Evaluating Map-Reading Capabilities of Vision-Language Models:

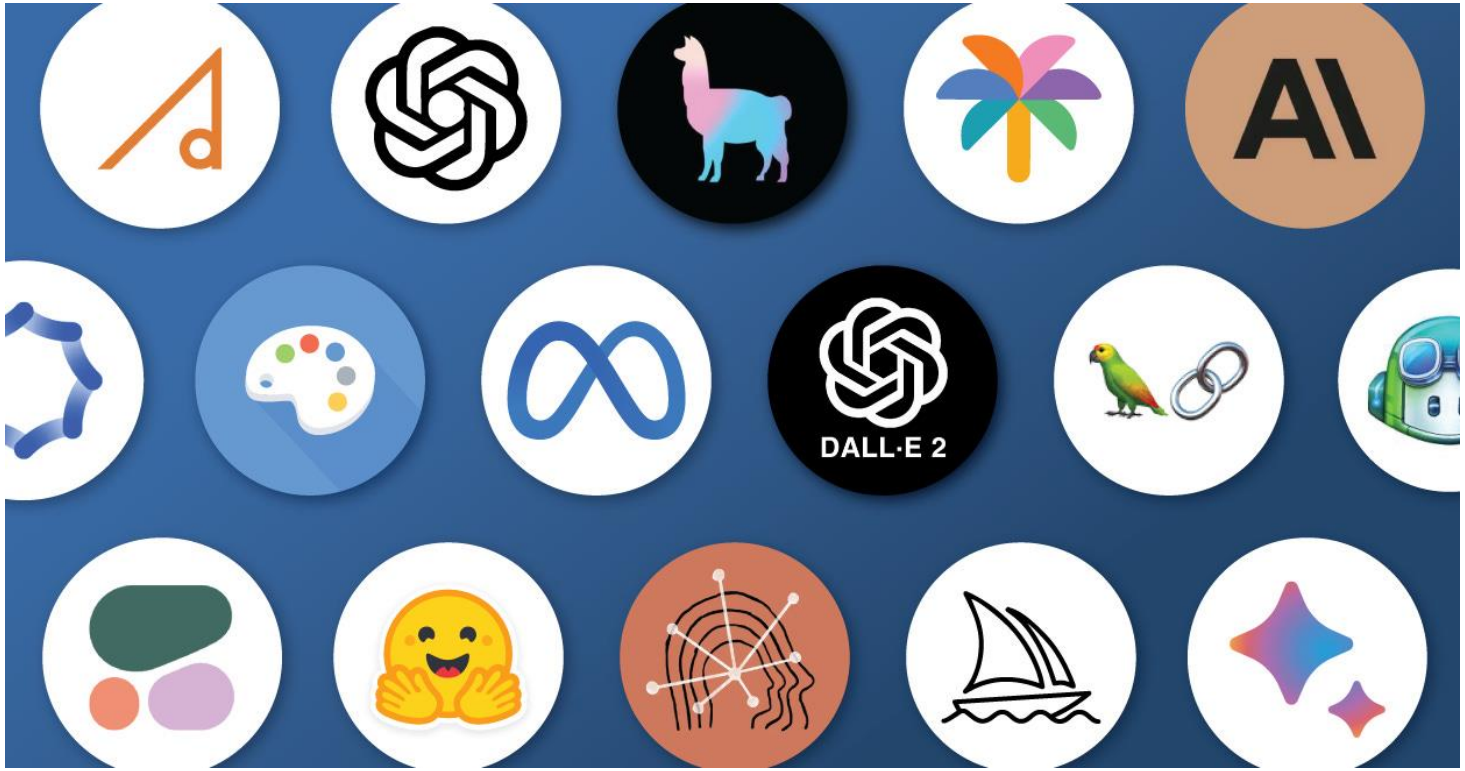
A Bloom's Taxonomy-Based Framework

Presenter: Yangyang Zhao
Affiliation: Chinese Academy of Surveying and Mapping, China
Ghent University, Belgium
Date:June 2026

1.Background and Research Gap

These models are increasingly being applied in geospatial and cartographic domains.

- map interpretation
- geographic question answering
- spatial reasoning
- …



1.Background and Research Gap

Can current VLMs truly read and understand maps in a human-like way?

Map reading is much more complicated.

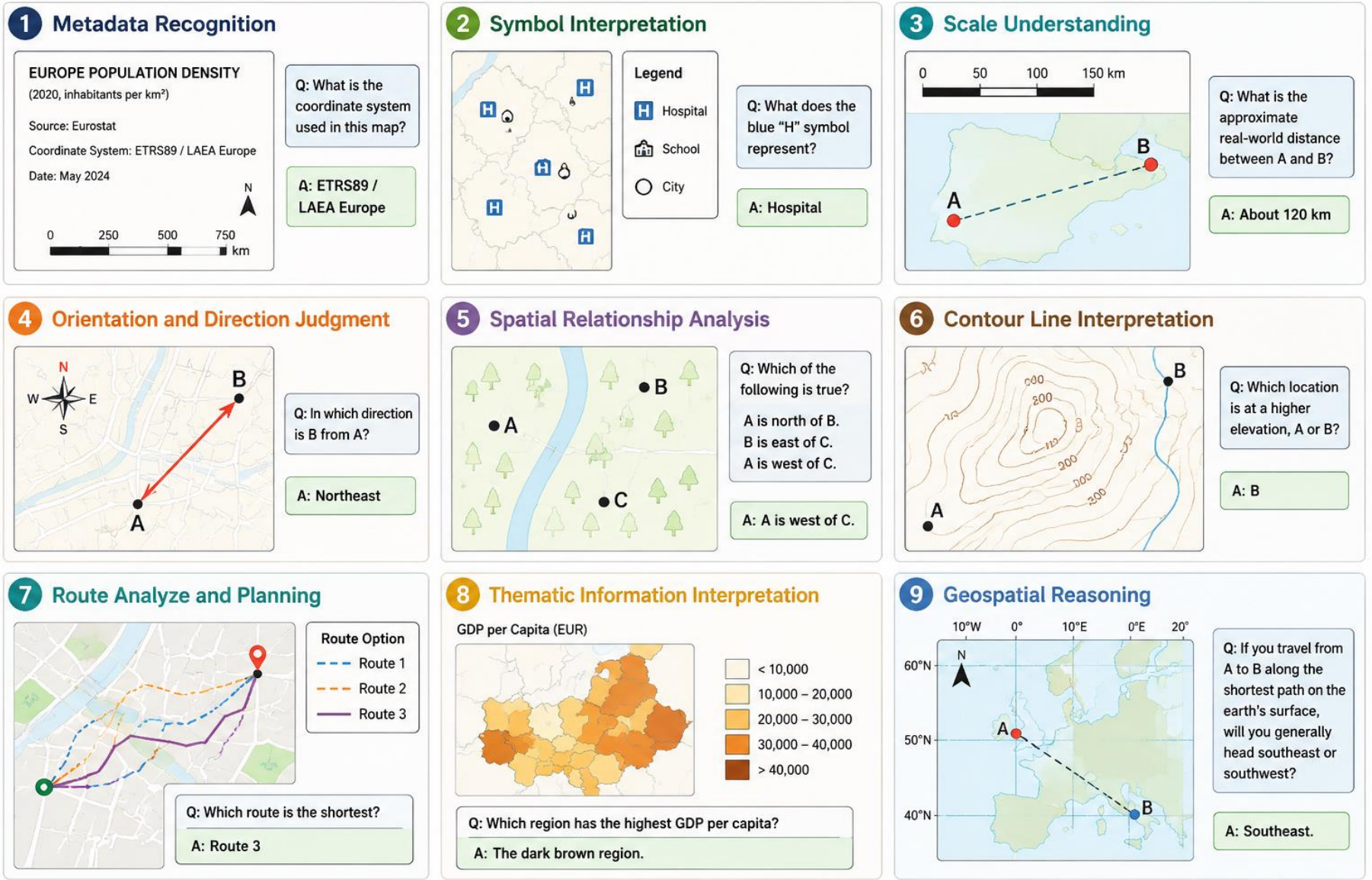
There is still a lack of a systematic framework for evaluating the map-reading capabilities of VLMs.

- How can we systematically define map-reading capabilities for VLMs?
- How can we construct an operational benchmark for evaluating these capabilities?
- What are the strengths and limitations of current VLMs in map-reading tasks?

2.Bloom-MR-VLM Cognitive Framework

Based on existing studies in cartography and map cognition, we summarize these literacy into several important skills.

- understanding metadata
- recognizing symbols and labels
- understanding map scale
- judging orientation and direction
- analyzing spatial relationships
- interpreting contour lines
- choosing and planning routes
- understanding thematic information
- performing spatial reasoning



2.Bloom-MR-VLM Cognitive Framework

This framework integrates classical theories from cartographic cognition, map literacy, and Bloom's taxonomy of cognitive processes.

- Map reading is broken down into:

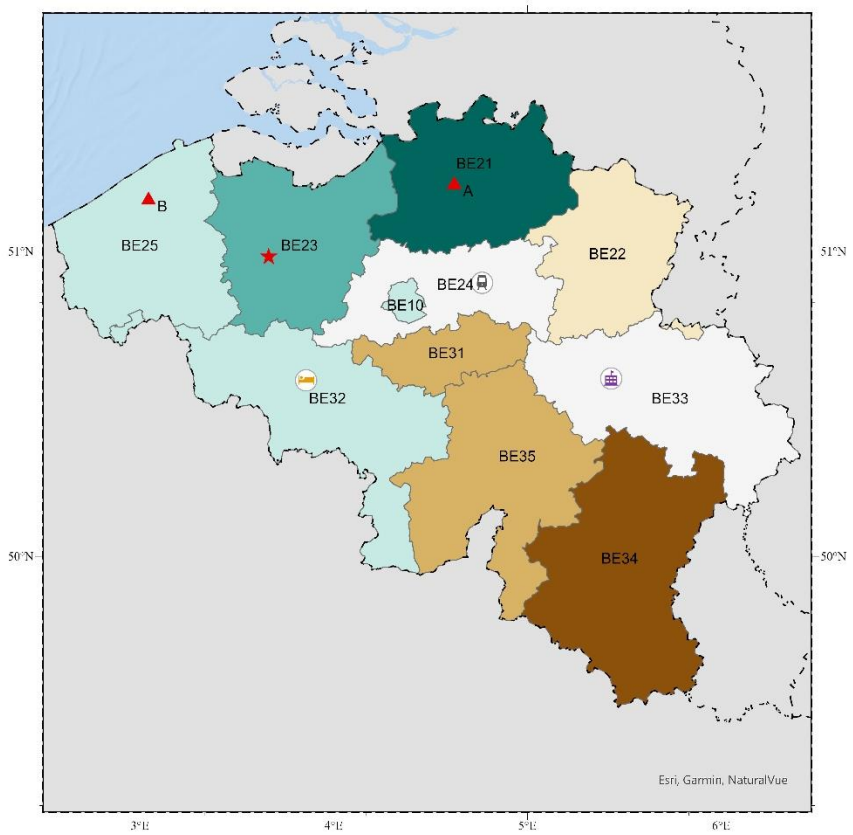
8 core skills

19 subskills

3. Benchmark Construction

The benchmark contains 102 custom-designed choropleth maps.

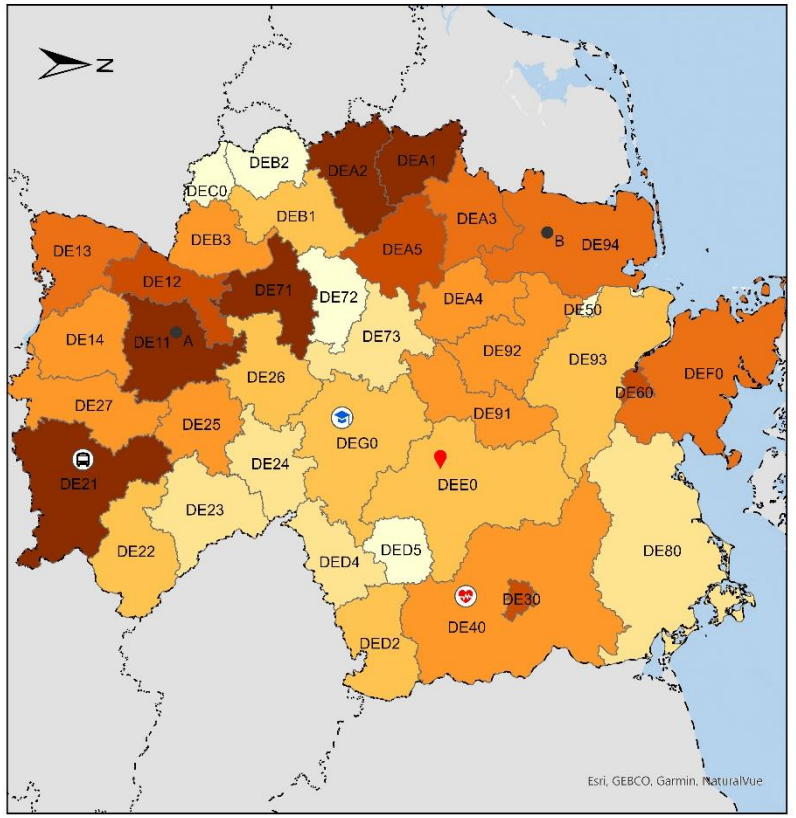
Population of België in 2023



Author: Kate X-Williams
Date: 09 March, 2025

Source: © EuroGeographics for the administrative boundaries
Esri, GEBCO, Garmin, NaturalVue

Gross Domestic Product (GDP) of Deutschland in 2020



Author: Zach X-Wilson
Date: 13 April, 2025

Source: © EuroGeographics for the administrative boundaries
Esri, GEBCO, Garmin, NaturalVue

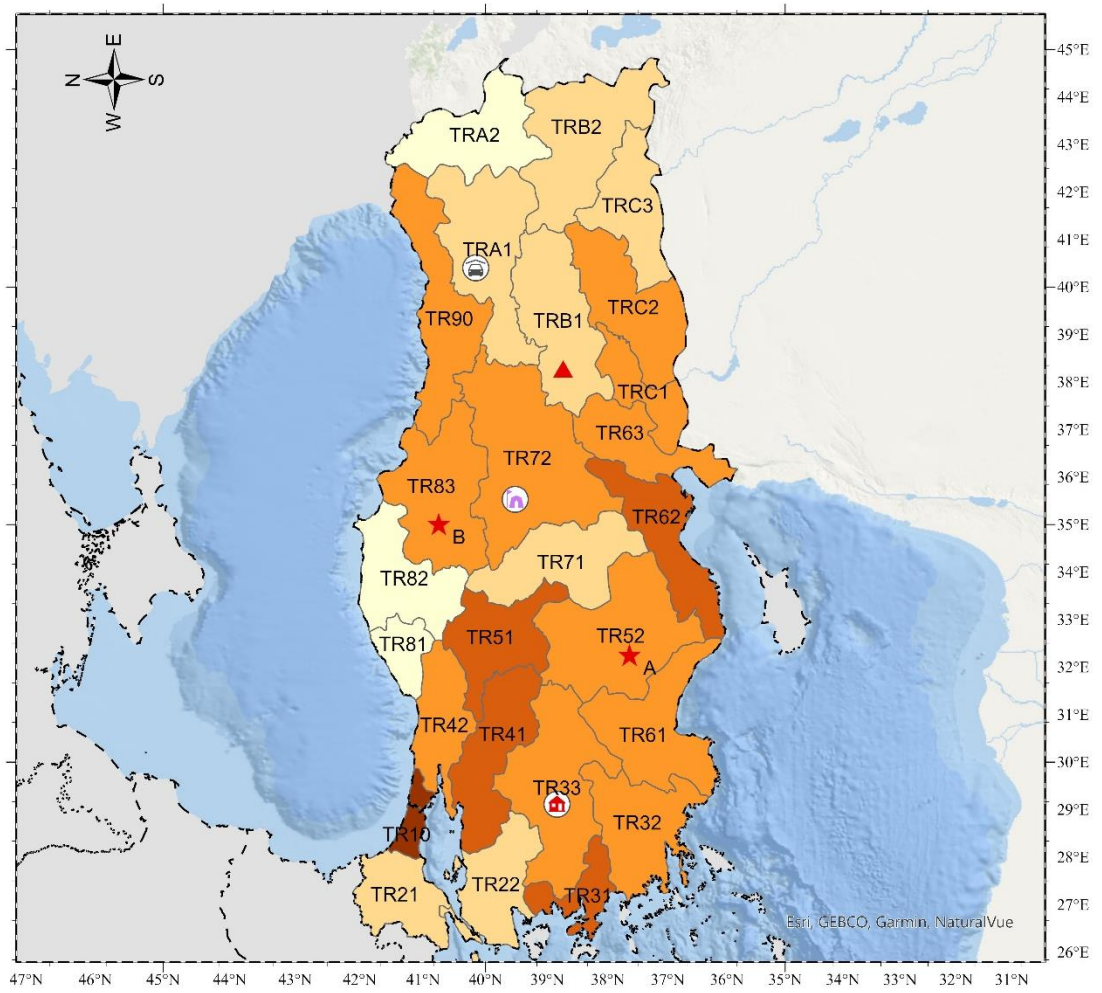
Females of Bulgaria in 2023

Author: Ursula X-Gonzalez
Date: 02/21/25

Source: © EuroGeographics for the administrative boundaries
Esri, GEBCO, Garmin, NaturalVue

3. Benchmark Construction

Available Beds in Hospitals of Türkiye in 2020



Scale: 1:13,225,580

Legend

Point of Interest

- Your location
- Mark
- Hospital
- School
- Subway
- <all other values>
- Boundary

Beds2020_NR

- 0 – 3474
- 3475 – 7038
- 7039 – 10339
- 10340 – 19834
- 19835 – 46382
- Null

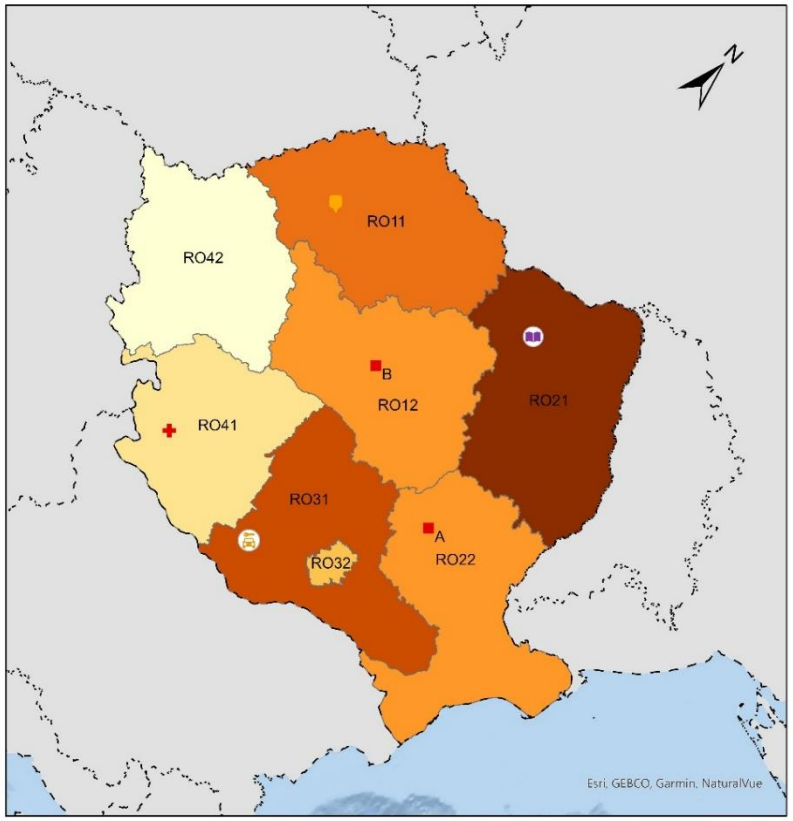
Author: Eve X-Miller

Source: © EuroGeographics for the administrative boundaries

Esri, GEBCO, Garmin, NaturalVue

Date: 2025-03-31

Males of România in 2023



Legend

Name

- Hospital
- Mark
- School
- Subway
- Your location
- <all other values>
- Boundary

Males_NR

- 678515 – 809515
- 809516 – 910271
- 910272 – 1071000
- 1071001 – 1145928
- 1145929 – 1233929
- 1233930 – 1387318
- 1387319 – 1592780
- Null

Scale: 1:6,696,155

Projection: Albers Equal-Area Conic

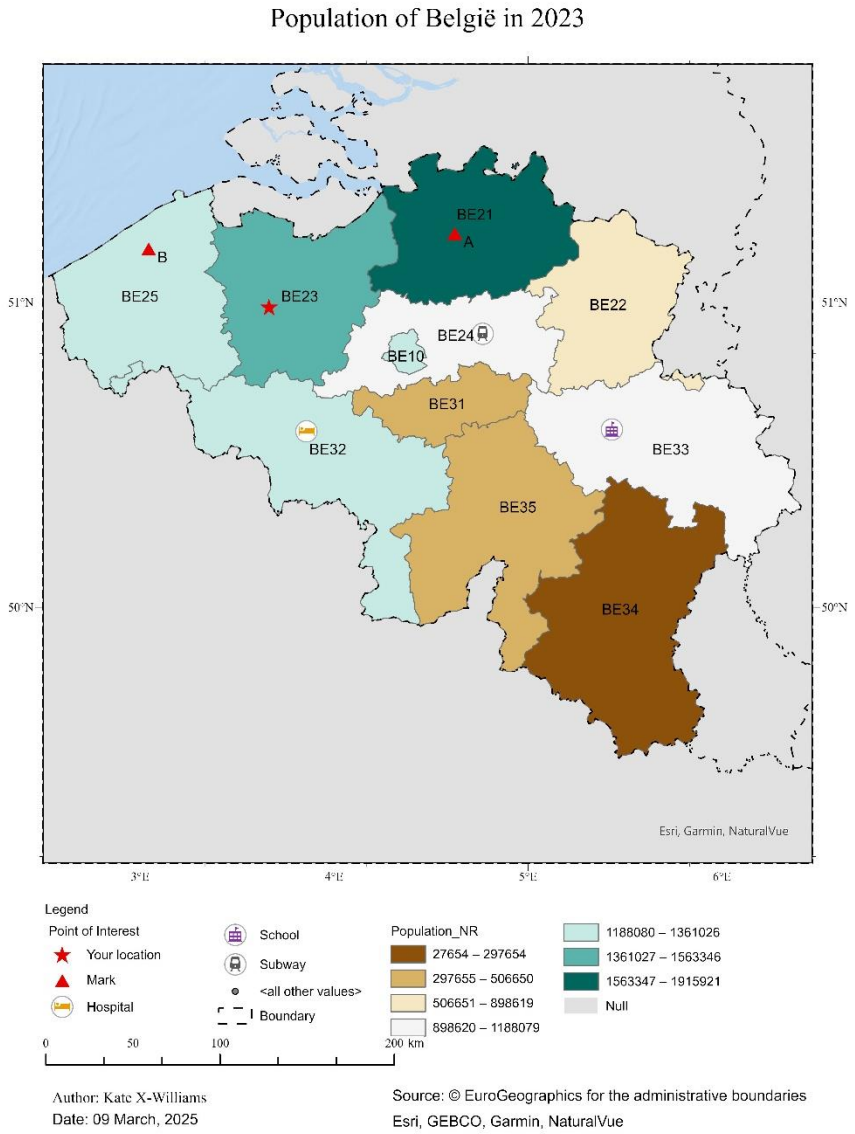
Author: Jack X-Brown

Date: March 02, 2025

Source: © EuroGeographics for the administrative boundaries

Esri, GEBCO, Garmin, NaturalVue

3. Benchmark Construction



Map: Populaiton of Belgium in 2023

Question: List all map elements included in the given map. Provide your answer in the format: [Title] [Legend] etc.

Answer: [Title], [Legend], [Scale bar/ Scale], [Author], [Date], [Source], [Coordinate grid/ Latitude and longitude lines/grid/Graticule/Geographic Grid/Latitude and Longitude Grid]

……

Map: Populaiton of Belgium in 2023

Question: What is the direction of Mark A in relation to Mark B? Enclose each answer in square brackets like this: [Answer].

Answer: [northeast]

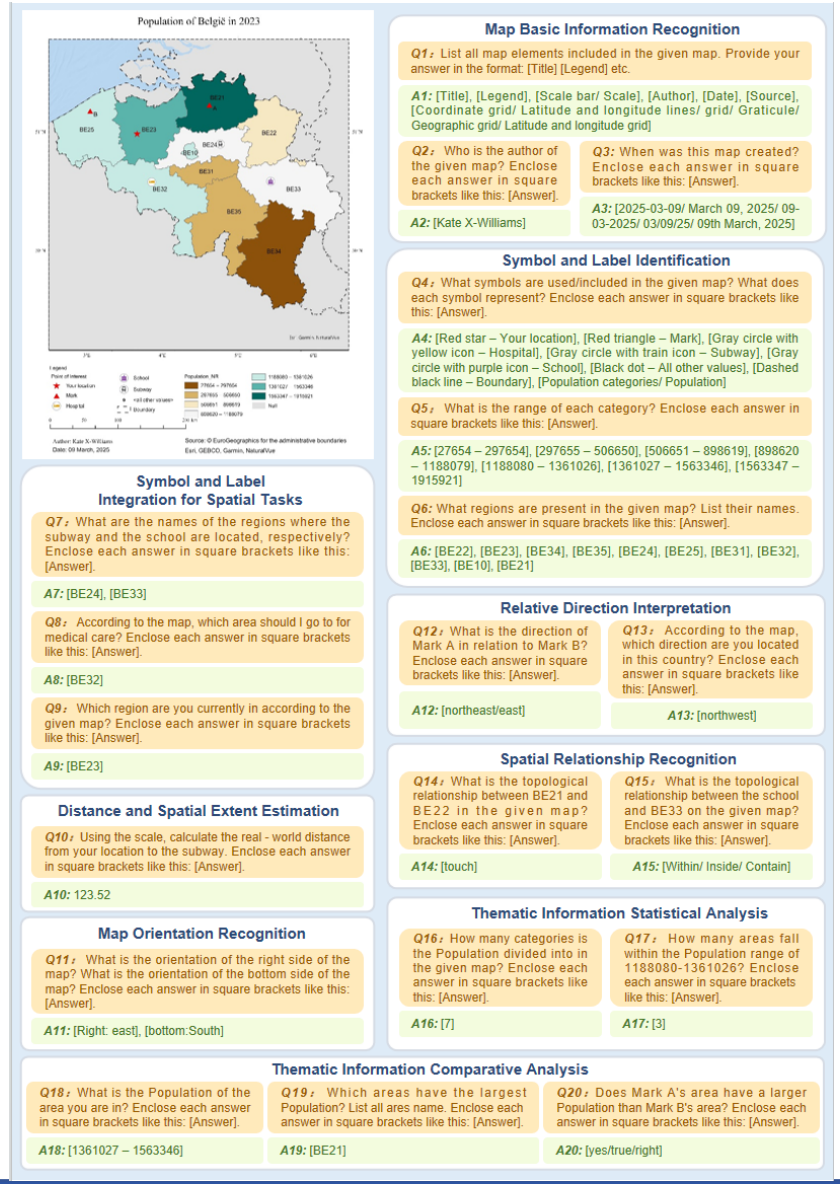
4. Experiment

Case Study

- 102 Choropleth maps
- 9 representative subskills
- 2040 MQA samples

Experiment

- Automated scoring combined with expert manual review
- 5 representative vision-language models tested：
- GPT-4.1, GPT-4o, Gemini 2.0 Flash, Gemini 2.5 Pro, and InternLM3-8B-Instruct.



5. Results

5.1 Structured Output and Fallback Behavior

Most models were able to generate stable structured outputs.

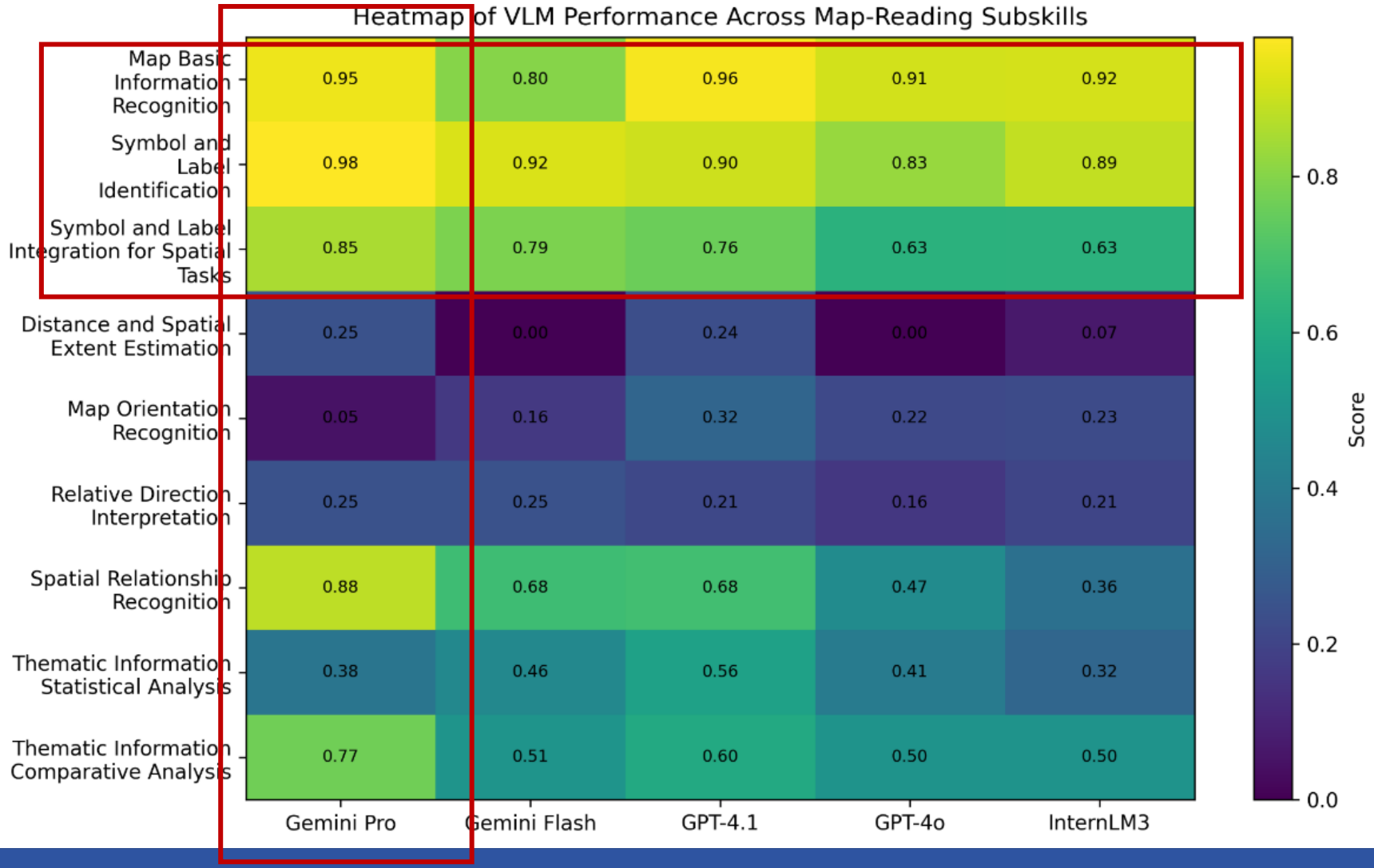
Table 1 Summary of answer regularity and fallback responses across models.

Model	Structured Output Rate (%)	Fallback Rate (%)
GPT-4o	90.88	6.81
GPT-4.1	99.90	0.10
Gemini Flash 2.0	98.09	0.88
Gemini Pro	100.00	0.00
InternLM3	99.95	0.05

This suggests that GPT-4o adopted a more conservative answering strategy, while other models tended to answer all questions even when uncertain.

5. Results

5.2 Overall Performance Across Subskills



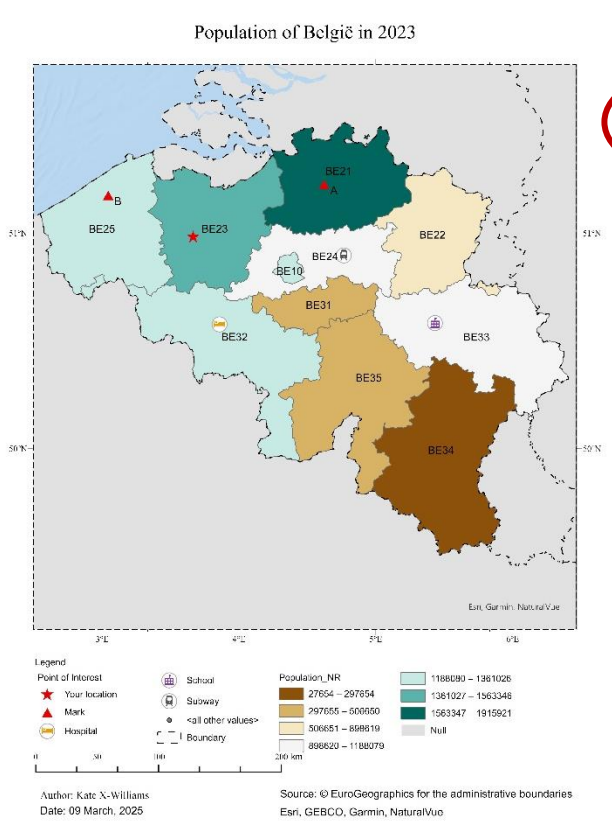
All models show similar performance patterns.

- perform well on low-level tasks
- performance drops significantly on advanced tasks
- Gemini Pro achieves the most stable and strongest performance

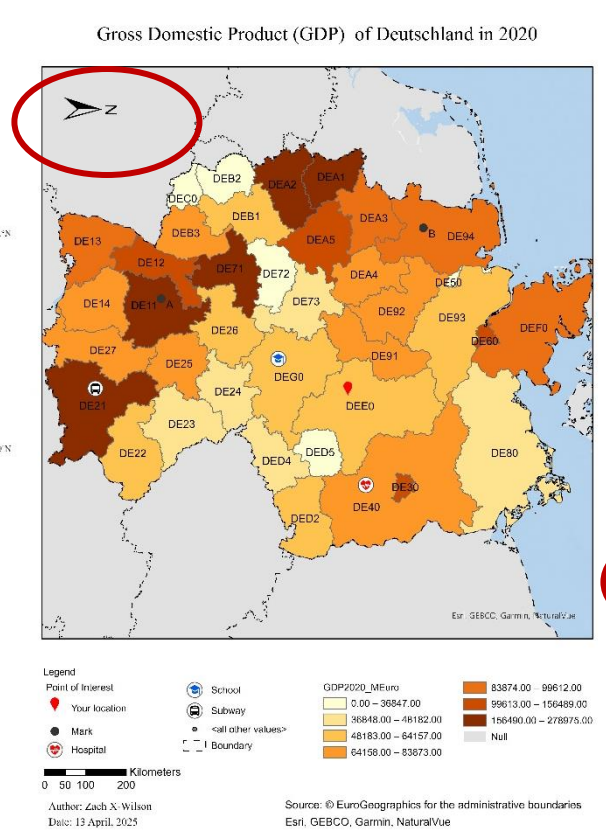
5. Results

5.3 Weakness in Spatial Reasoning

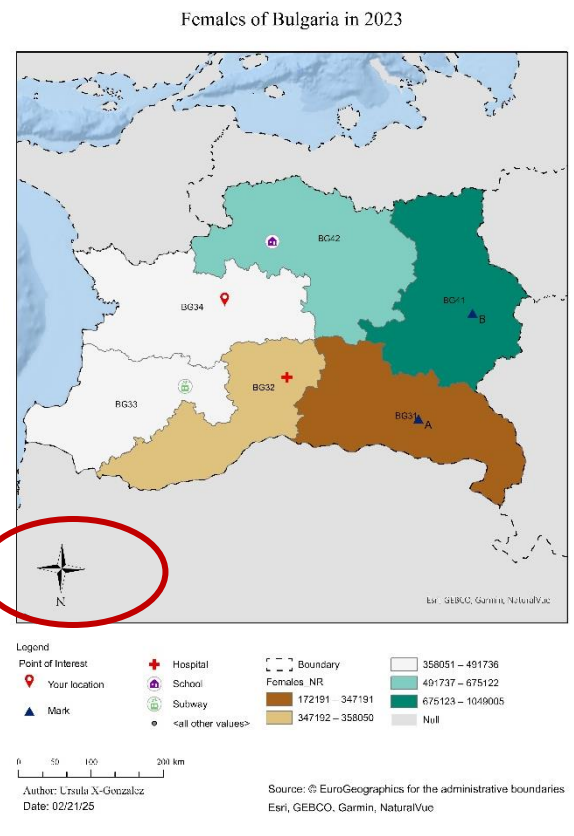
many models fail when the map orientation is changed.



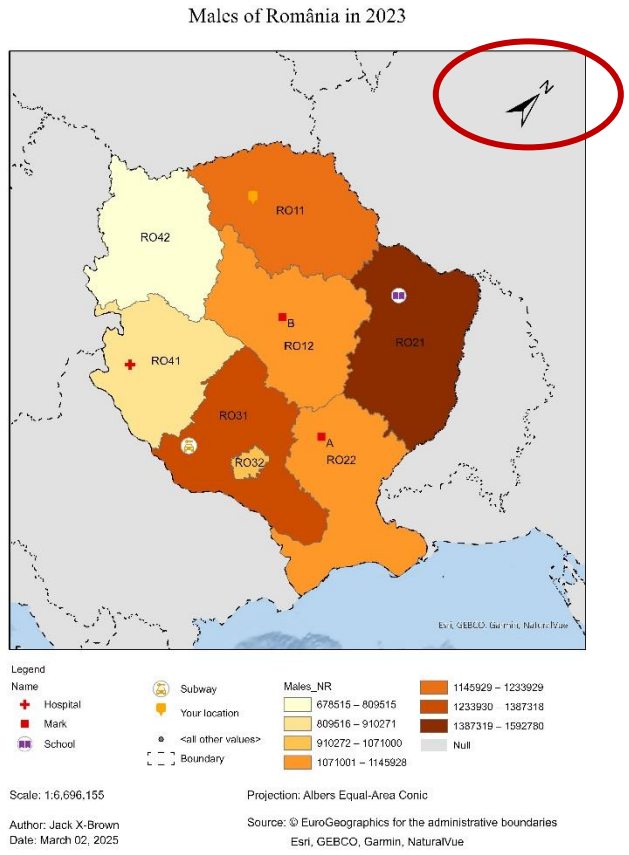
(√)



(×)



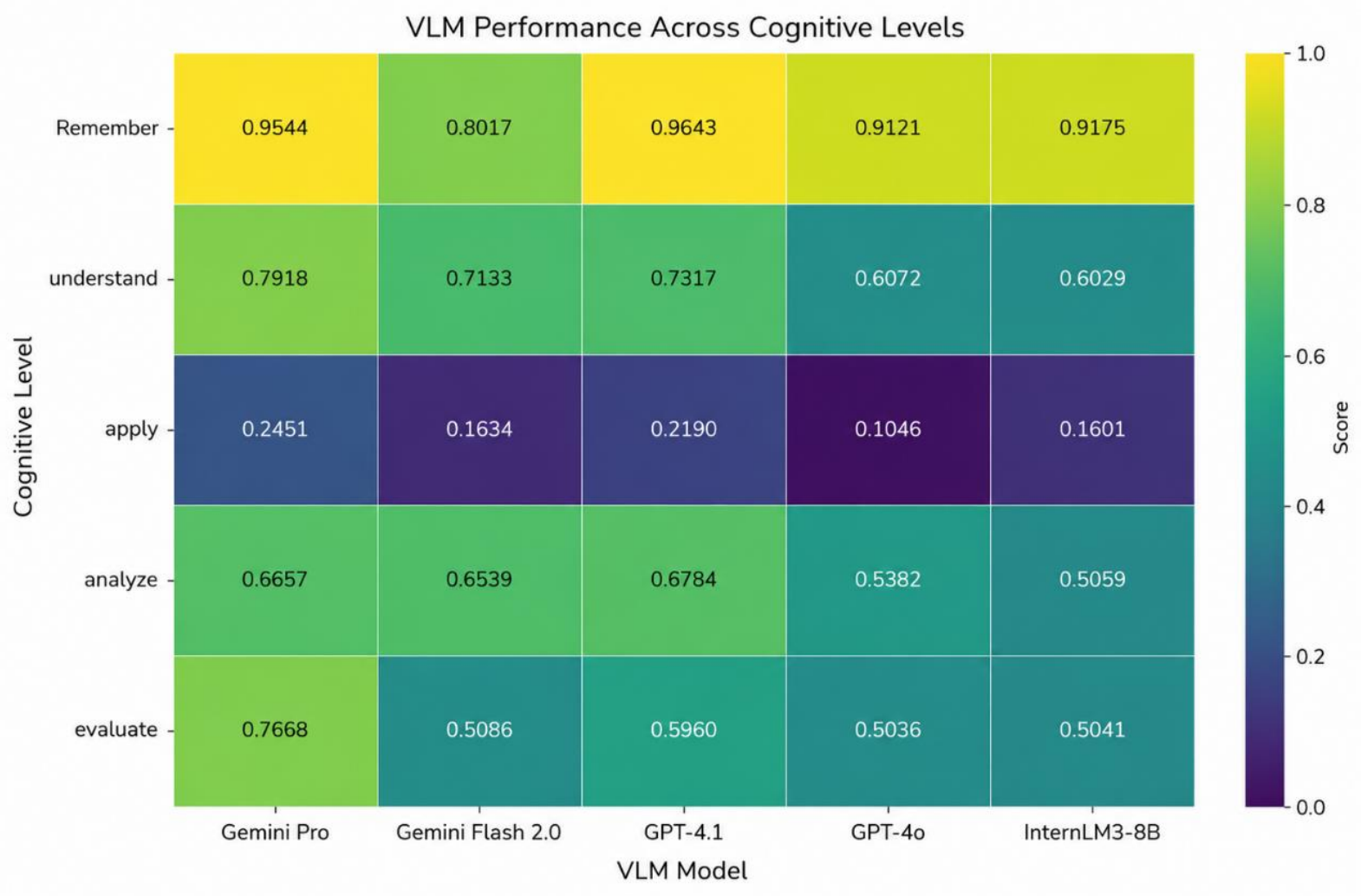
(×)



(×)

5. Results

5.4 Performance Across Cognitive Levels



The models perform very well at lower cognitive levels

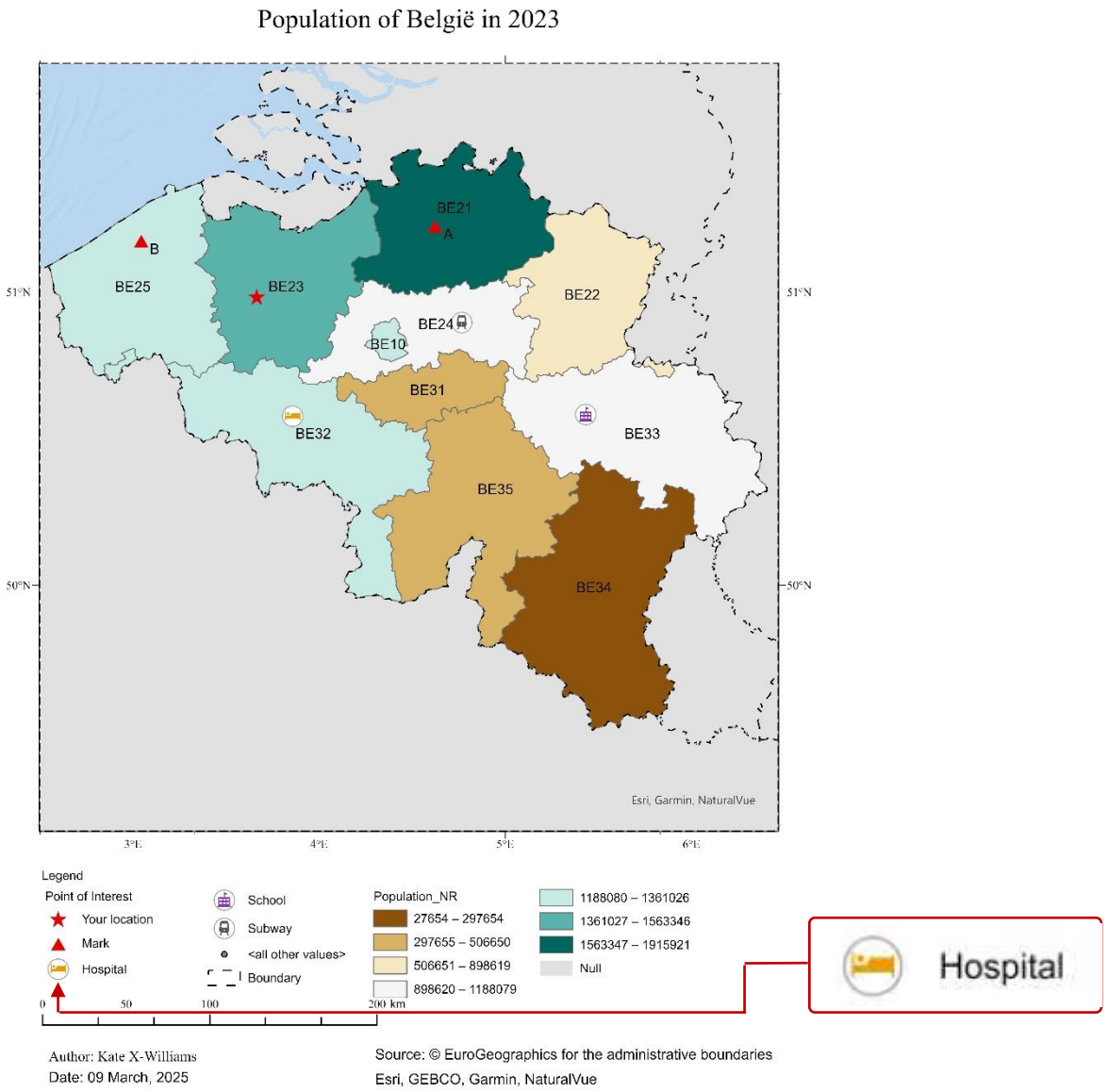
Performance decreases substantially at higher cognitive levels

This indicates that current VLMs are strong at semantic recognition but still limited in complex geographic cognition.

5. Results

5.5 Typical Error Patterns

- Cartographic Symbol Misunderstanding
- Spatial Orientation and Positioning Bias
- Metric Reasoning Failure
- Spatial Relation Reasoning Errors
- Hallucination in Map Interpretation
- Response Compliance Failure



6. Discussion and Conclusion

- Evaluating map-reading capability is much more complicated than general visual question answering tasks.
- Current VLMs still lack strong spatial understanding.
- The importance of integrating geospatial cognition into future multimodal AI systems.

Thank you!

Presenter: Yangyang Zhao

Affiliation: Chinese Academy of Surveying and Mapping, China

Ghent University, Belgium

Email： Zhaoyy@casm.ac.cn