SAI4EO: Generative AI for Earth Observation Tasks

Nicolò Taggio[1,*,†], Elio Musacchio[2], Pierpaolo Basile[2], Flavio D'Ippolito[1,†], Domenico Monaco[1,†], Nicola Nicastro[1,†], Marco Polignano[2], Lucia Siciliani[2] and Giovanni Semeraro[2]

[1]*Planetek Italia, via Massaua, Bari, 70123, Italy*
[2]*University of Bari Aldo Moro - Department of Computer Science, Via Orabona 4, Bari, 70125, Italy*

Abstract

This work explores the integration of Symbiotic Artificial Intelligence (SAI) and Earth Observation (EO) through the development of an assistant system capable of analyzing geospatial data via natural language. Starting from preliminary experiments, two use cases were defined to evaluate the approach: (1) container analysis in port areas using Very High Resolution (VHR) satellite imagery, and (2) environmental monitoring of European coastal zones using a large-scale geodatabase. In both scenarios, the system enables users to interact with complex EO data for tasks such as object detection, counting, and trend analysis through intuitive language queries. Early results demonstrate the system's capability to support effective, human-like exploration and interpretation of geospatial information.

Keywords

Symbiotic AI, Earth Observation, Artificial Intelligence, Smart Cities

1. Introduction

Artificial Intelligence (AI) is a branch of computer science dedicated to the development of systems and algorithms capable of simulating human cognitive functions such as reasoning, learning, problem-solving, and adapting to dynamic environments. Recent advancements in computational power have led to rapid progress in the field. Among the emerging paradigms within AI, Symbiotic Artificial Intelligence (SAI) is gaining increasing attention [1]. SAI focuses on fostering synergy between humans and intelligent systems, not to replace human abilities but to augment them through intelligent, adaptive, and collaborative interfaces.

At its core, SAI aims to create a bidirectional relationship between humans and machines, enabling mutual support in decision-making processes and the pursuit of shared goals. This vision is realized through methodologies such as collaborative learning, emotionally responsive interfaces, and models that adapt to user intent and context. In this landscape, Large Language

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
†These authors contributed equally.
taggio@planetek.it (N. Taggio); elio.musacchio@phd.unipi.it (E. Musacchio); pierpaolo.basile@uniba.it (P. Basile); dippolito@planetek.it (F. D'Ippolito); monaco@planetek.it (D. Monaco); nicastro@planetek.it (N. Nicastro); marco.polignano@uniba.it (M. Polignano); lucia.siciliani@uniba.it (L. Siciliani); giovanni.semeraro@uniba.it (G. Semeraro)
0009-0003-6392-3099 (N. Taggio); 0009-0006-9670-9998 (E. Musacchio); 0000-0002-0545-1105 (P. Basile); 0000-0002-3939-0136 (M. Polignano); 0000-0002-1438-280X (L. Siciliani); 0000-0001-6883-1853 (G. Semeraro)

Models (LLMs) [2] have emerged as foundational tools for enhancing human-machine interaction. Trained on vast textual datasets, models like GPT-3 can generate context-aware and coherent text, enabling natural and intuitive dialogue with users. Within the SAI framework, LLMs serve as powerful enablers of interactive, explainable, and accessible AI systems.

In this project, we investigate the potential of SAI in the context of Earth Observation (EO), building on the foundational technologies developed within the FAIR Foundation's Spoke6 project—such as the LLaMAntino [3, 4] and LLaVA-NDiNO [5] models, we adapt and extend these systems to develop an EO assistant capable of interacting with geospatial data through natural language.

To ground our research in practical scenarios, starting from the initial activities [6], we defined and explored two real-world use cases.

The main contributions of this work are: (i) the integration of SAI principles into the EO domain through a natural language-based assistant; (ii) the adaptation of MLLMs to multimodal geospatial tasks; and (iii) the evaluation of this approach in two representative case studies, showing the feasibility and effectiveness of natural language interfaces for EO data exploration and analysis.

2. State of the Art

2.1. Generative AI Methods

While deep learning (DL) has demonstrated promising results in Earth Observation (EO), particularly in tasks like object detection [7], current methodologies often rely on specialized architectures. These task-specific approaches, though effective, struggle with multi-sensor remote sensing data complexities, adapting to diverse tasks, and generalizing across scenarios. On the other hand, the rise of MLLMs has drawn attention to natural language-driven image querying, showcasing impressive multi-task reasoning capabilities in real-world settings. Unlike specialized models, MLLMs offer versatile performance and effectively generalize to new situations, enabling zero-shot capabilities across varied tasks. Key architectures in MLLMs include ViT and LLMs. LLMs continue to push the boundaries of natural language understanding, with the LLaMA [8] (Large Language Model Meta AI) family demonstrating exceptional contextual reasoning abilities. Open-source models like LLaMAntino have been adapted for Italians through linguistic fine-tuning, ensuring data privacy while expanding AI accessibility.

In this project, we will utilize and fine-tune two already pre-trained algorithms as part of the activities conducted within Spoke 6: the LLaMAntino and LLaVA-NDiNO models. These models will be further adapted and optimized using Earth Observation data, enabling them to extract meaningful insights and enhance their performance in processing and analyzing remote sensing information. The fine-tuning process will involve tailoring the models to specific geospatial applications.

2.2. Generative AI in EO

These models can interpret human requests with more context awareness, understand textual inputs, and associate them with images, enabling smoother and more intuitive interactions.

However, applying MLLMs to EO data presents significant challenges due to fundamental differences between natural (standard photo) images and those captured by remote sensing sensors. Consequently, there is a notable gap in the literature regarding MLLM applications for EO, which highlights a crucial area for future research and innovation. Among the few existing studies, notable examples include GeoChat [9], EarthGPT [10], and chat2geo.

GeoChat is the first Large Vision-Language Model designed explicitly for EO, employing a geospatially anchored approach.

EarthGPT is a universal multimodal Large Language Model tailored for EO. It introduces a practical cross-modal learning framework between visual and linguistic modalities, seamlessly unifying a wide range of remote sensing image analysis and multi-sensor interpretation tasks into multi-turn dialogues.

Chat2Geo is a Next.js 15 application offering a chatbot interface for geospatial analysis based on remote sensing. It utilizes the backend's Google Earth Engine (GEE) to process and analyze EO datasets in real time.

This project introduces the first Italian generative model for Earth Observation (EO), specifically designed to address the emerging need for environmental monitoring of European coastal zones. By fine-tuning the open-source LLaVA-NDiNO model, we focus on delivering a solution tailored to real-world user needs. At the heart of this initiative is a strong collaboration between a research center and a private company, demonstrating how joint efforts can create innovative, user-centric tools that bring the power of AI closer to those who need it most.

3. Use cases definition

3.1. Containers analysis

The design of this use case acknowledges that the nature and specificity of user queries can vary significantly depending on the user's expertise and familiarity with the domain. Users with limited technical knowledge may formulate vague or general questions, while experienced analysts may issue highly precise queries targeting specific outputs. To accommodate this variability, we adopt a conversational querying model that enables dynamic interaction: the system responds with progressively richer and more detailed information as the user's questions become more specific.

To enhance response consistency and ensure relevance across diverse query types, user questions have been categorized into thematic groups, each linked to a predefined type of response. This classification facilitates the training and evaluation of the assistant's behavior and allows for modular expansion of the system.

Furthermore, a comprehensive set of question-answer pairs has been prepared and enriched using external large language models. These pairs span a variety of informational needs, including qualitative descriptions, statistical summaries, positional queries, and change detection scenarios. This augmentation process ensures broader coverage of real-world use cases and improves the system's robustness when deployed in operational environments.

3.2. Europe coastal area monitoring

The second use case focuses on analyzing land use patterns and changes along European coastlines, leveraging a comprehensive geospatial database that captures temporal and thematic variations. This use case aims to assist users in interpreting complex environmental data through an intuitive natural language interface.

The proposed workflow involves the ingestion of a polygonal file—such as a GeoJSON, Shapefile, or similar format—defining the Area of Interest (AoI). Once the AoI is identified, the system performs automated statistical analyses over the corresponding spatial extent. If the user provides only the AoI without a specific query, the assistant can return a general overview of land use dynamics in that region. More specific and targeted questions will yield correspondingly detailed and structured responses.

As with the container analysis case, we developed a structured set of question-and-answer pairs designed to cover a wide spectrum of user intents and expected outputs. In this context, however, the responses are more data-rich and may include visual elements such as bar charts, summary tables, time series plots, and comparative diagrams. These multimodal outputs are tailored to support both expert users seeking precise metrics and non-experts exploring general trends or patterns. This design enhances interpretability and encourages broader adoption of EO tools in environmental planning and coastal zone management.

4. Preliminary results

Once defined the two use cases, a fine tuning is performed for the first use case. In particular, we formatted the xView dataset for visual instruction-tuning [11]. To do so, we convert remote sensing tasks (e.g. object detection) into a dataset consisting of conversational turns, where the user sends a request written in natural language and the model generates the appropriate answer. To leverage outputs from external models and algorithms, we want the model to generate properly formatted API requests, outputs of external tools for these requests are added to the conversation with a special role. An example instance from this training mixture is shown in Figure 1.

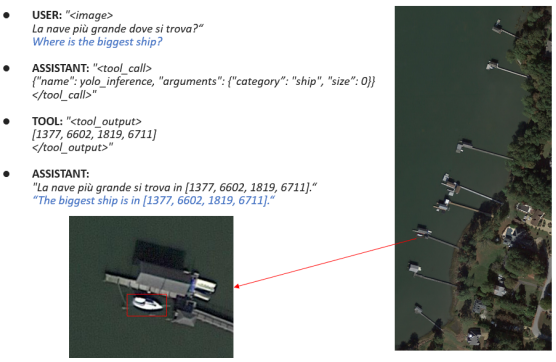

Figure 1: An example of instance from the xView dataset after formatting.

To create this mixture, we implemented a pipeline to automatically convert any remote sensing dataset into this format. The pipeline extracts relevant data from the dataset and applies a formatting for a randomly chosen tasks (e.g. object detection). Pre-defined textual formatting for the user requests and model answers are then applied to the extracted data. Annotation of textual formatting for user requests and model answers in Italian was manually performed by the authors of this work. After applying the pipeline to the xView dataset, the mixture consists of a total of 15,156 distinct instances. We perform fine-tuning of the Qwen 2.5 VL 3B model on this dataset. We choose this model since it supports the Italian language and has a very low amount of parameters. For fine-tuning we used LoRA with a rank of 16, alpha of 32, learning rate of 2e-5 and batch size of 16. We train for 1 epoch over the entire training dataset. At the time of writing, we only test the performance of the model in generating the correct external tool call. This is important since it is also required that the model generates appropriate arguments of the external tool call w.r.t. user requests, for example if the user requests the bounding box of the largest ship in the image the arguments of the tool call should contain a label for that specific entity of interest. We apply the same pipeline to the xView dataset to create a small test set consisting of 500 instances. We test model performance in generation of the correct tool call, using greedy decoding as sampling strategy, and report an outstanding 98% accuracy.

5. Conclusion

In the project's initial phase, two use cases will be defined in collaboration with EO experts. These use cases will be designed to align with the algorithms already developed within Spoke 6. A large language model, such as LLaVA-NDiNO, has been selected for the defined tasks. The use cases will address different aspects of EO, including object detection and the analysis of large-scale datasets—areas that often present challenges for typical end-users who are not EO specialists.

In the second phase, the defined use cases will be the foundation for prototyping an EO chatbot. Specifically, EO data will be prepared, and the LLaVA-NDiNO model will be fine-tuned based on a structured question-and-answer approach derived from the use cases. Given the computational demands of the fine-tuning process, resources from the Leonardo supercomputer at CINECA will be utilized, with an estimated 40,000 computing hours allocated for this task.

Finally, a validation phase will be conducted in collaboration with EO experts. For the first use case, accuracy will be assessed by comparing the model's outputs with ground-truth bounding boxes. In the second use case, a geodatabase will be developed to enable natural language queries, allowing users to extract relevant information. This system's performance will be evaluated qualitatively and quantitatively by EO experts using a prototype front-end developed during the project's second phase.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the

NextGenerationEU.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4.5 and Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] P. Marra, L. Pulito, A. Carnevale, A. Lombardi, F. Lisi, A. Dyoub, et al., Acceptability of symbiotic artificial intelligence: Highlights from the fair project, in: CEUR WORKSHOP PROCEEDINGS, volume 3762, 2024, pp. 112–117.

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 1 (2023).

[3] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, LLaMAntino: LLaMA 2 models for eective text generation in Italian language, arXiv preprint arXiv:2312.09993 (2023).

[4] M. Polignano, P. Basile, G. Semeraro, Advanced Natural-based interaction for the ITAlian language: LLaMAntino-3-ANITA, arXiv preprint arXiv:2405.07101 (2024).

[5] E. Musacchio, L. Siciliani, P. Basile, G. Semeraro, LLaVA-NDiNO: Empowering LLMs with Multimodality for the Italian Language, in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2024), CEUR-WS. org, 2024.

[6] N. Taggio, S. Samarelli, M. Simone, Sai4eo: Symbiotic artificial intelligence for earth observation, in: CEUR WORKSHOP PROCEEDINGS, volume 3762, 2024, pp. 24–28.

[7] T. Hoeser, C. Kuenzer, Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends, Remote Sensing 12 (2020) 1667.

[8] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[9] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, F. S. Khan, Geochat: Grounded large vision-language model for remote sensing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 27831–27840.

[10] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, X. Mao, Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain, IEEE Transactions on Geoscience and Remote Sensing (2024).

[11] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26296–26306.