ANNALS OF GEOPHYSICS, 2026; doi:10.4401/ag-9685
OPEN ACCESS

Exploring the Adoption of Generative AI in Geoscientific Workflows at INGV: A Structured Qualitative Case Study from the GdL AI Working Group

Daniele Bailo[1], Raffaele Di Stefano[1], Michele Matteo Cosimo Carafa[1], Fabio Di Felice[1], Valentino Lauciani[1], Valerio De Rubeis[1], Stefano Aretusini[*,1], Gabriele Tarabusi[1], Luca Malagnini[1], Marco Salvi[1]

(1) Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy

Article history: received August 14, 2026; accepted November 4, 2026

Abstract

Generative Artificial Intelligence is rapidly entering scientific practice and poses both an opportunity and a challenge for research institutions. Alongside the rapid diffusion of new tools, researchers are increasingly required to understand where these systems can provide real support, where their limits remain substantial, and under which conditions they can be used responsibly. Within this context, in May 2024 the Istituto Nazionale di Geofisica e Vulcanologia (INGV) established a working group on Generative AI (GdL AI) to explore its possible role in geoscientific workflows. This paper presents a structured account of that experience which lasted one year. The initiative examined a set of representative use cases, including macroseismic interpretation, extraction of information from historical earthquake documents, natural-language interaction with seismic web services, and AI-assisted scientific coding for seismological processing. The results suggest that Generative AI can already provide useful support in selected tasks, especially when it helps transform unstructured material into structured intermediate outputs, accelerates access to technical infrastructures, or shortens development and code-revision cycles in scientific software. At the same time, the experience consistently showed that these systems do not remove the need for expert judgment. Their outputs remained scientifically useful only when embedded in a human-in-the-loop process of validation, correction, and contextual interpretation. More broadly, the INGV case indicates that institutional adoption depends not only on access to tools, but also on training, privacy-aware governance, and shared practice. These findings support a cautious but constructive conclusion: Generative AI can strengthen selected geoscientific workflows when its use is documented, critically supervised, and integrated into a broader institutional framework of responsible experimentation.

Keywords: Generative AI; Large Language Models; Geoscientific Workflows; Qualitative Case Study; Human-In-The-Loop

1. Introduction

Generative Artificial Intelligence (Generative AI), and in particular chatbot systems based on Large Language Models (LLMs) such as ChatGPT, Gemini, and Claude[1], is increasingly entering scientific work not only as a writing aid but also as a practical interface for coding, data access, document analysis, and exploratory reasoning (Gemini Team et al., 2023; OpenAI et al., 2023). In geoscience, this transition is especially relevant because many research and operational workflows combine complex observational data, simulations, inverse problems, computational pipelines, and expert interpretation (Bergen et al., 2019), while recent reviews have explicitly identified LLMs and generative AI as an emerging research frontier in geoscience and geospatial science (Wang et al., 2024). As a result, even modest improvements in tasks such as extracting information from historical documents, interacting with web services, or accelerating scientific coding can have a disproportionate effect on research productivity.

At the same time, the scientific value of these systems cannot be inferred from the generic enthusiasm for new technologies. Their usefulness depends on context, task definition, validation procedures, and institutional conditions of use. For research organizations, the question is therefore not simply whether Generative AI is powerful, but under which conditions it becomes useful, trustworthy, and governable inside real scientific workflows.

Recent literature has begun to document institutional responses to the adoption of Generative AI in research and higher-education settings. Liu and Jagadish (Liu and Jagadish, 2024) argue that research organizations should not leave the adoption of rapidly evolving AI technologies entirely to individual researchers, but should provide shared support through training, tools, and responsible practices. More focused institutional initiatives include the private Generative AI proof of concept developed at Emory University to support research-administration activities (Wilson et al., 2025), and the AI Task Team process at the University of KwaZulu-Natal, which translated institutional deliberation into governance guidelines, data-protection provisions, and capacity-building measures across teaching and research (Thaldar et al., 2025). Survey evidence further indicates that Generative AI use among academic scientists is already widespread, although perceptions of benefits, risks, and intended uses remain heterogeneous (Arroyo-Machado et al., 2025). Compared with these experiences, the INGV initiative focused on an interdisciplinary working group that explored multiple geoscientific use cases through iterative experimentation, collective discussion, and expert validation.

Against this wider institutional landscape, the Istituto Nazionale di Geofisica e Vulcanologia (INGV) established a working group on Generative AI (GdL AI) with the dual objective of promoting informed experimentation and building shared competence across disciplinary boundaries. Unlike the institution-wide support, focused research-administration deployment, and governance-oriented processes described above, the INGV initiative focused on an interdisciplinary working group that explored multiple geoscientific use cases through iterative experimentation, collective discussion, and expert validation. Between May 2024 and May 2025, the initiative combined practical experimentation, recurring meetings, thematic discussions, training activities, and reflection on privacy, ethics, and organizational adoption to evaluate Generative AI-based workflows in a public geoscience research setting. Established a working group on Generative AI (GdL AI) with the dual objective of promoting informed experimentation and building shared competence across disciplinary boundaries. Between May 2024 and May 2025, the initiative combined practical experimentation, recurring meetings, thematic discussions, training activities, and reflection on privacy, ethics, and organizational adoption in order to evaluate-based workflows in a public geoscience research setting.

This paper adopts a structured qualitative case-study perspective on early Generative AI adoption in geoscientific workflows. The main contribution is threefold. First, it documents a set of concrete use cases that are directly relevant to geoscience, including macroseismic interpretation, historical earthquake document analysis, interaction with seismic web services, and scientific coding for seismological processing. Second, it identifies recurring patterns of utility and limitation, with particular emphasis on human-in-the-loop validation. Third, it shows that institutional adoption is not only a technical issue but also a matter of training, governance, and shared practice.

The paper therefore focuses on a specific question: how did the GdL AI initiative at INGV use Generative AI in selected workflows, what kinds of benefits and limitations emerged, and what lessons can be generalized for similar research environments? Within this scope, the paper documents an early phase of technically exploratory and institutionally organized experimentation with Generative AI at INGV.

1 For Claude, see the official Anthropic announcement of the Claude 3 model family: https://www.anthropic.com/news/claude-3-family (last accessed July 2026).

2. Background and Working-Group Organization

The GdL AI initiative was established at INGV as a structured response to the rapid diffusion of Generative AI tools in scientific and technical work. In its earliest phase, the initiative was connected to a practical institutional need: understanding whether protected access to Generative AI systems could support research and operational activities while also reducing risks linked to informal use of external AI services. The group was therefore conceived not simply as a forum for discussion, but as a working environment in which experimentation could be organized, shared, and critically assessed.

The working-group process combined recurring meetings, practical experimentation, and progressive consolidation of use cases. Early meetings had a strong alignment function: they created a shared baseline on what Generative AI is, how it differs from other forms of artificial intelligence, and where it might realistically be useful within our institute. Once access to licensed tools became available, the emphasis shifted toward concrete tasks drawn from participants' own scientific and technical practice. Members were encouraged to test one or more use cases, prepare short presentations or demonstrations, and return to the group with results, difficulties, and open questions.

This organization gave the initiative a recognizable internal structure. Meetings were not limited to generic discussion, but were used to review progress, compare experiences, assign follow-up tasks, and identify themes requiring deeper treatment. Between meetings, participants developed examples, refined prompts, tested workflows, and prepared materials that could be discussed collectively. In this sense, the group operated through cycles of experimentation, reporting, feedback, and renewed testing.

A second important feature of the process was periodic thematic deepening. Rather than treating Generative AI as a single undifferentiated topic, the group progressively organized attention around specific domains such as coding, macroseismics, workflow automation, historical earthquake documents, ethics, and training. This helped move the initiative from generic curiosity to task-oriented reflection. It also made visible that different use cases required different forms of validation and different levels of domain expertise.

The initiative evolved in parallel with a broader institutional conversation on training, privacy, and governance. Meeting discussions show that the group was aware from an early stage that AI adoption could not be reduced to tool acquisition alone. Questions of safe use, staff training, hardware needs, and future project development were explicitly tied to the experimentation process. The working group thus became one of the first places within the institute where technical, scientific, and organizational aspects of Generative AI were discussed together.

This trajectory can be described in three broad phases: an initial exploratory phase centered on proposal and curiosity; a consolidation phase in which use cases were discussed in relation to policy, privacy, and practical constraints; and a subsequent phase of thematic deepening and institutional dissemination. This phased evolution is summarized in Fig. 1. It is important because it shows that the initiative was not only about testing outputs, but also about building a structured environment for responsible experimentation.

Figure 1 synthesizes the main meetings that marked this evolution. The May 2024 meeting surfaced the first concrete proposals; the October 2024 meeting connected experimentation with privacy, licensed access, and a possible paper; the January 2025 meeting aligned the emerging cases with writing and training goals; and the February 2025



Figure 1. Timeline of the main milestones of the INGV GdL AI initiative: 16 May 2024, 23 October 2024, 14 January 2025, 19 February 2025, and the institutional-uptake phase of April-May 2025. The sequence highlights the transition from exploratory use-case proposal to policy-aware consolidation, thematic deepening, and broader institutional capability building.

thematic meeting deepened the technical side, especially around coding and workflow automation. The later phase represented in the figure corresponds to broader institutional uptake through training and internal dissemination.

These were the main reference meetings, but they were not the whole process. The initiative also relied on several smaller and more informal exchanges in which participants discussed technical details, compared intermediate results, clarified task requirements, and prepared material for subsequent collective sessions.

3. Methods

3.1 Study design

This work adopts a qualitative case-study design. Its aim is not to measure the comparative performance of LLMs on standardized benchmarks, nor to estimate causal effects on productivity through a formal experimental protocol. Instead, it reconstructs and synthesizes the experience of the INGV GdL AI initiative through a set of representative use cases developed, discussed, and refined within the working group. The unit of analysis is therefore the use case, understood as a task or workflow for which Generative AI was explored as an operational aid. The case was bound to the activities of the INGV GdL AI initiative conducted between May 2024 and May 2025.

3.2 Working method of the GdL AI initiative

The working method of the initiative was iterative and case-based. Participants met periodically, tested concrete tasks between meetings, and reported their experiences back to the group. These experiences were not restricted to final successful outputs: they also included failed attempts, limitations encountered, prompt refinements, and reflections on validation requirements. The recurring meetings therefore functioned both as moments of exchange and as checkpoints in an ongoing experimentation cycle. Progress was collected in minutes, presentations, and working documents, which made it possible to preserve continuity between meetings and to trace how the experimentation developed over time.

A practical feature of the method was the progressive assignment of follow-up activities between meetings. Participants were encouraged to explore one or more use cases in their own domain, prepare examples or short presentations, and return with material that could be discussed collectively. This procedure made the initiative cumulative: each meeting did not restart the discussion from zero, but built on previously shared experiments, new questions, and partial results.

3.3 Use-case selection

For the purposes of this manuscript, the broad set of activities carried out in the group was distilled into a smaller number of representative use cases. Selection favored cases that satisfied three conditions: (i) they addressed a concrete geoscientific or technical task, (ii) they were discussed with sufficient specificity to reveal both utility and limitation, and (iii) they illustrated a recurring pattern rather than an isolated curiosity.

For presentation and comparative analysis, the six representative use cases were organized into four broad domains (Fig. 2): (i) macroseismics and historical earthquake analysis (UC1 and UC2); (ii) data access and workflow automation (UC3); (iii) scientific coding and software maintenance (UC4 and UC5); and (iv) training and institutional uptake (UC6). Out of eleven were retained for the main analytical inventory: macroseismic intensity attribution from historical summaries (UC1), extraction of structured information from historical earthquake documents (UC2), natural-language interaction with seismic web services (UC3), AI-assisted seismological coding pipelines (UC4), support for legacy-code conversion and software maintenance (UC5), and training-oriented institutional uptake (UC6).

3.4 Analytical criteria

The analysis adopted a structured qualitative case-study approach in which all representative use cases were examined through a common analytical matrix. Our framework-informed structured synthesis approach is consistent

with qualitative case-study methodology, where a shared analytical protocol supports systematic cross-case comparison (Yin, 2018). Our approach draws on the principles of Framework Analysis, which advocates the use of predefined analytical dimensions to facilitate comparison across heterogeneous qualitative cases while preserving contextual interpretation (Ritchie and Spencer, 1994). In the present study, the matrix served to organize interpretation, comparison, and synthesis across the institutional use cases discussed within the GdL AI initiative.

The dimensions themselves were defined for the purposes of this work, but in a way consistent with that qualitative comparative logic rather than as imported performance metrics. Each use case was examined according to the same dimensions: (i) scientific domain and relevance; (ii) task type; (iii) role played by Generative AI; (iv) expected or observed output; (v) human validation requirements; (vi) observed benefits; (vii) limitations and risks; and (viii) type and strength of available evidence. These criteria were intended to support structured qualitative interpretation and comparison, not quantitative evaluation, scoring, or performance measurement.

Table 1. Analytical dimensions used to compare the representative use cases.

Analytical dimension	Purpose in the analysis
Scientific domain and relevance	Clarify why the case matters for geoscientific or institutional practice.
Task type	Identify the operational problem being addressed, such as extraction, querying, coding, or training.
Role played by Generative AI	Specify whether the model acted as translator, drafting aid, coding assistant, interface layer, or exploratory support.
Expected or observed output	Describe the form of the resulting artifact, query, script, classification, or organizational outcome.
Human validation requirements	Make explicit what expert checks were required before outputs could be considered usable.
Observed benefits	Capture the main practical value perceived in the case, such as time reduction, easier access, or first-pass structuring.
Limitations and risks	Record recurring weaknesses, including errors, ambiguity, prompt dependence, privacy constraints, or overinterpretation risk.
Type and strength of available evidence	Distinguish between stronger and thinner cases without turning the comparison into numerical scoring.

3.5 Validation and limitations

A central methodological point emerging from the initiative is that validation remained human-led in all scientifically meaningful cases. In coding tasks, validation took the form of code review, execution checks, and scientific plausibility assessment. In document-analysis and macroseismic tasks, validation required expert assessment of extraction completeness, interpretation, and consistency with established domain knowledge. The present paper therefore treats the evidence as qualitative and reconstructive, with claims of utility expressed in bounded terms and linked to the documented working process.

4. Results

4.1 Overview of the use-case set

The six use cases described above emerged progressively during the experimentation and are presented here in the order in which the experience developed. to synthesize the common patterns that emerged across these activities, we subsequently organized them into four broad domains: (i) macroseismics and historical earthquake

analysis; (ii) data access and workflow automation; (iii) scientific coding and software maintenance; and (iv) training and institutional uptake. This domain structure is therefore an inductive summary of the experimental experience, rather than a predefined classification, and is summarized in Fig. 2. Table 2 applies the comparative analytical matrix introduced in Section 3.4 (Table 1) to the six representative use cases.

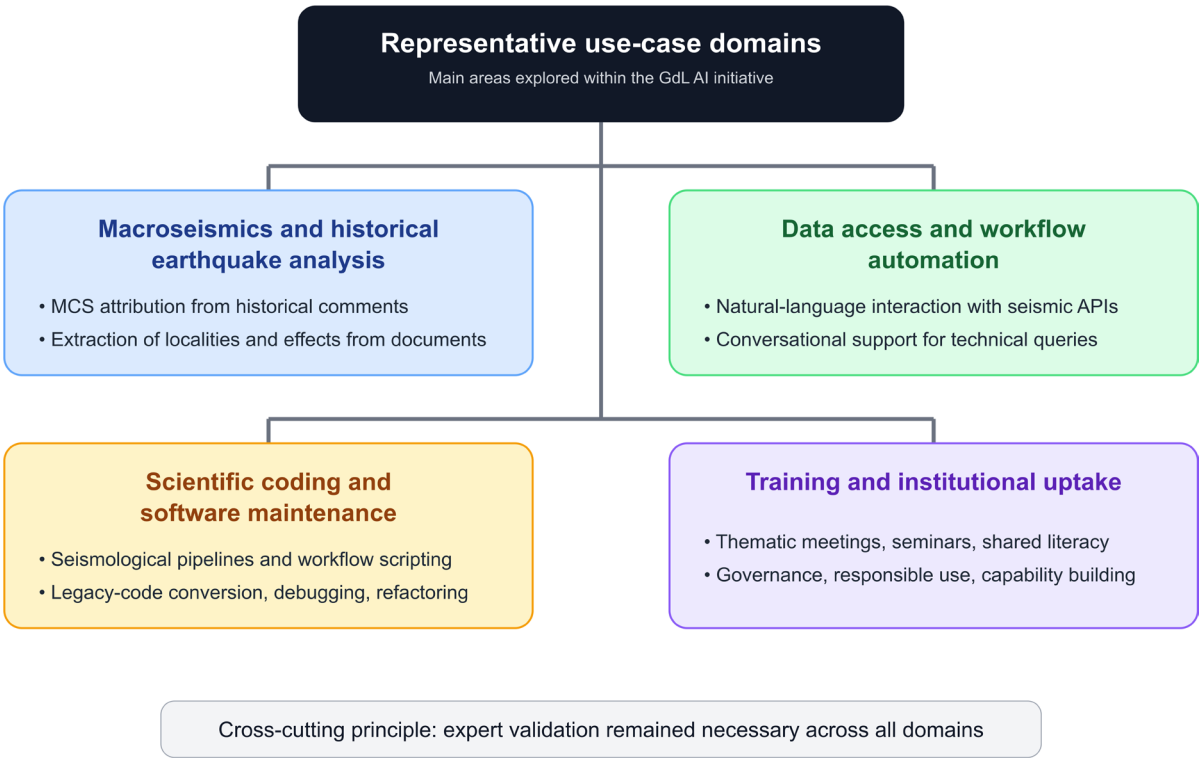


Figure 2. Taxonomy of the representative use cases identified within the GdL AI initiative. The four domains capture the main areas in which Generative AI was explored, while also emphasizing that scientifically meaningful outputs remained dependent on expert validation.

4.2 Macroseismics and historical earthquake analysis

One of the clearest geoscience-specific lines of experimentation concerned historical earthquake information. In the CFTI context (Guidoboni et al., 2019)[2], use case UC1 explored whether an LLM could assist in assigning Mercalli-Cancani-Sieberg (MCS) intensity values from historical-critical comments (Musson et al., 2010). The proposed workflow ranged from a relatively conservative use, in which the model supports quality ranking and consistency checking of existing summaries (as the CFTI's historical-critical comments), to more ambitious scenarios in which the model attributes intensity directly from textual descriptions. This use case is scientifically relevant because it targets a well-defined interpretive problem in historical seismology, already embedded in a curated historical-earthquake framework, while keeping expert oversight explicit.

A related use case, UC2, focused on extracting structured information from documentary sources used in historical earthquake studies. The proposed workflow included the extraction of localities, descriptions, temporal information, and perceived effects from PDF documents, followed by geolocation and support for epicentral inference. This case belongs to the same broader historical-seismology domain, but emphasizes a different step in the workflow: moving from narrative documentary evidence toward structured tables that can support further interpretation. Here too, extraction completeness, location disambiguation, and interpretation of damage descriptions remain dependent on expert review.

2 For the Catalogo dei Forti Terremoti in Italia (CFTI5Med), see the official resource at http://storing.ingv.it/cfti/cfti5/ (last accessed July 2026).

Table 2a. Comparative analytical matrix applied to the representative use cases. Operational focus.

ID	Domain	Task type	Role of GenAI	Expected or observed output
UC1	Macroseismics/ historical earthquakes	Intensity attribution from historical summaries	Interpretive support and first-pass classification	Proposed MCS intensity classes and summary-quality ranking
UC2	Macroseismics/ historical earthquakes	Information extraction from documentary sources	Extraction and structuring of narrative evidence	Tables of localities, effects, timing, and candidate epicentral clues
UC3	Data access/ workflow automation	Natural-language interaction with seismic web services	Conversational interface to APIs and query formulation	Structured service calls, parameterized queries, and reusable query patterns
UC4	Scientific coding/ software maintenance	Event-to-waveform processing pipeline	Coding assistant and workflow integrator	Processing scripts, diagnostics, seismological outputs
UC5	Scientific coding/ software maintenance	Legacy-code conversion and project navigation	Code conversion, explanation, and implementation support	Converted code, code explanations, project maps, and modification suggestions
UC6	Training/ institutional uptake	Literacy building and internal dissemination	Organizational support, shared practice, and capability building	Thematic meetings, training materials, and shared responsible-use practices

Table 2b. Comparative analytical matrix applied to the representative use cases. Evaluation focus.

ID	Human validation requirements	Observed benefits	Limitations and risks	Type and strength of†evidence
UC1	Expert review of assigned intensity, consistency, and historical plausibility	Faster first-pass screening of summaries; support for consistency checking	Risk of overinterpreting ambiguous descriptions; strong dependence on expert judgment	Clearly defined exploratory case presented in the CFTI context
UC2	Check extraction completeness, geolocation, and interpretation of damage descriptions	Transforms unstructured historical documents into more analysable intermediate material	Ambiguity in place names, incomplete extraction, and possible distortion of historical nuance	Clearly described workflow proposal with explicit analytical steps
UC3	Check syntax, parameters, service semantics, and returned data	Easier access to technical infrastructure; reduced friction between question and query	Incorrect parameterization or misleading outputs if service logic is not checked	Recurrent institutional use case with clear practical logic
UC4	Code review, execution checks, and scientific plausibility assessment	Strong acceleration of implementation, revision, and iterative refinement	Requires expert supervision throughout; output quality depends on context and feedback	Strongest technical case, supported by a detailed presentation and reconstructed workflow
UC5	Iterative code review, implementation checks, and execution-based verification	Lowers the entry barrier to unfamiliar languages and codebases	Risk of partial or misleading code transformations without close review	Multiple concrete examples discussed in coding activities
UC6	Organizational coherence, alignment with responsible-use principles, and internal review	Extends value beyond isolated experiments; builds shared capability	Not evaluable as a technical performance case; outcomes are organizational rather than benchmarkable	Strong process evidence across recurring meetings and training activities

4.3 Data access and workflow automation

Use case UC3 concerned the use of Generative AI to support interaction with INGV seismic web services, in continuity with the institute's broader data acquisition and distribution infrastructure (Mazza et al., 2012). In this workflow, natural-language requests were translated into structured calls to web services, with the practical effect of simplifying access to seismic data. This case is strategically important because it shows how conversational interfaces can become an orchestration layer for otherwise complex technical workflows, shortening the path from a scientific question to a technically correct query while leaving expert control over the interpretation of returned data. In this sense, Generative AI acts less as an autonomous analyst than as an interface between scientific intent and machine-readable service calls.

4.4 Scientific coding and software maintenance

In use case UC4, Generative AI was used to help design and refine a pipeline that starts from an INGV seismic event identifier and traverses multiple steps: retrieval of QuakeML event data, extraction of picks and station metadata, download of waveform data through FDSN services, signal-to-noise and frequency analysis, response removal, Wood-Anderson simulation, and export of summary outputs (Fig. 3). The workflow combines institution-specific web services with community standards such as QuakeML (Danijel Schorlemmer et al., 2011) and FDSN-style service access within the INGV operational environment (Mazza et al., 2012). This case is both domain-specific and technically articulated. It shows that the model was not simply asked for generic code snippets, but was embedded in an iterative loop of specification, code revision, refinement, and expert review. Group discussion around this case emphasized that the adopted Generative AI tool worked best when provided with detailed context, explicit constraints, and repeated feedback during implementation.

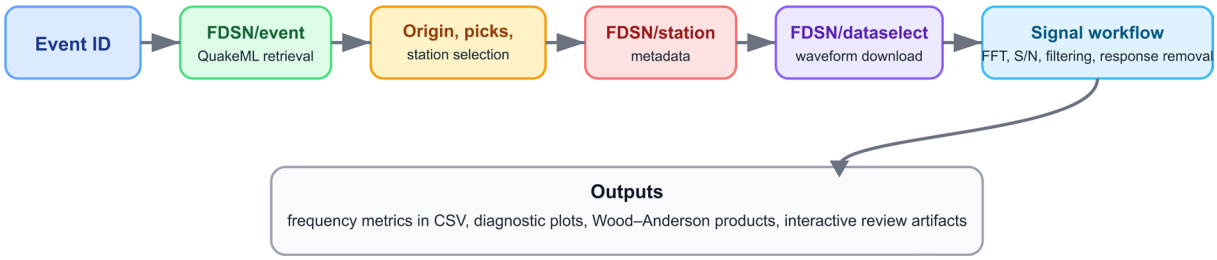


Figure 3. Conceptual workflow of the AI-assisted seismological coding pipeline. The diagram emphasizes the integration between event metadata, seismic web services, waveform processing, and derived outputs, as well as the role of Generative AI in accelerating implementation and code revision rather than replacing scientific interpretation.

The same line of work also included broader coding use cases, grouped in use case UC5, including navigation and modification of a large Java/JavaFX project and conversion of legacy Fortran code to Python. These cases extend the argument beyond single scripts and show that the perceived utility of Generative AI also lies in lowering the entry barrier to complex or unfamiliar software ecosystems. Across these examples, success depended on careful prompting, explicit contextualization of the task, and iterative validation.

4.5 Training and institutional uptake

A final result emerging from the initiative is that the GdL AI process gradually evolved into a capability-building effort, represented by use case UC6. Introductory meetings, thematic seminars, and training activities show that experimentation was accompanied by a deliberate attempt to build shared literacy on AI tools, ethical issues, and responsible use. Group discussions further indicate that participants linked day-to-day experimentation with

wider institutional goals: broader availability of protected licenses, AI literacy for scientific and administrative staff, thematic training on coding and macroseismics, and longer-term positioning of INGV in emerging AI-related funding and project opportunities. One of the outcomes of the project was therefore the creation of a small but structured institutional environment in which experimentation, reflection, and dissemination reinforced one another.

5. Discussion

5.1 Where Generative AI was most useful

The experience of the GdL AI suggests that the most convincing uses of Generative AI were those in which it reduced friction in intermediate workflow layers. Three areas were particularly relevant. First, the interaction with unstructured textual material, especially historical earthquake documentation and macroseismic summaries, shows clear potential for transforming narrative information into structured intermediate artifacts such as tables, ranked comments, and candidate classifications. Second, natural-language interaction with web services and technical interfaces appears valuable because it lowers the access barrier to data infrastructures without changing the scientific criteria by which those data are later interpreted. Third, scientific coding emerges as a particularly strong domain of use, especially when the task requires rapid prototyping, code refactoring, navigation of unfamiliar software projects, or iterative debugging.

A common feature of these successful cases is that the model acts as an accelerator of translation between representations: from prose to structure, from task description to code, from human request to machine-readable query, or from legacy implementation to more maintainable software. This is an important result because it helps distinguish realistic from unrealistic expectations. In the GdL AI experience, the model is most useful when it supports transformation, organization, and first-pass implementation; it is less convincing when asked to serve as a fully autonomous scientific analyst.

5.2 LLMs as a semantic translation layer and support for workflow construction

The documented use cases can be interpreted through a recurring functional distinction between semantic translation and workflow-construction support. Although the cases differ in scientific domain and operational purpose, several of them involve the reformulation of information or task descriptions across heterogeneous representations. Historical narratives were converted into candidate structured information; natural-language requests were translated into web-service queries; scientific processing requirements were reformulated as source code; and legacy code was converted into alternative software implementations. In these cases, the contribution of the LLM did not consist in independently establishing scientific results, but in generating intermediate artifacts that could facilitate subsequent technical operations and expert assessment.

This function is described here as semantic translation, meaning the reformulation of content between representations that differ in structure, syntax, or level of formalization. The term does not imply that scientific meaning was necessarily preserved without alteration. The documented cases indicate that omissions, ambiguities, incorrect parameterization, and technically plausible but scientifically inappropriate outputs remained possible. Model-generated tables, queries, and code must therefore be considered provisional intermediate artifacts whose usability depends on comparison with the original sources, service documentation, software behaviour, and domain-specific knowledge.

In the web-service and coding cases, semantic translation also contributed to workflow construction by assisting with task specification, parameter formulation, code development, and the organization of processing steps. The available evidence does not indicate autonomous orchestration in the stricter sense of independent tool selection, invocation, execution monitoring, and error management. Rather, the LLM supported experts in formulating workflows that were subsequently executed through domain-specific services and software and reviewed for technical correctness and scientific plausibility. The appropriate object of assessment is consequently the reviewed LLM-assisted workflow, including its iterations and expert corrections, rather than the isolated model output.

5.3 Why human validation remained indispensable

Across all the use cases, human validation was not a secondary precaution but the central condition of usefulness. In macroseismic and historical-document workflows, expert judgment remains necessary to evaluate whether extracted localities are complete and reliable, whether damage descriptions have been interpreted correctly, and whether assigned intensity classes are scientifically defensible. In API-related tasks, the correctness of parameters, query structure, and returned data must still be checked against the service specification. In coding tasks, the key issue is not only whether code runs, but whether the proposed solution remains scientifically plausible, technically reviewable, and aligned with the intended workflow. In the meetings, participants repeatedly described the process as one of exposing the problem carefully, constraining the task, inspecting outputs, feeding corrections back into the interaction, and refining the result step by step. This recurring pattern is summarized in Fig. 4. In this sense, the relevant operating model is best described as human-in-the-loop augmentation, in line with broader trustworthy-AI guidance that emphasizes human oversight and traceable responsibility[3].

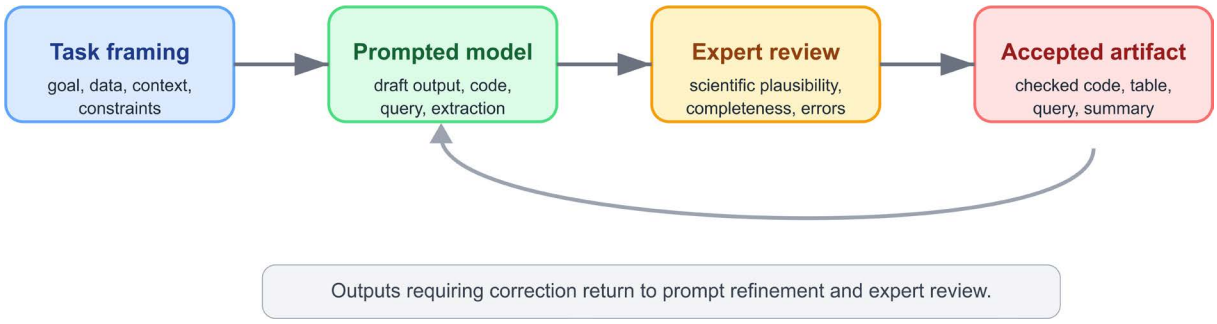


Figure 4. Human-in-the-loop validation pattern emerging across the documented cases. Although the specific tasks vary, the same workflow logic recurs: task framing, AI-assisted draft generation, expert review, and iterative refinement before any output is accepted as scientifically usable.

A further result follows from this validation pattern. In the strongest cases, usefulness did not depend on a single model response, but on the combination of prompting, revision, correction, and expert checking that made the output scientifically usable. The productive unit was therefore the reviewed interaction, not the raw answer alone. This is consistent with recent calls for transparent reporting of AI use in scientific practice (Van Dis et al., 2023).

5.4 Implications for reproducibility and evidence

One of the main challenges raised by this experience is reproducibility. Many of the reported outcomes did not arise from a single prompt-response exchange, but from iterative interactions in which the prompt was progressively refined, errors were diagnosed, and the task was reformulated. This makes the process productive, but also harder to reproduce in the strict sense expected of conventional computational methods. The problem is not unique to the working group described in this work: it is intrinsic to conversational systems whose outputs depend on task framing, context accumulation, model version, and limited explainability.

Indeed, comparable institutional experiences also suggest that these challenges are not confined to the INGV case. Liu and Jagadish (2024) emphasize the need for shared institutional support, training, and validation practices when researchers adopt rapidly evolving Generative AI tools. The focused deployment described by Wilson et al. (2025) similarly relied on iterative development, expert involvement, and secure handling of institutional information, while Thaldar et al. (2025) situate capacity building and governance within a broader institutional response.

3 Ethics Guidelines for Trustworthy AI, European Commission, Brussels. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (last accessed July 2026).

For this reason, reproducibility cannot be claimed here in a deterministic sense. What can be provided, however, is a transparent reconstruction of the process: representative cases, explicit workflow descriptions, and clear indication of where expert validation entered the loop. This does not eliminate the problem, but it makes the reported experience inspectable and interpretable. More broadly, the tension between performance, transparency, and explainability has been recognized in major AI ethics and governance frameworks, which explicitly treat transparency, traceability, and explicability as central requirements rather than optional additions[4] (Floridi and Cowls, 2019).

5.5 Institutional adoption as a socio-technical process

Another important finding is that the GdL AI initiative was not only a collection of isolated experiments. The repeated emphasis on thematic meetings, shared licenses, training, privacy settings, and ethical discussion indicates that adoption was being treated as an institutional process. Group discussions make it clear that participants were aware of the risk that many staff members were already using public AI services informally, potentially with work-related data, and saw managed licenses as a partial institutional response to that risk. This is important because the practical value of Generative AI in research settings depends heavily on surrounding conditions: whether users know how to formulate tasks, whether data-handling boundaries are explicit, whether examples and practices are shared, and whether experimentation is socially legitimized inside the organization. In this respect, the GdL AI experience suggests that the introduction of Generative AI should be understood as a socio-technical capability-building exercise. Technology alone does not generate value. Value emerges when a research group develops a common language around use cases, validation, risks, and good practice.

5.6 Privacy, governance, and responsible use

Privacy and governance concerns were present from an early stage. The discussion of Generative AI Team, non-use of team data for model training, and explicit attention to handling of sensitive material indicate that technical experimentation was accompanied by reflection on boundaries of acceptable use. In a public research institution, unpublished scientific material, operational data, and internal workflows cannot simply be treated as neutral inputs for external platforms. For this reason, responsible adoption emerged as part of the result rather than as an external constraint. Protected access, clearer rules on suitable data, and shared discussion of risks made experimentation more credible and more usable. This is also coherent with broader trustworthy-AI frameworks, in which human oversight, privacy and data governance, transparency, and accountability are treated as core requirements for acceptable use[5] (Floridi and Cowls, 2019).

5.7 Considerations on scope and limitations

The qualitative nature of this study should be kept in view when interpreting its findings. The aim was not to produce a benchmark or a controlled quantitative evaluation, but to reconstruct and compare a set of institutional use cases in a structured way. As in other qualitative studies, the emphasis therefore falls on interpretive clarity, process transparency, and contextual understanding rather than on deterministic repeatability or statistical generalization. A second point concerns timing. The experience described here started in the first half of 2024 and emerged in a phase strongly shaped by the rapid diffusion and evolution of ChatGPT, but the broader landscape already includes other relevant systems such as Gemini, Claude, and open models such as Mistral 7B (Gemini Team et al., 2023; Jiang et al., 2023; OpenAI et al., 2023). The picture should therefore be read as a snapshot of an evolving technological moment rather than as a stable endpoint.

Since the period documented in this study, agentic AI systems have begun to extend LLM-based support beyond response generation, enabling the selection and invocation of external tools, the execution of multi-step

4 Ethics Guidelines for Trustworthy AI, European Commission, Brussels. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (last accessed July 2026).

5 Ethics Guidelines for Trustworthy AI, European Commission, Brussels. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (last accessed July 2026).

procedures, and the monitoring of intermediate results. In particular, contemporary coding agents, including commercial environments such as Claude Code[6] (Anthropic, 2025) and open-source alternatives such as OpenCode (OpenCodeAI, 2025) and pi coding agent[7] (Zechner, 2025), can inspect project context, modify files, invoke development tools, run tests, and iterate on the basis of execution feedback, subject to the permissions and tool access granted by the user. Recent work on scientific agents likewise illustrates the potential for multi-tool orchestration in complex research workflows (Ding et al., 2025). In the INGV cases, this evolution could transform UC3 from a conversational interface for formulating web-service queries into supervised orchestration of sequential data-access operations, and could extend UC4 and UC5 from code drafting and debugging assistance toward bounded execution, testing, and workflow maintenance. These possibilities do not reduce the importance of human oversight; rather, as agents act across more steps and systems, validation, provenance, access controls, and responsibility become even more central.

This matters especially in Europe, where open or more controllable models may become increasingly relevant for simpler and more repetitive tasks, or wherever institutional constraints make dependence on large commercial platforms less desirable. Within these bounds, the contribution of the paper remains clear: it documents an early institutional phase in which several geoscience-relevant workflows were tested, some recurring patterns of usefulness became visible, and a more explicit culture of validation and responsible experimentation began to take shape.

6. Conclusions

This paper has reconstructed the early experience of the INGV GdL AI initiative as a structured qualitative case study of Generative AI adoption in geoscientific workflows. The strongest cases show that current LLM-based systems can already provide meaningful support in tasks that sit between raw scientific material and final expert interpretation, especially where information must be reorganized, translated into structure, or operationalized into code and queries.

At the same time, the study provides no basis for claims of autonomous scientific replacement. Across macroseismics, historical-document analysis, API interaction, and scientific coding, the decisive factor was the continued presence of domain experts who validated outputs, corrected errors, and decided what could be accepted as scientifically usable. The most defensible conclusion is therefore that Generative AI is presently best understood as an augmentation layer within a human-led workflow.

A second conclusion concerns institutional practice. The INGV experience suggests that the value of Generative AI depends not only on access to tools, but also on shared training, privacy-aware governance, and explicit documentation of use cases. In this sense, the initiative was important not simply because it tested new tools, but because it began to organize a more structured internal culture around experimentation, validation, and responsible adoption.

Future work should move in two parallel directions: first, deeper documentation of representative workflows with stronger supporting material; and second, broader comparison with other models, tasks, and institutional settings. Even within its limits, the present case study shows that early institutional experimentation with Generative AI can generate both practical workflow benefits and strategic organizational learning when it is conducted critically and transparently.

Data availability statement. The material underlying this paper consists of internal working documents produced during the GdL AI initiative, including meeting notes, presentation slides, and related reporting material. No dedicated scientific dataset was created for the purposes of this study. These materials are not deposited as a public package, but selected non-sensitive items may be made available upon reasonable request to the corresponding author, subject to institutional constraints.

Acknowledgements. The authors thank the participants of the INGV GdL AI initiative for their contributions to the experimentation process, in particular: Umberto Fracassi, Giuseppe Falcone, Emanuele Casarotti, Barbara Angioni,

6 Claude 3.7 Sonnet and Claude Code official announcement: https://www.anthropic.com/news/claude-3-7-sonnet (last accessed July 2026).

7 Zechner, M. (2025, November 30). What I learned building an opinionated and minimal coding agent. https://mariozechner.at/posts/2025-11-30-pi-coding-agent/

Ludovico Vitiello, Paola Bordoni, Kety Giuliacci, Rossana Paciello, Maria Grazia Ciaccio, Aladino Govoni, Manuela Sbarra, Alessandro Pignatelli, Massimo Crescimbene, Gianluca Chessa, Gianfranco Vannucci, Rita Chiara Taccone, Rossella Fonzetti, Giovanna Maracchia, Luca Merucci, Gaia Soldati, Salvatore Barba. The authors thank INGV Roma 1 Director Giuseppe Di Stefano for having supported and facilitated the initiative. Generative AI tools were used during the drafting and revision of parts of this manuscript to improve language, structure, and overall clarity. All AI-assisted outputs were reviewed, revised, and validated by the authors, who take full responsibility for the final text and its scientific content.

References

Arroyo-Machado, W., J. Ma, T. Chen, T. P. Johnson et al. (2025). Generative AI and academic scientists in US universities: Perception, experience, and adoption intentions, PLOS ONE, 20, 8, e0330416, doi:10.1371/journal.pone.0330416.

Bergen, K. J., P. A. Johnson, M. V. De Hoop and G. C. Beroza (2019). Machine learning for data-driven discovery in solid Earth geoscience, Science, 363, 6433, eaau0323, doi:10.1126/science.aau0323.

Ding, K., J. Yu, J. Huang, Y. Yang et al. (2025). SciToolAgent: A knowledge-graph-driven scientific agent for multitool integration, Nat. Comput. Sci., 5, 10, 962-972, doi:10.1038/s43588-025-00849-y.

Floridi, L. and J. Cowls (2019). A Unified Framework of Five Principles for AI in Society, Harvard Data Sci. Rev., doi:10.1162/99608f92.8cd550d1.

Gemini Team, R. Anil, S. Borgeaud, J. B. Alayrac et al. (2023). Gemini: A Family of Highly Capable Multimodal Models, Version 5, arXiv, doi:10.48550/ARXIV.2312.11805.

Guidoboni, E., G. Ferrari, G. Tarabusi, C. Sgattoni et al. (2019). CFTI5Med, the new release of the catalogue of strong earthquakes in Italy and in the Mediterranean area, Sci. Data, 6, 1, 80, doi:10.1038/s41597-019-0091-9.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford et al. (2023). Mistral 7B, Version 1, arXiv, doi:10.48550/ARXIV.2310.06825.

Liu, J. and H. V. Jagadish (2024). Institutional efforts to help academic researchers implement generative AI in research, Harvard Data Sci. Rev., Special Issue 5, doi:10.1162/99608f92.2c8e7e81.

Mazza, S., A. Basili, A. Bono, V. Lauciani et al. (2012). AIDA – Seismic data acquisition, processing, storage and distribution at the National Earthquake Center, INGV, Ann. Geophys., 55, 4, doi:10.4401/ag-6145.

Musson, R. M. W., G. Grünthal and M. Stucchi (2010). The comparison of macroseismic intensity scales, J. Seismol., 14, 2, 413-428, doi:10.1007/s10950-009-9172-0.

OpenAI, J. Achiam, S. Adler, S. Agarwal et al. (2023). GPT-4 Technical Report, Version 6, arXiv, doi:10.48550/ARXIV.2303.08774.

Ritchie, J. and L. Spencer (1994). Qualitative data analysis for applied policy research, In Analyzing Qualitative Data, Routledge, 22, doi:10.4324/9780203413081.

Schorlemmer, D., F. Euchner, P. Kästli and J. Saul (2011). QuakeML: status of the XML-based seismological data exchange format, Ann. Geophys., 54, 1, doi:10.4401/ag-4874.

Thaldar, D., M. Botes, A. Badru, H. Chenia et al. (2025). Generative AI governance in higher education: A case study from Africa, Front. Polit. Sci., 7, 1666661, doi:10.3389/fpos.2025.1666661.

Van Dis, E. A. M., J. Bollen, W. Zuidema, R. Van Rooij et al. (2023). ChatGPT: five priorities for research, Nature, 614, 7947, 224-226, doi:10.1038/d41586-023-00288-7.

Wilson, L. A., B. Konsynski and T. Yisrael (2025). Advancing research administration with AI: A case study from Emory University, J. Res. Admin., 55, 52-73, https://www.srainternational.org/blogs/srai-jra2/2025/05/22/advancing-research-administration-with-ai-emory.

Wang, S., T. Hu, H. Xiao, Y. Li et al. (2024). GPT, large language models (LLMs) and generative artificial intelligence (GAI) models in geospatial science: a systematic review, Int. J. Digit. Earth, 17, 1, 2353122, doi:10.1080/17538947.2024.2353122.

Yin, R. K. (2018). Case study research and applications: design and methods, Thousand Oaks, California, SAGE Publications, Inc., 352, ISBN:9781506336169.

***CORRESPONDING AUTHOR: Stefano ARETUSINI**,
Istituto Nazionale di Geofisica e Vulcanologia, Sezione Roma 1, Roma, Italy
e-mail: stefano.aretusini@ingv.it

© 2026 the Author(s).
Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License