AIST
Create the Future, Collaborate Together

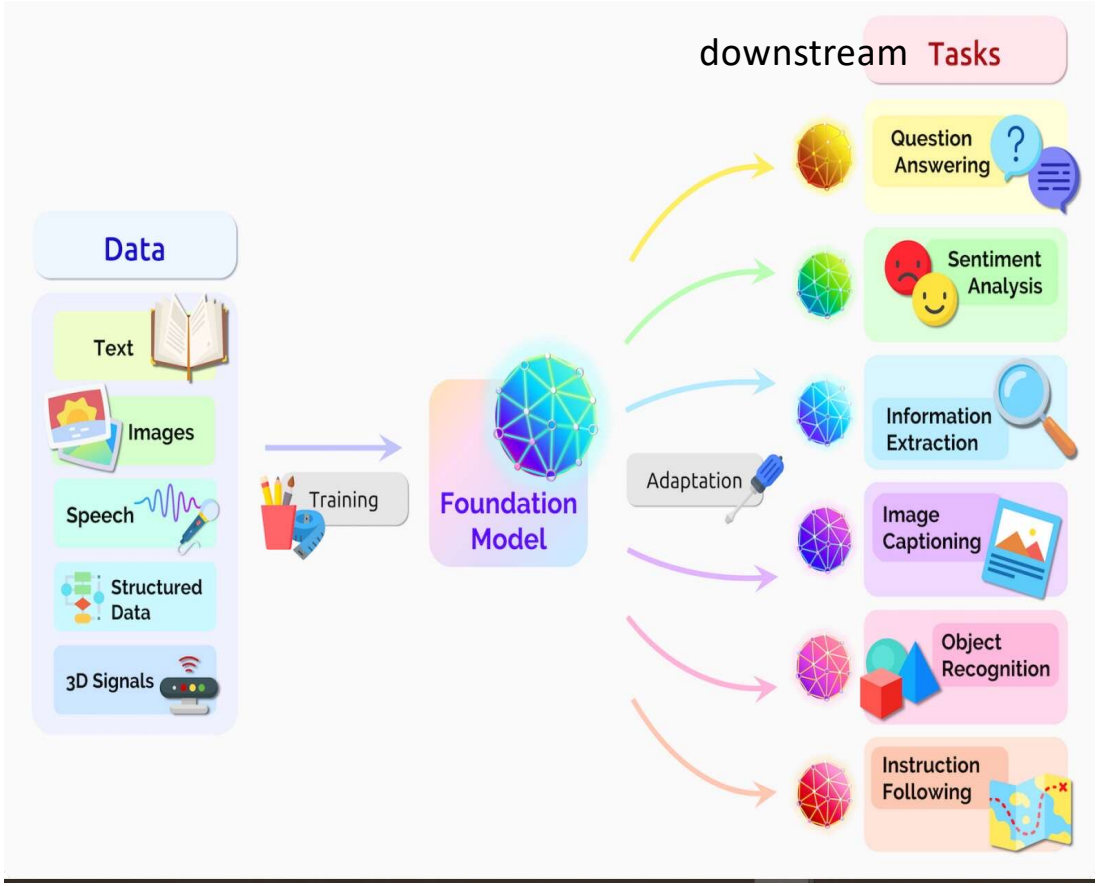
Evaluation of Geospatial Foundation Models

Kyoung-Sook KIM (ks.kim@aist.go.jp)
Deputy Director
Intelligent Platforms Research Institute (IPRI)
AIST, Japan

2026.01.28
TPC @ SCA / HPC Asia 2026, Osaka, Japan

The Success of Foundation Models

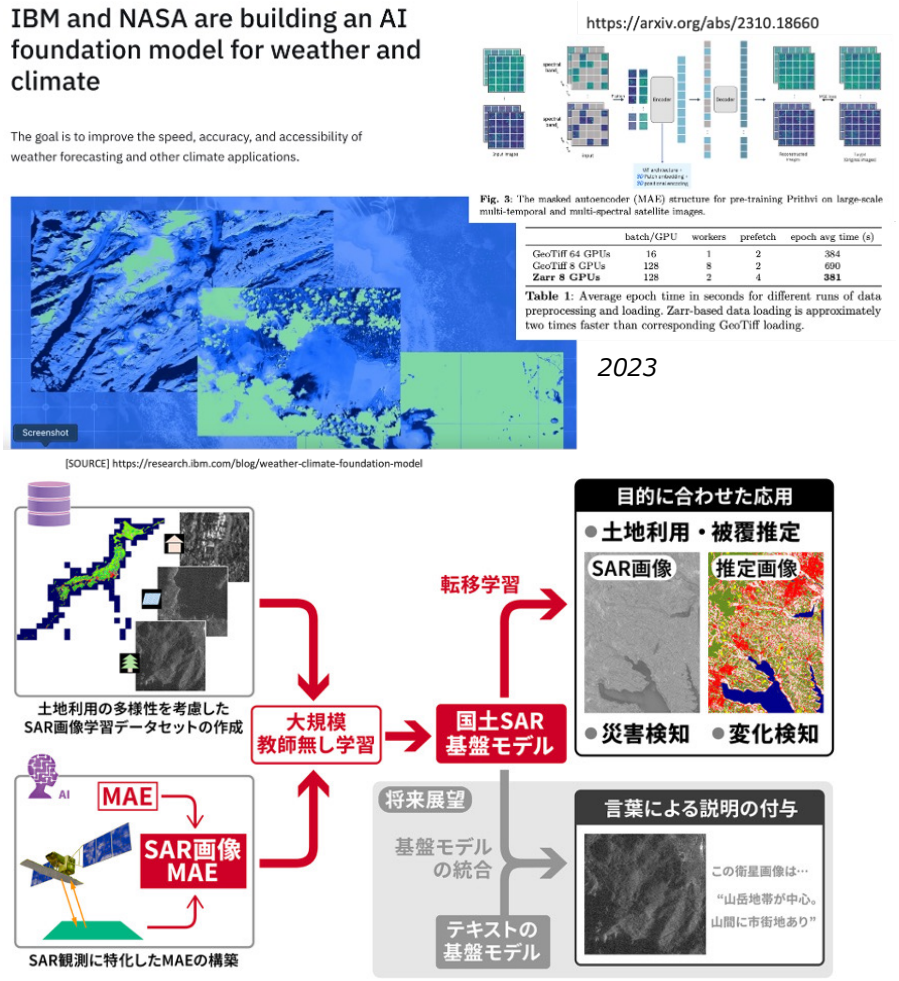
- Foundation models
  - **Large-scale** pretrained models on vast amounts of data adapted to a **wide range of downstream tasks.**
  - Achieved breakthrough performance across many domains.
    - Natural language processing (NLP), computer vision, multi-modal systems, and specialized domains.
  - Success recipe in NLP
    - Transferable representations of their core units (e.g., words, pixels).
    - SSL aligned with massive unlabeled datasets.
    - Adaptability of backbone to many downstream tasks.
    - Already human-generated data.
    - Broad population contribution.



SOURCE：On the Opportunities and Risks of Foundation Models, 2021, https://arxiv.org/abs/2108.07258

Geospatial Foundation Models

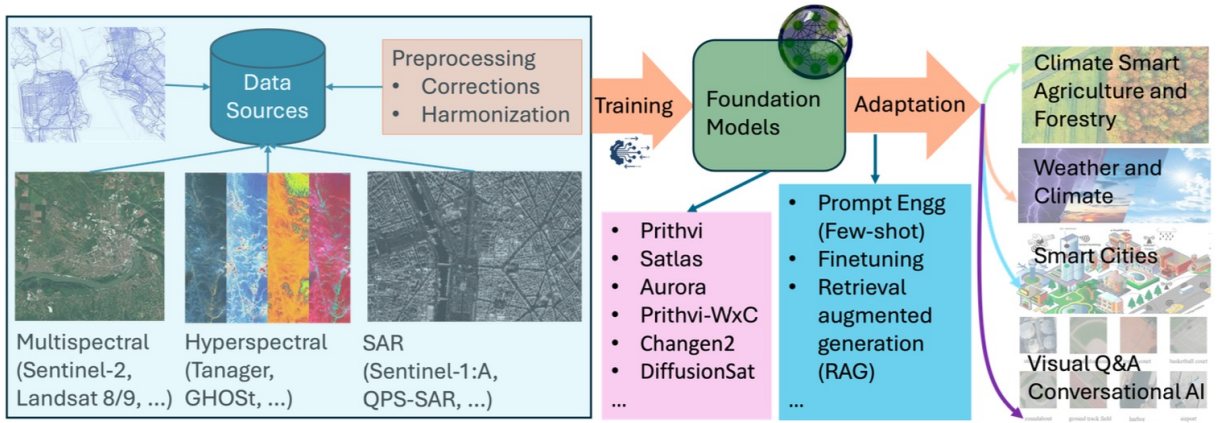
- Motivations
  - **Massive data growth**
    - Satellite (e.g., Sentinel, Landsat): TBs/day globally
    - Includes imagery, sensors, maps, GPS, social media, etc.
  - **Manual processing cost**
    - Requires automated understanding and analysis
  - **Limited labeled data**
    - Reusability and transferability are essential
  - **Dynamic environments (e.g., disasters, cities)**
    - Scalability, efficiency, and adaptability are critical



神山 研究グループ: 人工衛星「だいち2号」の観測データを活用して国土に特化したSAR基盤モデルを構築
[SOURCE] https://www.aist.go.jp/aist_j/press_release/pr2025/pr20250603/pr20250603.html

Geospatial Foundation Models

- Large-scale, pre-trained artificial intelligence systems that learn universal spatial, temporal, and semantic representations, enabling scalable and data-efficient adaptation across diverse Earth observation and geospatial tasks—including segmentation, change detection, land-cover classification, retrieval, biophysical estimation, and vector-based spatial reasoning.
- **Key requirements**
  - Heterogeneity of data
  - Scale and coverage
  - Transferability and scalability
  - Modal fusion and alignment



[SOURCE] Ranga Raju Vatsavai. 2024. Geospatial Foundation Models: Recent Advances and Applications. In Proceedings of the 12th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (BigSpatial '24). Association for Computing Machinery, New York, NY, USA, 30–33. https://doi.org/10.1145/3681763.3698478

Geospatial Foundation Models

- Not just pixel-based vision models for remote sensing images
  - classification, segmentation, and object detection tasks.

Detection of Solar PV power plants from satellite imagery

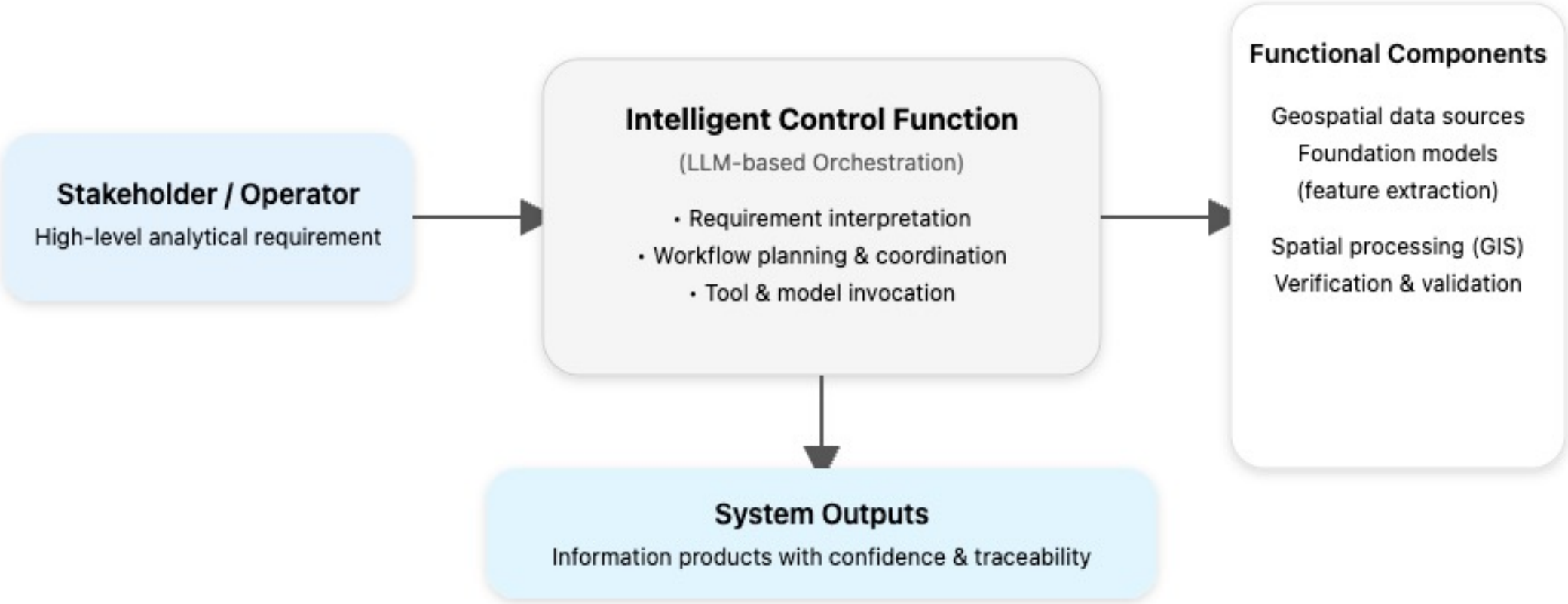
- Not just multimodal models (VLM) from images and text
  - from visual question answering to image captioning.

How many building are?

Left side: dense residential neighborhood
→ small, closely packed houses
Right side: industrial / utility area
→ fewer but larger structures (tanks, warehouses)
- Residential area (left): ~210–230 buildings
- Industrial / utility area (right): ~25–35 buildings

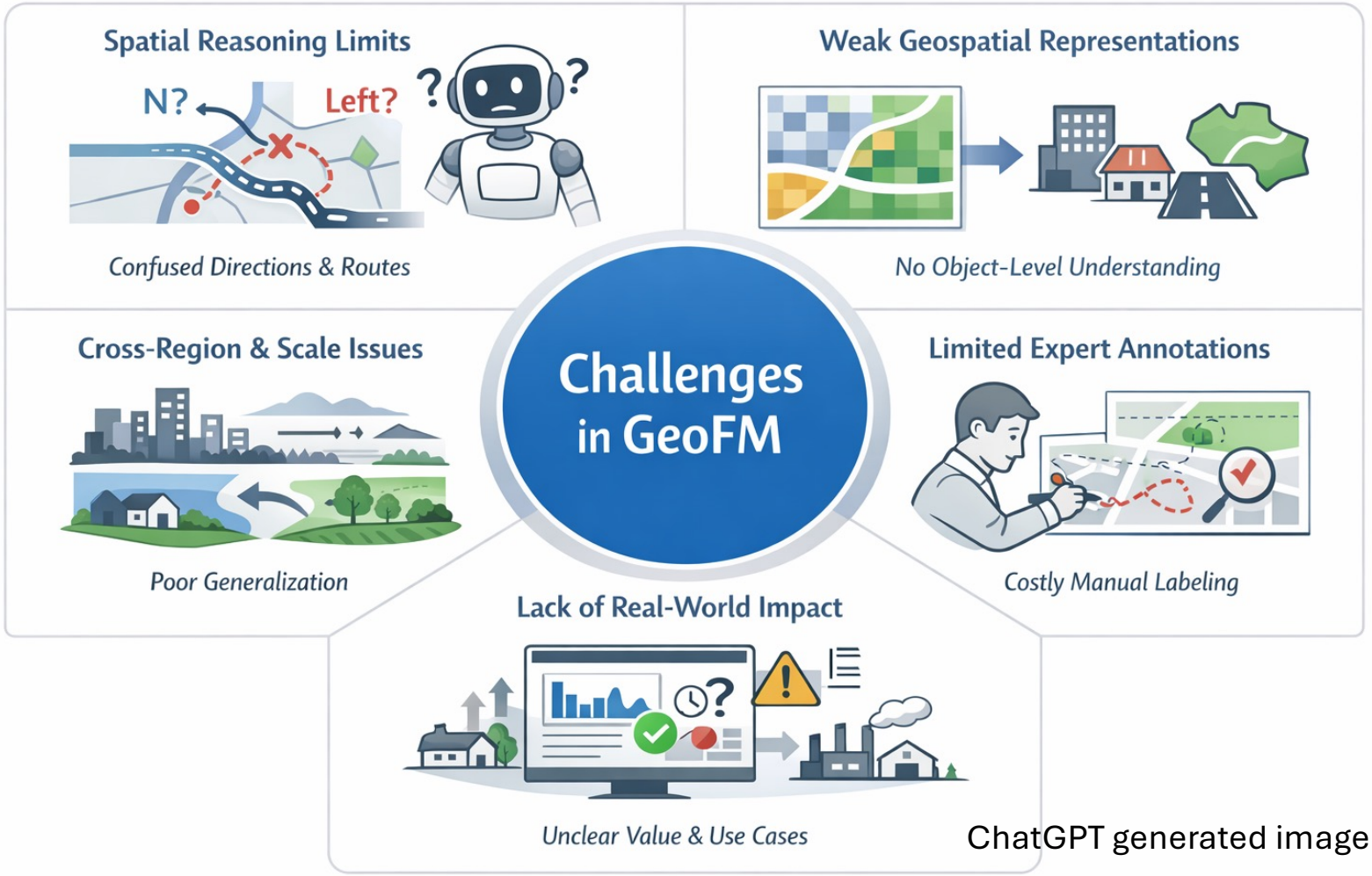
Total estimated buildings: ≈ 240–260 buildings

Agentic AI (LLM orchestrating geospatial tools + foundation models)



Challenges in Geospatial Foundation Models

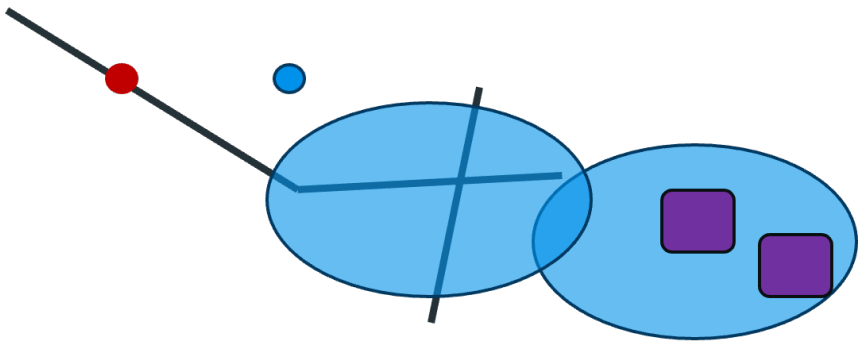
- Weak spatial reasoning
- Multi-modal data heterogeneity
- Cross-region generalization
- Limited annotations
- Lack of real-world deployment examples



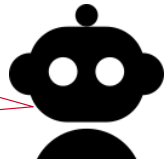
ChatGPT generated image

Geospatial Reasoning Limtits

- LLMs struggle to reason over metric space, topology, and spatial relationships.
  - Confusion of left/right, north/south, adjacency, containment
  - Route planning errors or misinterpreted spatial relations

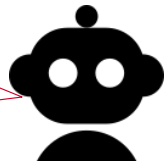


“The hospital is north of the river. The school is east of the hospital. Which direction is the school from the river?”



The school is northeast of the river.

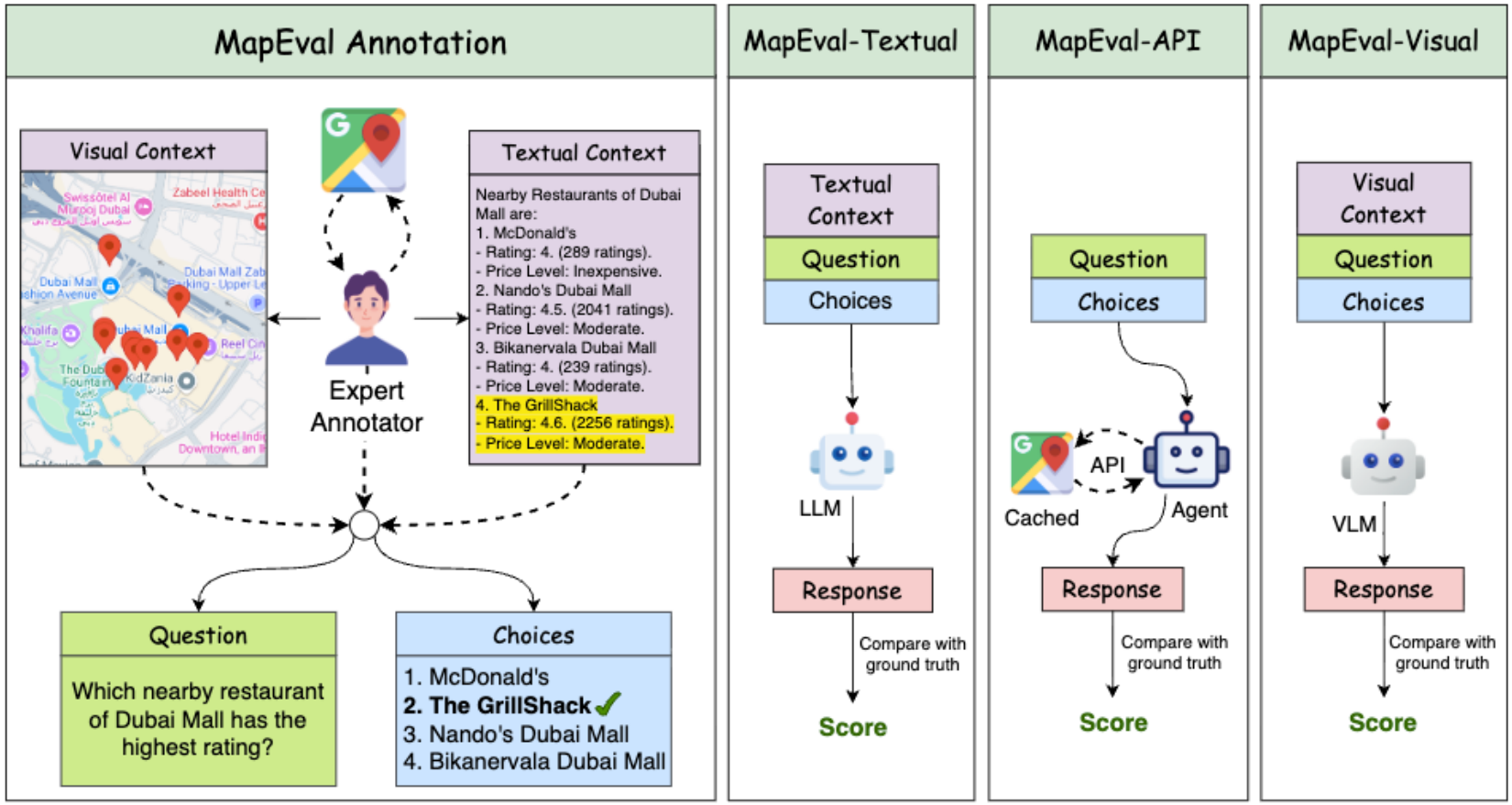
A park is inside District A. A library is next to the park. Is the library inside District A?”



The library could be outside District A.

Geospatial Reasoning Limtits

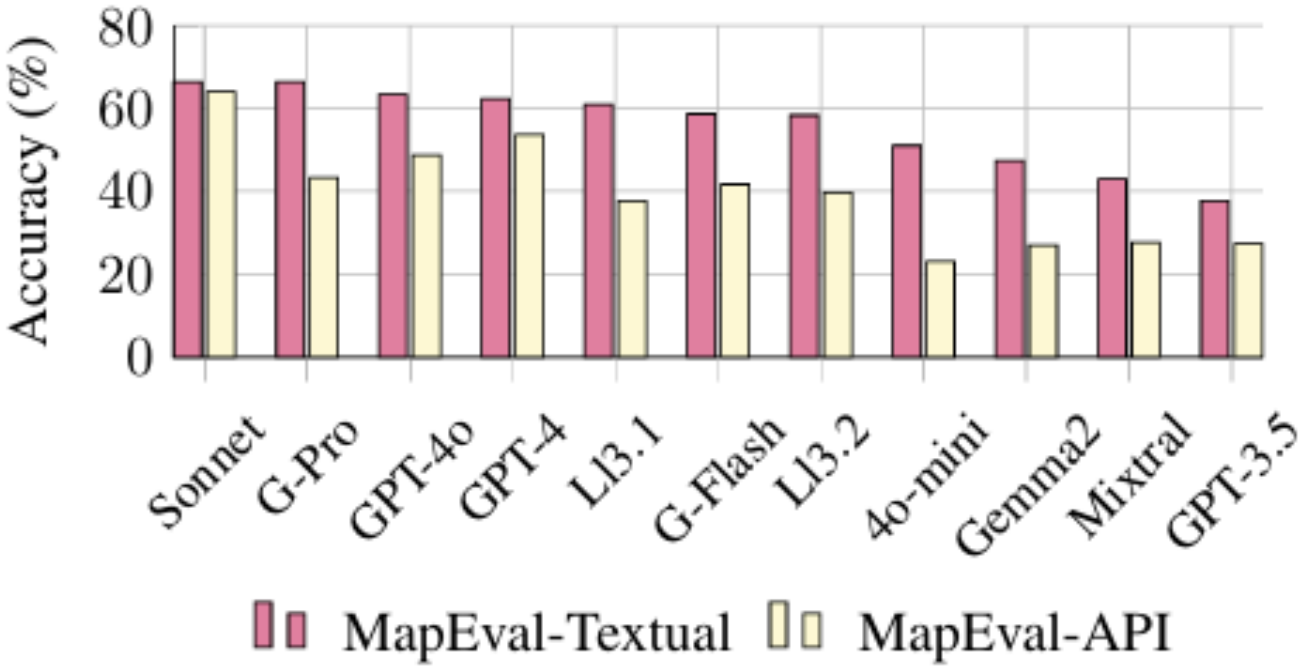
Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models
ML Dihan et al. (2024)



SOURCE: https://mapeval.github.io/

Geospatial Reasoning Limtits

Type	Task	Question Example	Count
Place Info	Textual/API	What is the direction of Victoria Falls from Harare?	64
	Visual	Is there any Hospital marked with a star symbol on the tourist map of Rome?	121
Nearby	Textual/API	Find restaurants nearby Louvre Museum above 4.0 rating.	83
	Visual	I stayed at SpringHill Suites by Marriott Portland Hillsboro. Can you recommend the	91
Routing	Textual/API	I am driving to Lockbourne Rd	
	Visual	What is the fast	
Unanswerable	Textual/API	Which road sh roads in heavy	
	Visual	Which way sh KONO so that	
Trip	Textual/API	I have an aftern Museum of Art nearby cafe, an ensure I have e	
Counting	Visual	How many hos	



https://arxiv.org/pdf/2501.00316

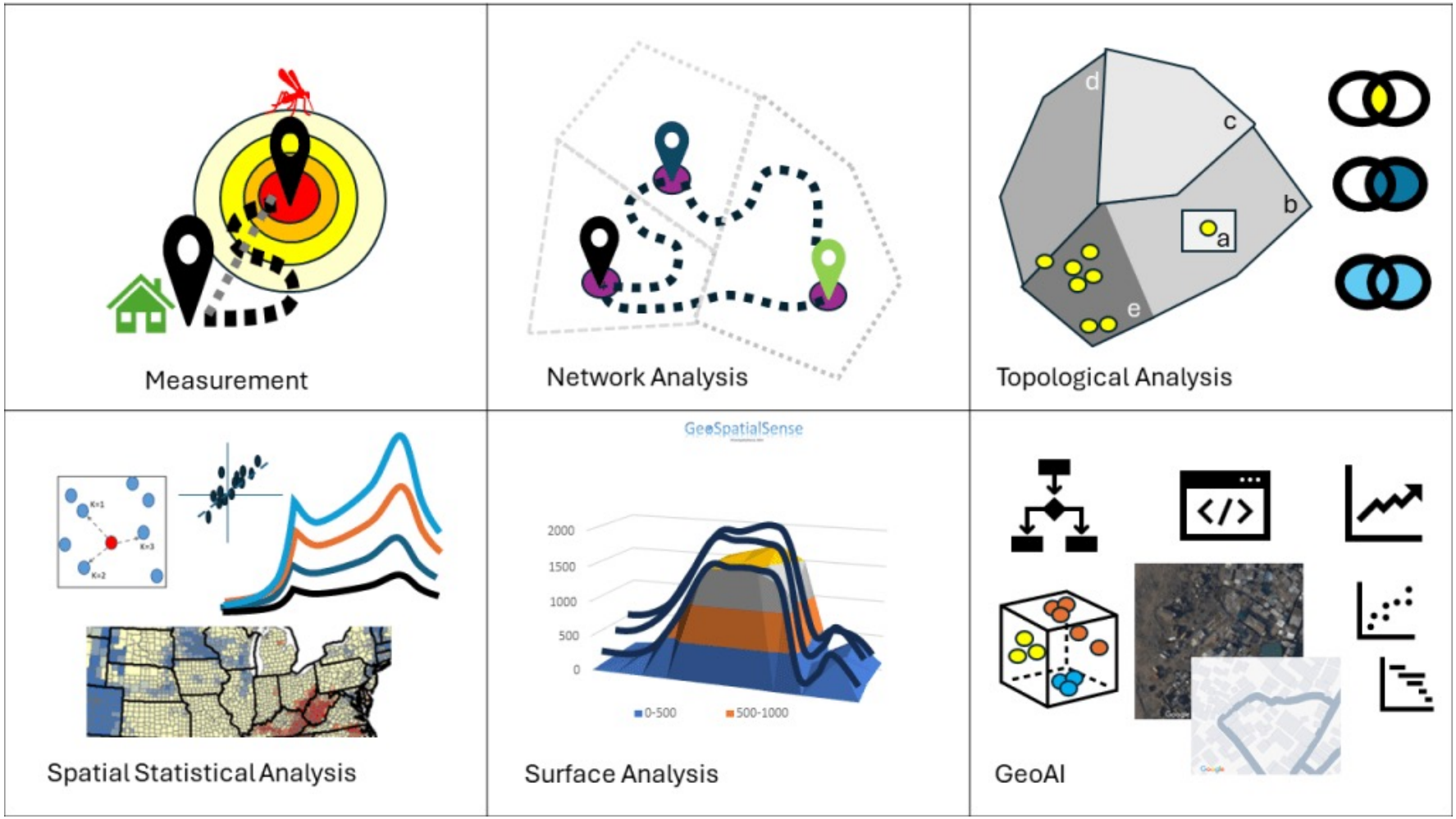
Geospatial Reasoning Limtits

MapEval-Visual evaluation

Model	Overall (#400)	Place Info (#121)	Nearby (#91)	Routing (#80)	Counting (#88)	Unanswerable (#20)
Close-Source (Proprietary) VLMs						
Claude-3-5-Sonnet (Anthropic, 2024)	**61.65**	**82.64**	55.56	45.00	**47.73**	**90.00**
GPT-4o (OpenAI, 2024)	58.90	76.86	**57.78**	**50.00**	**47.73**	40.00
Gemini-1.5-Pro (Team et al., 2024a)	56.14	76.86	56.67	43.75	32.95	80.00
GPT-4-Turbo (OpenAI, 2023)	55.89	75.21	56.67	42.50	44.32	40.00
Gemini-1.5-Flash (Team et al., 2024a)	51.94	70.25	56.47	38.36	32.95	55.00
GPT-4o-mini (OpenAI, 2024)	50.13	77.69	47.78	41.25	28.41	25.00
Open-Source VLMs						
Qwen2.5-VL-72B (Bai et al., 2025)	**60.35**	**76.86**	**54.44**	43.04	**52.33**	**90.00**
Qwen2-VL-7B (Wang et al., 2024)	51.63	71.07	48.89	40.00	40.91	40.00
Llama3.2-90B-Vision (AI@Meta, 2024)	50.38	73.55	46.67	41.25	36.36	25.00
Glm-4v-9b (GLM et al., 2024)	48.12	73.55	42.22	41.25	34.09	10.00
InternLm-Xcomposer2 (Dong et al., 2024)	43.11	50.41	48.89	**43.75**	34.09	10.00
MiniCPM-Llama3-V-2_5 (Yao et al., 2024)	40.60	60.33	32.22	32.50	31.82	30.00
Llama-3-VILA1.5-8B (Lin et al., 2023)	32.99	46.90	32.22	28.75	26.14	5.00
Llava-v1.6-Mistral-7B-hf (Liu et al., 2024b)	31.33	42.15	28.89	32.50	21.59	15.00
DocOwl1.5 (Hu et al., 2024)	31.08	43.80	23.33	32.50	27.27	0.00
Paligemma-3B-mix-224 (Beyer et al., 2024)	30.58	37.19	25.56	38.75	23.86	10.00
Llava-1.5-7B-hf (Liu et al., 2024a)	20.05	22.31	18.89	13.75	28.41	0.00
Human Performance						
Human	82.23	81.67	82.42	85.18	78.41	65.00

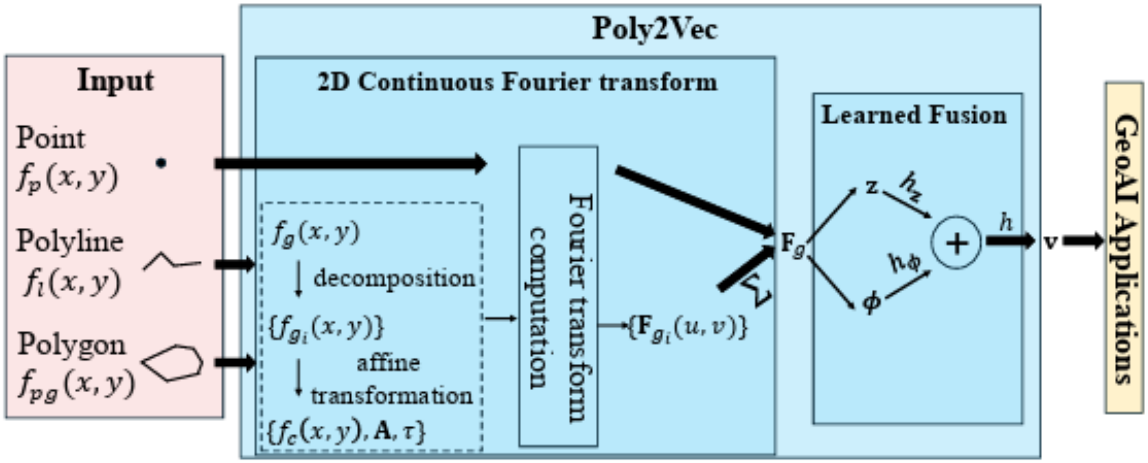
https://arxiv.org/pdf/2501.00316

Weak Geospatial Representations



ChatGPT generated image

Encodings of Geospatial Objects



(a) The workflow of POLY2VEC.

Maria Despoina Siampou, Jialiang Li, John Krumm, Cyrus Shahabi, Hua Lu.
Poly2Vec: Polymorphic Fourier-Based Encoding of Geospatial Objects for GeoAI Applications
(https://arxiv.org/abs/2408.14806)

Chen Chu, Cyrus Shahabi.
Geo2Vec: Shape- and Distance-Aware Neural Representation of Geospatial Entities
(https://arxiv.org/pdf/2508.19305)

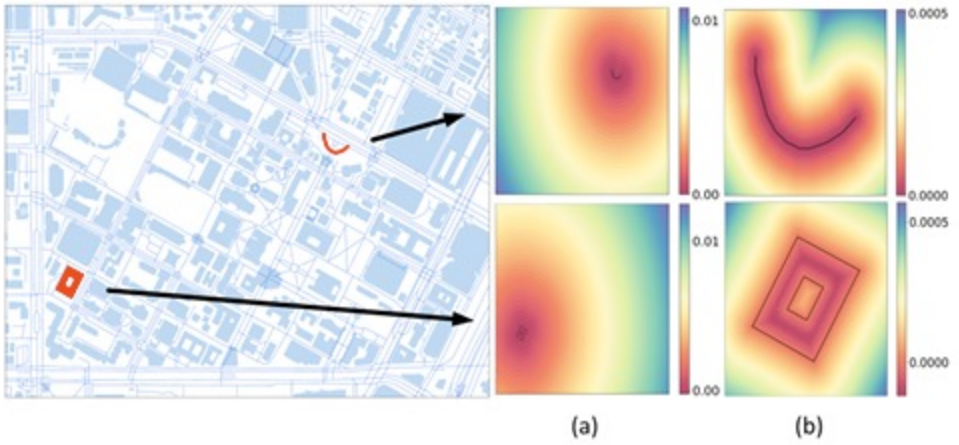
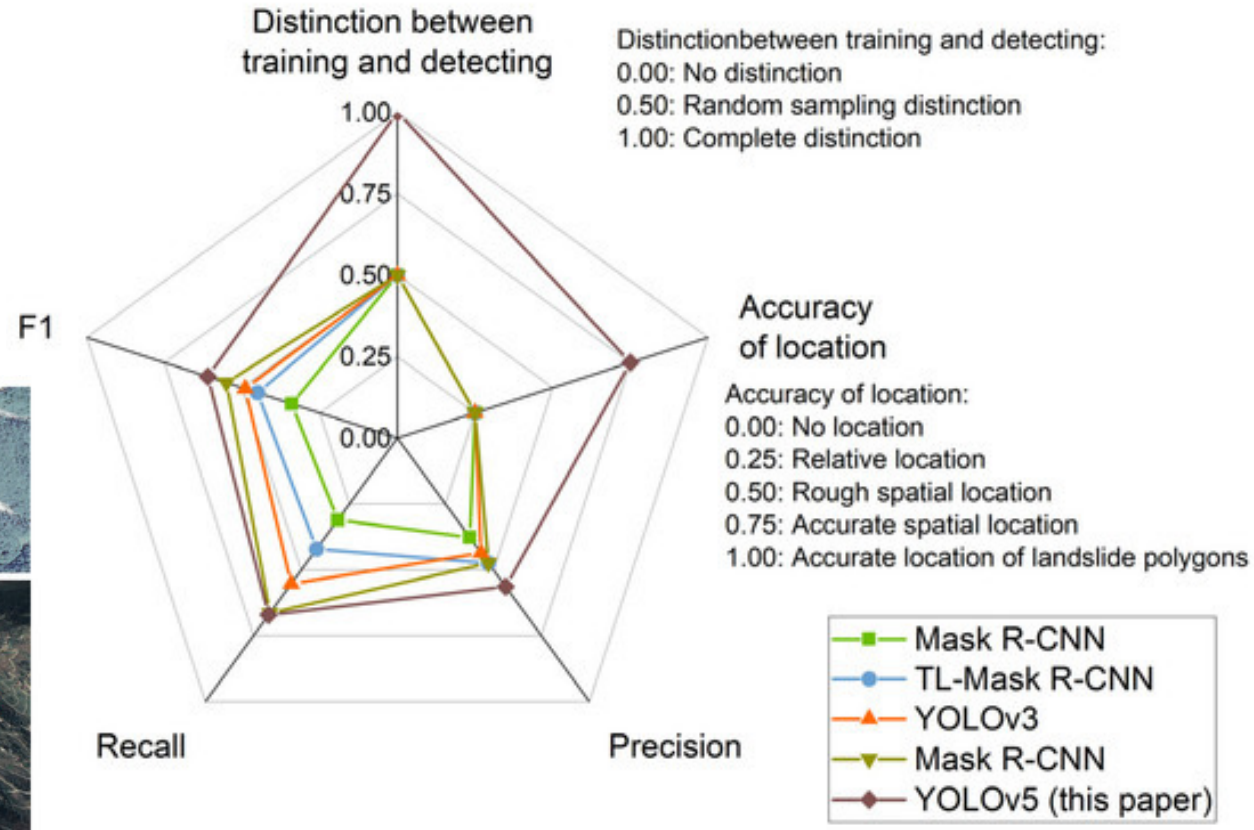


Figure 1: Signed Distance Fields for two types of geo-entities at spatial scales: (a) coarse scale, (b) fine scale.

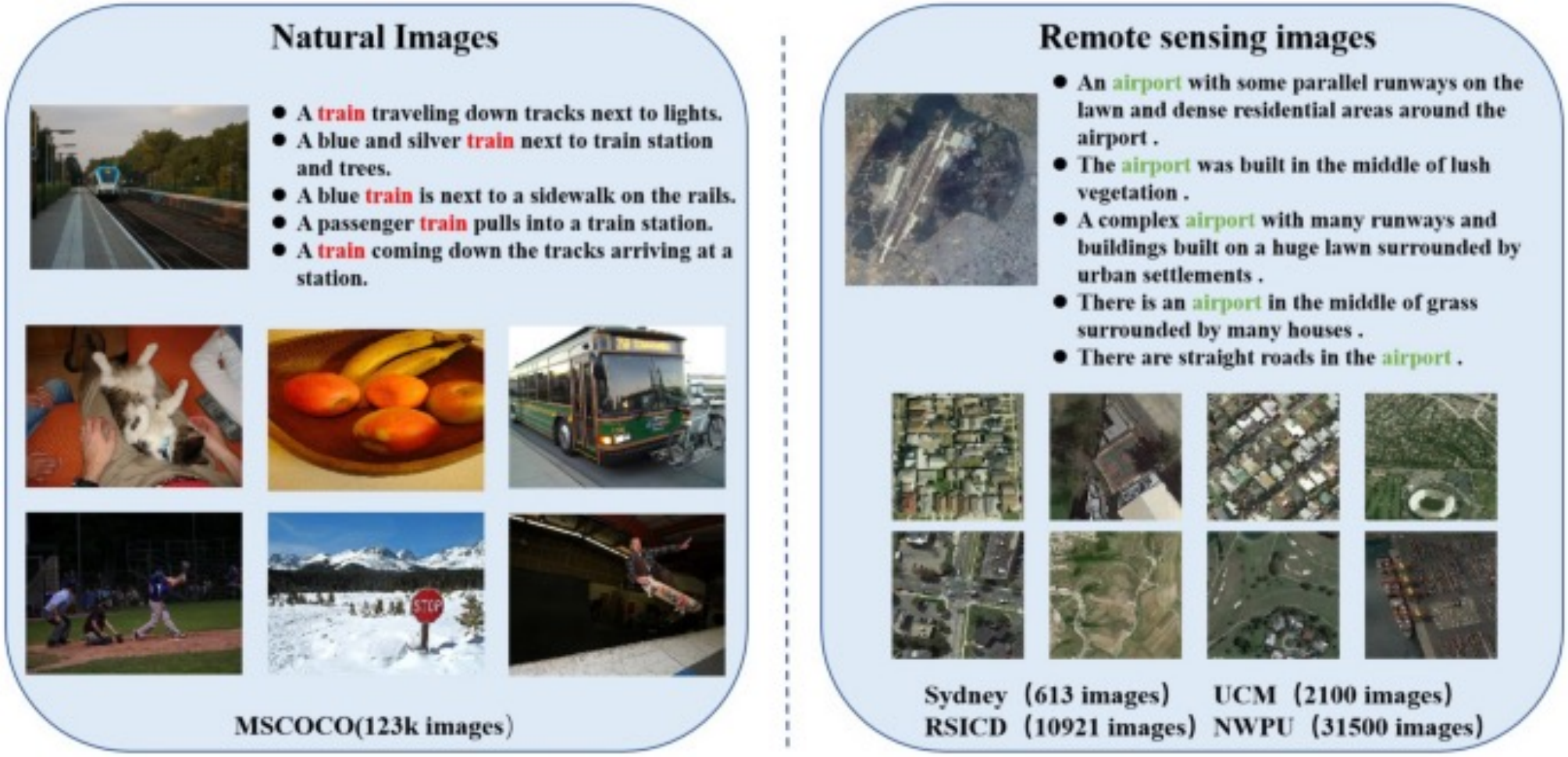
Cross Region & Scale Issues

- A YOLOv5-based landslide detector achieved F1 = 0.96 in the training region but only mediocre precision and recall in a different region under a cross-regional test scenario



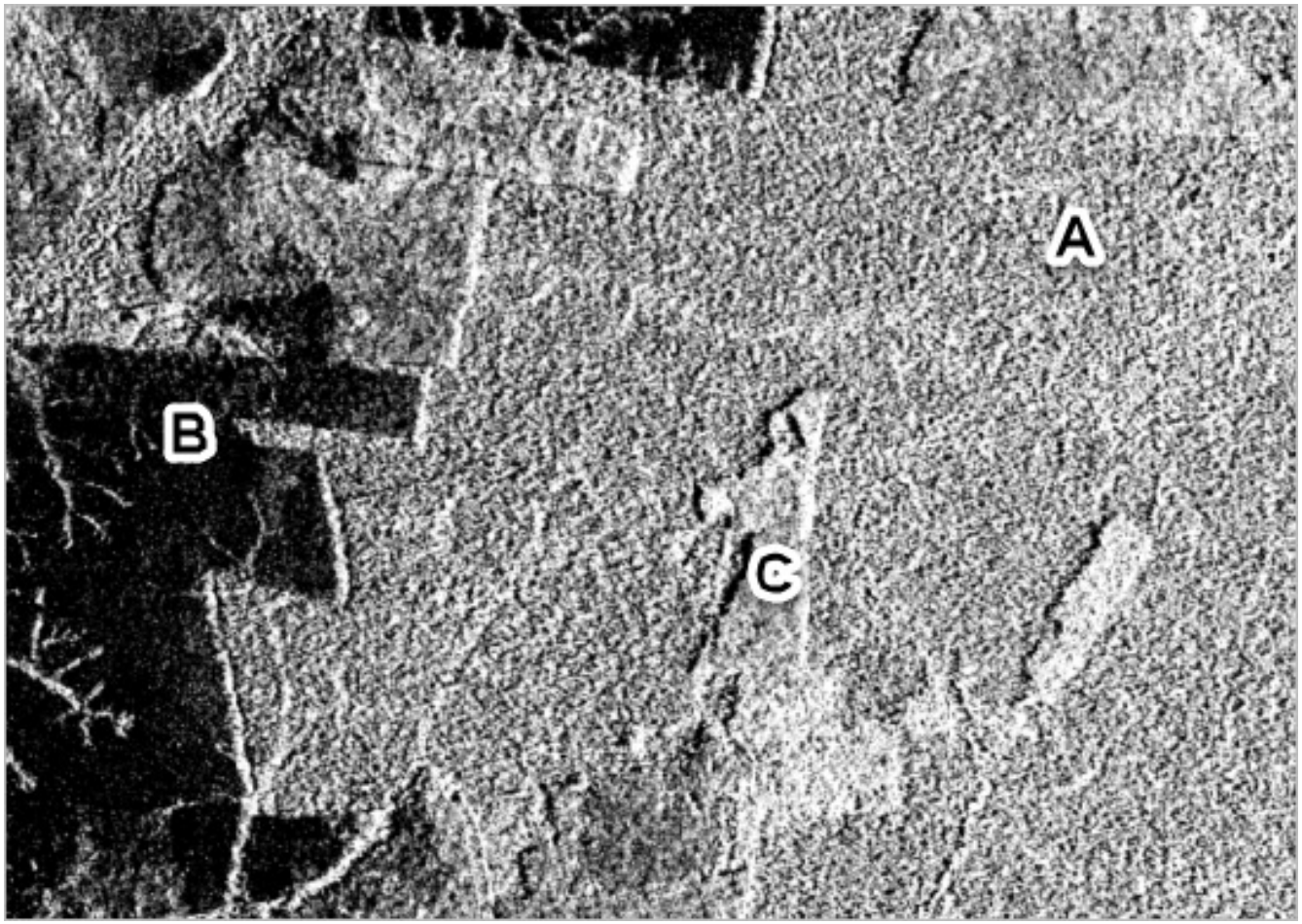
Xie, X.; Li, D.; Liang, X.; Chen, Q.; Yin, K.; Miao, F.
Cross-Regional Detection and Precise GIS Localization of Old Landslides Using High-Resolution Remote Sensing Imagery and YOLOv5. *Remote Sens.* **2026**, *18*, 13.
https://doi.org/10.3390/rs18010013

Limited Expert Annotations



Cheng, Q., Cheng, H., Yuan, L. *et al*. Training strategies for semi-supervised remote sensing image captioning. *Sci Rep* **15**, 25254 (2025). https://doi.org/10.1038/s41598-025-09853-8

Limited Expert Annotations

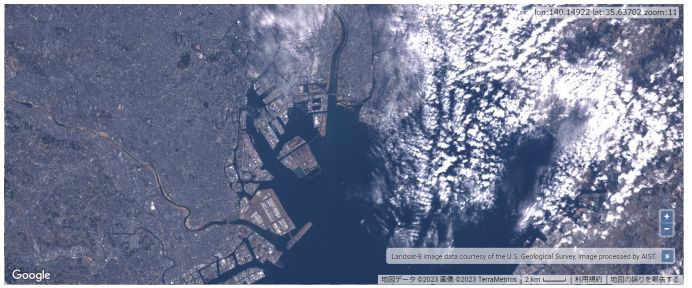


https://learn.arcgis.com/en/projects/explore-sar-satellite-imagery/

Limited Expert Annotations

Issue		Research Strategy
Few expert labels	Limits scale & generalization	Semi-supervised learning, self-supervised learning
Annotation cost	Expensive, slow	Active learning
Sparse labels	Coarse or minimal labels only	Weak/few-shot supervision
Unlabeled data abundant	Hard to use effectively	SSL, self-supervision
Domain mismatch	Labels don’t transfer directly	Domain adaptation

Unique Satellite Data Archives in AIST



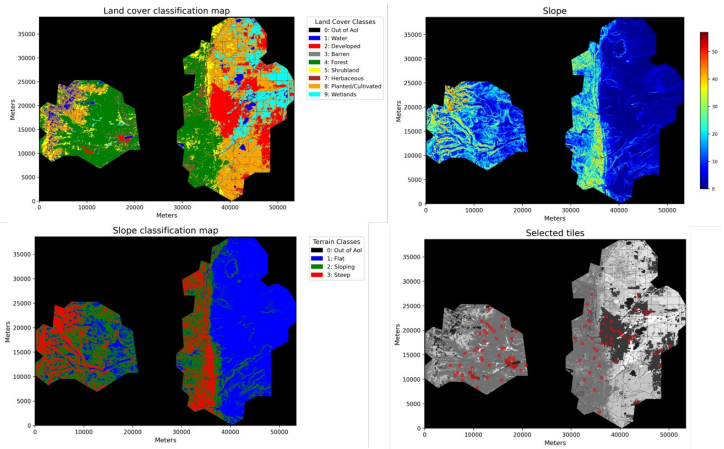
PB-scale satellite image archiving

- **Optical： ~1 PB, ~1TB/day**
  Terra/ASTER (since 1999): 100 TB
  Landsat8 / 9 (since 2013): 500 TB
  Sentinel 2A/B: 200 TB (Japanese territory)

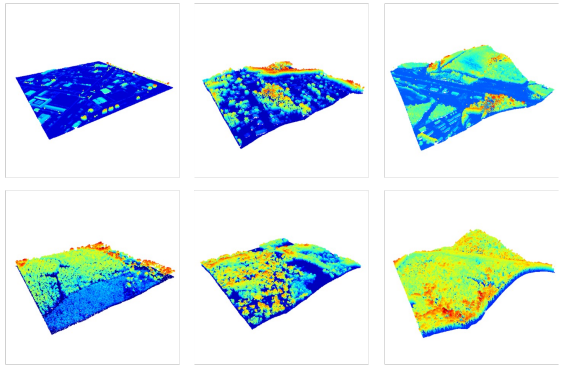
- **SAR : ~5 PB**
  ALOS/PALSAR (2005 ~ 2011): 4 PB
  L1 (raw level data), L2 (Image data)
  ALOS2/PALSAR2 (JAXA-AIST collaboration) (since 2018): 1.2 PB + α
  (Japan＋Tropical region, Amazon, Africa, etc…)

Strategy for Point Cloud Dataset

- Better point cloud dataset to **balance land cover and terrain type distribution** by considering the **class balance.**
- Develop a pre-trained model by training a masked autoencoder (MAE) model on the developed dataset.



Geospatial sampling based on land cover and terrain (slope) classification maps



Examples of sampled point cloud tiles with varying land covers and terrain types

Dataset	Year	Coverage	#Points
ISPRS [38]	2014	–	1.2 M
DublinCity [39]	2019	$2 \times 10^6\,\mathrm{m}^2$	260 M
LASDU [40]	2020	$1.02 \times 10^6\,\mathrm{m}^2$	3.12 M
DALES [3]	2020	$10 \times 10^6\,\mathrm{m}^2$	505 M
ECLAIR [77]	2024	$10.3 \times 10^6\,\mathrm{m}^2$	582 M
IDTReeS [78]	2021	$3440\,\mathrm{m}^2$	0.02 M
PureForest [2]	2024	$339 \times 10^6\,\mathrm{m}^2$	15 B
ISPRS filtertest [79]	–	$1.1 \times 10^6\,\mathrm{m}^2$	0.4 M
OpenGF [1]	2021	$47.7 \times 10^6\,\mathrm{m}^2$	542.1 M
3DEP (Ours)	-	$17691 \times 10^6\,\mathrm{m}^2$	184 B

comparison of the dataset statistics

Lack of Real-world Impacts

- Barriers for deployments that materially improve outcomes, workflows, or decisions for users outside research lab

ML/DL algorithms

Labs

Science, technology, engineering, and mathematics (STEM)

academic benchmark sets

Deployment challenges

Difficulty translating research prototypes into operational systems that improve real-world workflows or decision-making.

High stakes and complexity

Real-world geospatial problems involve large scale, significant risk, and diverse stakeholders.

Limited cross-sector collaboration:

Sustainable partnerships between academia, government, and industry are still insufficient to address complex societal challenges.

AI-based products & services in the real world

[SOURCE]OpenAI. (2024). ChatGPT (4o)

Lack of Real-world Impacts

- AIST Efforts

ML/DL algorithms

Labs

Science, technology, engineering, and mathematics (STEM)

academic benchmark sets

Deployment challenges → **AI Bridging Cloud Infrastructure (ABCI)**

High stakes and complexity → **AI Safety Project**

Limited cross-sector collaboration: → **AI Quality Management Initiative**

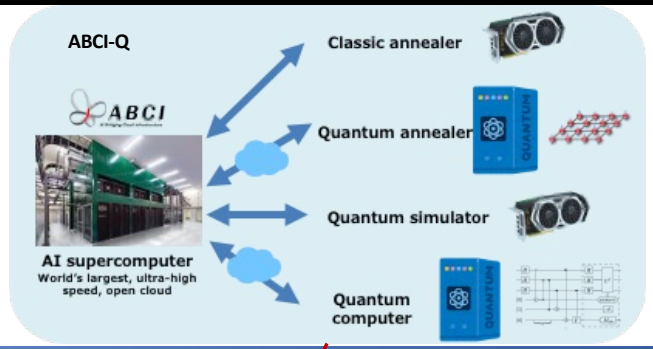
AI-based products & services in the real world

[SOURCE]OpenAI. (2024). ChatGPT (4o)

ABCI Next Generation (https://abci.ai/)

AIST
Create the Future, Collaborate Together

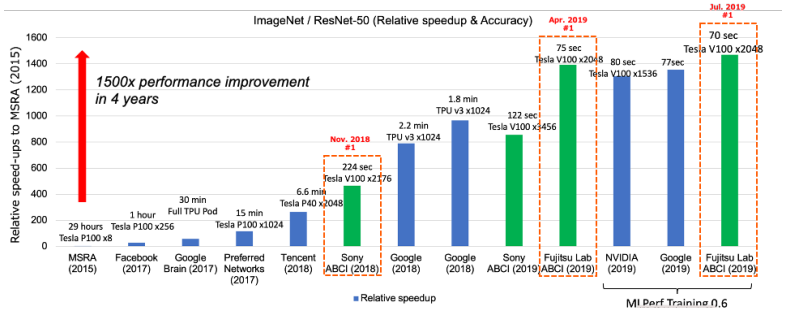
ABCI (AI Bridging Cloud Infrastructure) continues to evolve to meet industrial demands and emerging technology trends such as generative AI and quantum computing.



1208 servers（5312 GPU）:
0.85Exa AI-FLOPS

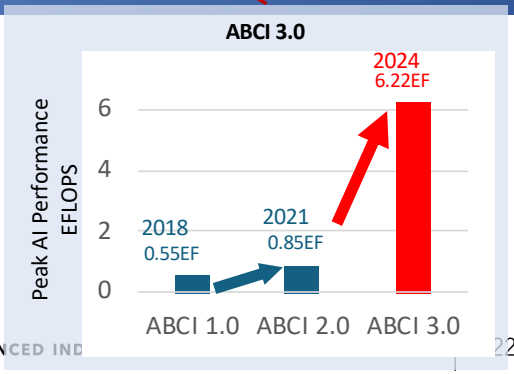
2011 D-Wave
2012 AlexNet
2015 AlphaGo
2018 ABCI
2019 Google Quantum Supremacy
2021 ABCI2.0
2023 QuEra 48 Logical Qubits
2023 ChatGPT
2024 ABCI3.0
2025 ABCI-Q

World records in "MLPerf training" benchmark



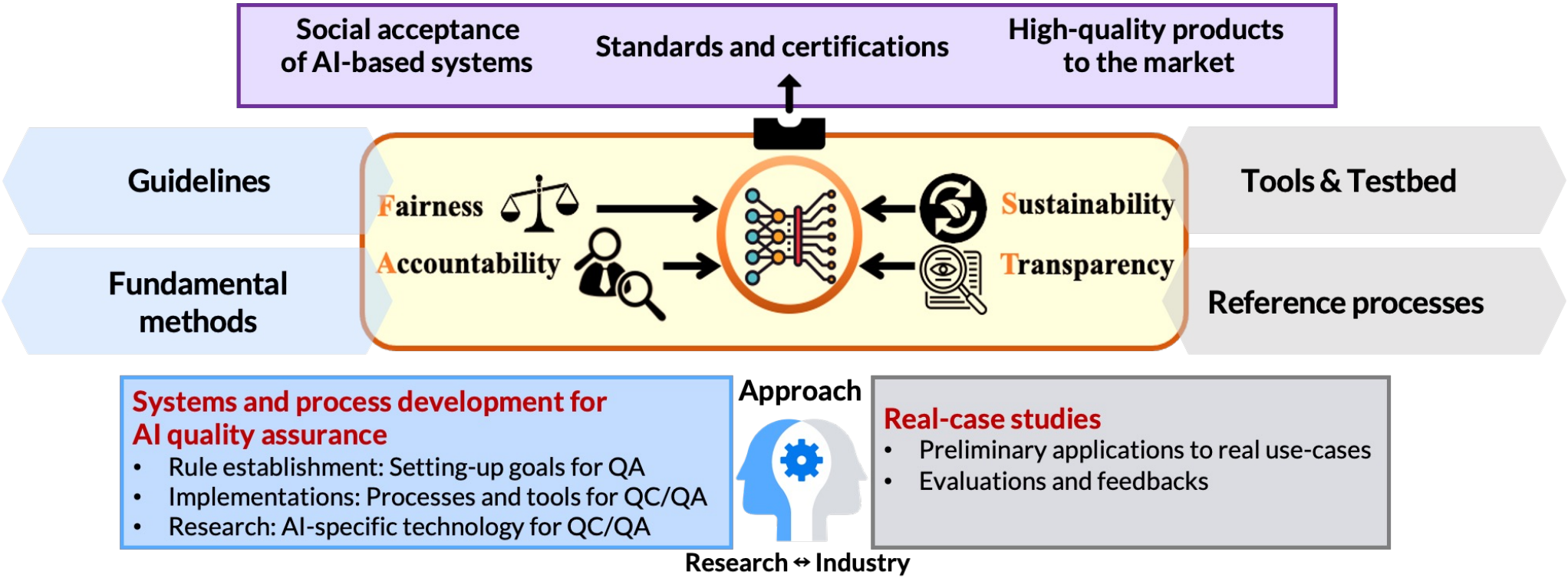
LLM Building Support Program in 2023

Preferred Networks
NII 国立情報学研究所 National Institute of Informatics
産総研
東京工業大学 Tokyo Institute of Technology
ELYZA
LLM 勉強会 LLM-jp



AI Safety Project

- Research and development of holistic approaches to analyze risks and to evaluate the effectiveness and reliability of AI and establish guidelines, methods, and toolsets



AI Quality Management Initiative

AIQMI

Industry members help each other to promote AI Quality management

70⁺ Menbers from 40⁺ Companies

Chairman: Yoshiki Seo (AIST)
Secretary General: Koichi Konishi (AIST)

To be Updated

Regulations
Guidelines
Technologies
Authentications
…

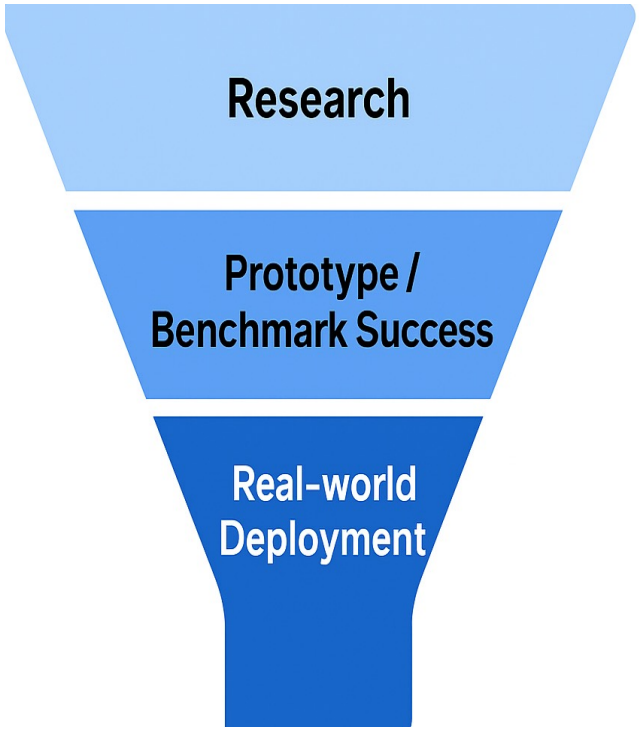
Problem Solving

Biz Requirement
Evaluation
Testing
Authentications
…

Ecosystem

Cooperation Framework
Authentication
Education

Standardized, Inclusive Benchmarks

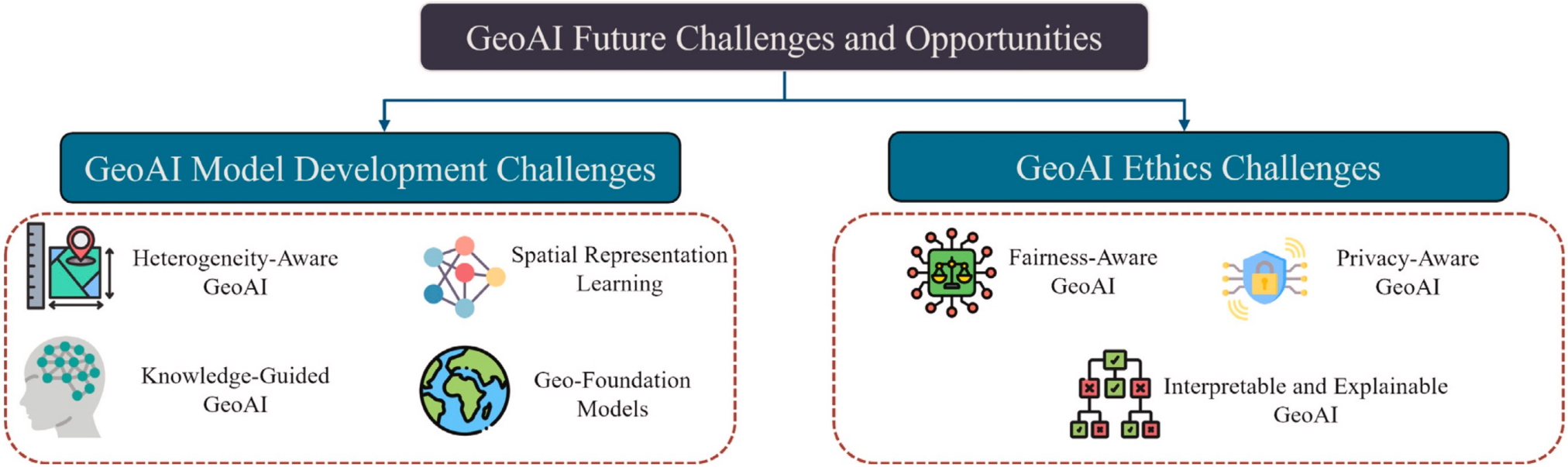


- Remote Sensing–Centric Benchmarks
  - GEO-Bench (Lacoste et al., 2023): A large benchmark for evaluating remote sensing vision foundation models.
  - PANGAEA (Marsocci et al., 2024): A comprehensive remote sensing benchmark emphasizing generalization.
  - VRSBench (Li et al., 2024b): A benchmark for vision–language remote sensing models.
- Vision–Location and Geographic Bias Benchmarks
  - TorchSpatial (Wu et al., 2024): Evaluates vision–location GeoFMs and quantifies geographic bias.
- Map and POI Question Answering Benchmarks
  - MapQA (Chang et al., 2022): A benchmark for understanding maps through QA.
  - MapEval (Dihan et al., 2024): A map-based QA dataset for vision–language GeoFMs.
  - POI-QA (Han et al., 2025): A question-answering dataset focusing on point-of-interest information.
  - GeoBenchX (Krechetova, 2025):: A recently proposed benchmark for evaluating LLM-based geographic inference and map understanding.
- Spatial Cognition Evaluation for LLMs
  - SpatialBenchmark (Xu et al., 2025): Designed to test the spatial reasoning and cognition abilities of large language models.
- Climate and Multimodal Grid-Based Benchmark
  - GeoGrid-Bench (Jiang et al., 2025): A multimodal grid-based geospatial benchmark for evaluating climate-oriented vision–language FM performance.

Standardized, Inclusive Benchmarks

- Spatiotemporal change prediction is not yet systematically evaluated.
- Multimodal fusion tasks (e.g., imagery + time series + text) lack standardized benchmarks.
- Spatial fairness and uncertainty quantification are largely absent from current evaluation frameworks.

Summary



[SOURCE] Mai, G., Xie, Y., Jia, X., Lao, N., Rao, J., Zhu, ,Q., Liu, Z., Chiang, Y.Y., Jiao, J. (2025) Towards the next generation of Geospatial Artificial Intelligence, International Journal of Applied Earth Observation and Geoinformation, vol. 136, 104368.

Thank you for your attention!