich's road" extract both "Munich" and "Munich's road"
"Technical University of Munich" extract both "Munich" and "Technical University of Munich"
- Output a complete and non-redundant list of key entities. You may merge semantically similar entities into a single one.

Output format (JSON only):

```
[  
{"entity": "KEY ENTITY", "type": "TYPE"}  
...  
]
```

Example:

Input I see an image featuring a highway with separated lanes, probably in Europe, with tall residential buildings on the left, which indicates typical German Plattenbau. There's a tall white building displaying a red banner that might say "Augsburger".

Output

```
[
{"entity": "highway", "type": "v"}
{"entity": "separated lanes", "type": "v"}
{"entity": "Europe", "type": "l"}
{"entity": "tall residential buildings", "type": "v"}
{"entity": "German Plattenbau", "type": "i"}
{"entity": "Germany", "type": "l"}
{"entity": "red banner 'Augsburger'", "type": "v"}
]
```

Important:

- Strictly follow the example output format
- Stay faithful to the original text and do not omit
- Please output only raw JSON. Do not use any Markdown syntax

Your task:

A.2.4. Clue Entities Granularity and Parent Assignment Prompt

Please categorize the clue words into three granularity levels as follows and assign a parent node:

Granularity: 1 -> Entities of type v, representing broad-scope entities or entity groups (e.g., buildings, vegetation, roads, signs, sky)

Parent node guideline: Assign parent to null.

Granularity: 2 -> Entities of type v, representing entity details (e.g., windows on a building, shrubs in vegetation, markings on a road)

Parent node guideline: Select a node from granularity 1 as the parent node to indicate its closest association, if no suitable parent is found, leave the parent as null.

Granularity: 3 -> All entities of type i. AND All left entities of type v, representing characteristic of entity details (e.g., glass color of the windows on a building)

Parent node guideline: Select a node from granularity 2 or 1 as the parent node to indicate its closest association, if no suitable parent is found, leave the parent as null.

Rules for assigning parent:

- Each entity only have one parent node
- Nodes with granularity = 1 must have "parent": null
- The parent must be a node from a higher level, i.e., granularity-1, granularity-2, etc.
- First search through all nodes in granularity-1. If no suitable parent is found, continue to granularity-2, and so on until granularity=1

Output format (JSON only):

```
[
{"entity": "ENTITY", "type": "TYPE", "granularity": LEVEL_NUMBER, "parent": PARENT
  ENTITY}
...
]
```

Example:

Input:

```
[
{"entity": "highway", "type": "v"}
{"entity": "separated lanes", "type": "v"}
{"entity": "lanes color", "type": "v"}
{"entity": "Europe", "type": "l"}
{"entity": "tall residential buildings", "type": "v"}
{"entity": "German Plattenbau", "type": "i"}
{"entity": "Germany", "type": "l"}
{"entity": "red banner 'Augsburger'", "type": "v"}
]
```

Output:

```
[
{"entity": "highway", "type": "v", "granularity": 1, "parent": null}
{"entity": "separated lanes", "type": "v", "granularity": 2, "parent": "highway"}
{"entity": "lanes color", "type": "v", "granularity": 3, "parent": "separated lanes"}
{"entity": "tall residential buildings", "type": "v", "granularity": 1, "parent": null}
{"entity": "German Plattenbau", "type": "i", "granularity": 3, "parent": "tall residential building"}
{"entity": "red banner 'Augsburger'", "type": "v", "granularity": 2, "parent": "tall residential building"}
]
```

Important:

- Strictly follow the example output format
- Do not assign a parent with the same granularity level! Forbidden!

- Stay faithful to the original text and do not omit anything
 - Please output only raw JSON. Do not use any Markdown syntax
- Your task:

A.2.5. Location Entities Granularity and Parent Assignment Prompt

Your task is to first classify all location entities into five hierarchical levels based on their geographic scope and output them in order.

Granularity Level Definitions:

- 1 = continent (e.g., continents, oceans; ~2500 km scale)
- 2 = country (e.g., countries, transnational geographic features like mountain ranges; ~750 km scale)
- 3 = area (e.g., administrative regions within a country: states, provinces, regional landforms; ~250 km scale)
- 4 = city (e.g., cities, towns, villages, small water bodies, peaks; ~25 km scale)
- 5 = street (e.g., urban districts, neighborhoods, landmarks, streets; ~110 km scale)

Note: For ambiguous geographic terms (e.g., water bodies, mountain ranges, plains), use the approximate km range above to determine the level.

Then, for each word, you must assign an appropriate parent node to indicate its closest association:

- Nodes with granularity = 1 must have "parent": null
- The parent must be a node from a higher level, i.e., granularity-1, granularity-2, etc.
- First search through all nodes in granularity-1. If no suitable parent is found, continue to granularity-2, and so on until granularity=1
- If no suitable parent is found after reaching granularity=1: Leave the parent as null
- Parents must not be from the same or lower level (forbidden: granularity, granularity+1, ...)

Exmaple:

Input:

```
[
{"entity": "Marienplatz"},
{"entity": "Alps"},
{"entity": "Europe"},
{"entity": "Germany"},
{"entity": "Tokyo"},
{"entity": "Innsbruck"},
{"entity": "Southern Germany"},
{"entity": "Bavaria"},
{"entity": "Munich"},
```

```
{"entity": "Altstadt"},
]
```

Output

```
[
{"entity": "Europe", "granularity": 1, "parent": null},
{"entity": "Alps", "granularity": 2, "parent": "Europe"},
{"entity": "Germany", "granularity": 2, "parent": "Europe"},
{"entity": "Southern Germany", "granularity": 3, "parent": "Germany"},
{"entity": "Bavaria", "granularity": 3, "parent": "Germany"},
{"entity": "Tokyo", "granularity": 4, "parent": null},
{"entity": "Innsbruck", "granularity": 4, "parent": "Alps"},
{"entity": "Munich", "granularity": 4, "parent": "Bavaria"},
{"entity": "Altstadt", "granularity": 5, "parent": "Munich"},
{"entity": "Marienplatz", "granularity": 5, "parent": "Munich"}
]
```

Important:

- Strictly follow the example output format
- Stay faithful to the original text and do not omit anything
- Think carefully before assigning parents, follow the rules exactly!
- Please output only raw JSON. Do not use any Markdown syntax

Your task:

A.2.6. Semantic Matching Prompt

You are responsible for handling a semantic entity matching task. Given a paragraph of text and a list of pre-defined entity terms labeled with their types (including visual elements, inference/knowledge terms, and locations), your task is to semantically match the content of the paragraph with the given entity terms and complete the following two subtasks:

Step 1. Clue Extraction:

List all words or phrases in the paragraph that semantically match any entity terms of type v (visual elements) or i (inference/knowledge). These matched items will serve as clues.

Step 2. Location Mention Detection:

List all words or phrases in the paragraph that semantically match any entity terms of type l (locations). For each matched location:

- Determine in what context the model mentions or considers this location and classify the location into one of the following statuses by assigning a number:
 1. Excluded: Explicitly ruled out, with contradictions pointed out, impossible
 2. Included: Explored, recalled, relevant knowledge listed and compared, possible candidate

3. Concluded: Narrowed down, confirmed, specified, high likely based on evidence
- Also, list the corresponding clue terms (from step 1 Clue Extraction, and only those) that support this judgment.

Important Constraints:

You must only use the given list of entity terms for matching. Do not add, modify, or infer terms beyond the provided list.

Matching should allow for fuzzy semantic recognition: an exact textual match is not required as long as the semantic meaning clearly corresponds.

Important: Do not miss any word!

Important: You must **copy exactly** the string from the provided "entity" field in the input list, without changing, shortening, or removing any part, including parentheses and text inside them. The match must be letter-for-letter identical.

The output structure must strictly follow the JSON format below:

```
{
  "paragraph":1,
  "clue":["clue1","clue2","clue3"],
  "loc-clue":
  [
    {
      "loc":"loc1",
      "status":1,
      "related_clue":["clue1","clue2"]
    },
    {
      "loc":"loc2",
      "status":3,
      "related_clue":["clue3"]
    }
  ]
}
```

The following is the list of entity terms to be used for matching:

A.2.7. Reasoning Features Identification Prompt

You are a research expert specializing in analyzing LLM reasoning processes. Your task is to annotate and analyze LLM geolocation reasoning trajectories based on a defined theoretical framework.

1. You will analyze a series of reasoning steps. After careful analysis, classify each paragraph according to its main content into one of the following two patterns:

a) **Breadth-First**:

This is a divergent exploration mode, which can appear at different stages of reasoning:

- * **Clue collection**: When there is no specific hypothesis, in order to gather information, the model broadly and descriptively observes the image, lists clues, and obtains an overall impression.
- * **Hypothesis generation**: When key clues are lacking, it jumpingly mentions multiple hypothetical locations, briefly recalls their features, and performs quick, shallow comparisons for verification.
- * **Reference keywords**: consider, observe, try, wonder, notice, might, could

b) **Depth-First**:

This is a convergent verification analysis mode, which can also appear at different stages of reasoning:

- * **Clue collection**: When there is a specific hypothesis, in order to verify it, the model observes the image specifically and purposefully, searching for particular detail clues.
- * **Hypothesis generation**: Based on key clues, the model derives a hypothesis, carefully recalls location features or searches online for more information, and after comparison, either confirms or eliminates that location hypothesis.
- * **Reference keywords**: focus, detail, lead to, indicate, hint, confirm, specific, narrow down, point

2. In addition, you also need to annotate all paragraphs where a **switch** occurs. Please refer to the following definition:

c) **Breadth-Depth Switch**:

This is the node where the reasoning mode changes, usually occurring in the following scenarios:

i) **Switch to breadth mode**:

- * When high-confidence clues are lacking, turning to divergent exploration
- * After verification confirms the current granularity hypothesis, shifting to explore finer-grained locations
- * After verification encounters contradictions, eliminating the current hypothesis and turning to search for another

ii) **Switch to depth mode**:

* Attention in breadth mode gradually converges on local key clues, or a specific hypothesis is proposed, thereby turning to the verification analysis of depth mode

3. Finally, briefly explain your annotation choices.

Please refer to some examples:

a) ****Breadth-First****:

- i) The user is interested in pinpointing a neighborhood from a street photo. The image features European-style houses with red-brown tiled roofs and white plaster walls, typical of southern Germany or Austria. The architecture suggests it could be Bavaria, with steep gabled roofs and pastel-color facades. Notably, there's a tree and a parked Mercedes with a likely German license plate. There are distinct characteristics like wooden picket fences and horizontal slats on windows, enhancing the local vibe.
- ii) I'm considering options like Eichenau or Puchheim near Munich, but Puchheim has more modern buildings. Germering seems too large, so I'm thinking smaller suburbs like Grbenzell or Karlsfeld, yet they appear standard. Could it be in southwestern Germany, such as towns in Baden-Wrttemberg? That would align with certain architectural styles. I see ivy and dormer windows on houses, which further narrows it down. There's a faint sign with "Gaststube," suggesting a tavern. It might be in a small village like Emmering, in Frstenfeldbruck, but I'm unsure. The blue recycling bins suggest a local standard, likely confirming it's in Munich. The license plate indicating "M" supports this too. My guess could point towards Untermenzing or Obermenzing, particularly along smaller streets.

b) ****Depth-First****:

- i) I'm analyzing the area of Dorfstrasse in Obermenzing, particularly noting traditional buildings like the Gasthof zum Alten Wirt. The photo shows small houses, yet not the larger inn. The nearby area features St. Georg church, but no church tower is visible. I believe the tall poplar tree behind may indicate farmland. It seems this isn't Obermenzing but Untermenzing, where Eversbuschstrasse has preserved gable-style houses. I'll look up images for Eversbuschstrasse in Untermenzing to confirm details.
- ii) The wiki mentions that typical gable-facing buildings line Eversbuschstrae, but our photo seems to show a narrow side street instead. In the image, gables face us, indicating the street runs perpendicular to the ridge. Eversbuschstrae does have gable-facing houses, though. I wonder if we need specific house numbers like 62 or 54

since theyre listed as former farmhouses. Yet, the modern look might complicate identification. Its possible this is a different area, maybe in Bavaria, but not in Munich.

c) ****Switches****:

* ***Switch to Breadth***: Its possible this is a different area, maybe in Bavaria, but not in Munich. Im considering if this photo could be from villages near Ingolstadt.

* ***Switch to Depth***: The blue recycling bins suggest a local standard, likely confirming its in Munich. The license plate indicating "M" supports this too. My guess could point towards Untermenzing or Obermenzing, particularly along smaller streets.

****Your output format must be JSON and strictly follow the form below:****

```
{
  "Breadth-First": [{"Step":1,"Explanation":"Explain shortly your decision here"}, {"Step":3,"Explanation":"Explain shortly your decision here"}],
  "Depth-First": [{"Step":2,"Explanation":"Explain shortly your decision here"}, {"KeyElement":"List key hypothesis or clue to be verified or analyzed"}],
  "Breadth-Depth Switch": [{"FromStep":1,"ToStep":2,"SwitchType":"ToDepth", "Explanation":"Explain shortly your decision here"}, {"FromStep":2,"ToStep":3,"SwitchType":"ToBreadth", "Explanation":"Explain shortly your decision here"}]
}
```

Important:

* Do not omit anything

* Strictly follow JSON format, do not output in markdown format

A.2.8. Granularity Score Prompt

You are an helpful assistant to evaluate the accuracy of location conclusions. Given a text, for each paragraph, compare the hypothesis or conclusion by the end of it with the ground truth location, and rate accuracy:

0: No clear hypothesis or conclusion, or completely wrong at all levels.

1: Correct at country level.

2: Correct at city level.

3: Correct at street/neighborhood level.

A. Appendix

Output strictly in the following JSON format: [{"step":1,"location":"Hypothesis or Conclusion location","accuracy":0},{"step":2, "location":"Hypothesis or Conclusion location","accuracy": 1}]

Important:

- Strictly follow the output format
- Please output only raw JSON. Do not use any Markdown syntax