GALICIA

May 2025

Alberto Stefanini (ed., Novareckon), Eleonora Dorissi (Novareckon)

Jérémie Farret (Mind in a Box), Lorenzo Vandoni (HAL Service)

KPI13 – Platform Evaluation and Verification

Funded by the European Union

Funded by the **European Union**. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. Funded within the framework of the **NGI Sargasso** project under grant agreement No **101092887**

Project Name: GALICIA
Report Title: KPI 13 – Platform Verification and Evaluation
Authors: Alberto Stefanini (Novareckon), Eleonora Dorissi (Novareckon), Jérémie Farret (Mind in a Box), Lorenzo Vandoni (HAL Service)
Revised by: François Monette (Mind in a Box)
Date: 29/05/2025
Version: 1
Distribution: Public

TABLE OF CONTENTS

SUMMARY

This report summarizes the second evaluation of the GALICIA platform, following a prior assessment conducted by a qualified expert panel using the System Usability Scale (SUS) questionnaire.

The current evaluation aims to build on that foundation by gathering more structured feedback and exploring the practical usability of the platform through a brief hands-on exercise. Specifically, the evaluation included:

- The completion of a mini-use case: a small-scale application of the GALICIA platform, designed to verify its core functionalities in action.
- The submission of a tailored Evaluation Questionnaire: designed to collect targeted feedback on user experience and perceived platform value.

This second evaluation was useful to assess the relevance and applicability of GALICIA in both practical and educational contexts, particularly in domains such as manufacturing, critical infrastructure, cybersecurity, and trustworthy AI.

INTRODUCTION

This report presents the findings of the second evaluation of the GALICIA platform, conducted during the final phase of the project in mid-2025. It builds upon the results of the previous assessment exercise—documented in KPI10—whose outcomes highlighted both promising features and critical areas in need of refinement. The first round of evaluations, held in April 2025, offered a mixed perspective on GALICIA's usability and effectiveness, praising elements such as ease of use and interface design, while identifying several unsatisfactory aspects that undermined the perceived maturity and reliability of the platform.

In particular, users expressed strong concerns about the opacity of the generation and validation process, the occasional misalignment between input prompts and outputs, and the difficulty in understanding the system's internal logic. These issues, combined with a lack of detailed feedback on code validation outcomes, significantly impacted trust in the tool—especially for users accustomed to traditional formal verification environments.

The present evaluation was therefore motivated by the need to reassess GALICIA's performance following these criticisms. A second round of testing was planned to determine whether the platform was on track to meet its objectives and to measure the impact of initial corrective actions. Between April and May 2025, a limited set of improvements was introduced—mostly targeting minor interface clarifications, better error reporting, and the resolution of a few inconsistencies in code generation. While this set of changes was not extensive, it aimed at resolving the most urgent problems that had emerged in the first evaluation.

Accordingly, the second evaluation took place in May and June 2025, involving a revised set of consultations with domain experts and testers. The goal was not only to verify whether specific technical issues had been resolved, but also to explore deeper systemic problems in the platform's logic, usability, and transparency.

The structure of this document reflects this intent. Following this introduction, the report outlines the evaluation methodology and computation framework, describes the consultation process and participant profiles, and presents an analysis of each response. The document concludes with an overall evaluation summary, and a synthesis of the key takeaways and recommendations for future development.

1. METHODOLOGY

The evaluation methodology applied to the GALICIA platform in this second round represents a significant departure from that adopted during the KP10 milestone. At that time, the evaluation involved a panel of external evaluators, composed of practitioners, research fellows, and stakeholder experts. Their assessment was based on the System Usability Scale (SUS) questionnaire.

In contrast, the current methodology aims at obtaining a more structured and meaningful picture of the platform's usability and functionality, with particular attention to the users' experience in realistic conditions. To this end, the evaluation was:

- Opened to a broader and more diversified set of external evaluators, including professionals with different levels of experience in software verification and cybersecurity.
- Organized around the execution of a practical test case, in which each evaluator was required to:
  * Develop a coding sample of their choice;
  * Produce a corresponding formal model using GALICIA;
  * Verify and validate the model using the platform;
  * Reflect on the interaction and performance of the system.

After completing the hands-on testing, each evaluator was required to fill in a custom, structured questionnaire, aimed at assessing user satisfaction, clarity of results, and perceived value of the tool. The questionnaire was organized into five sections:

- Platform Functionality & Performance
- Workflow and Transparency
- Potential Use and Value
- Engagement & Contribution
- Open Feedback

Although qualitative in nature, the questionnaire was designed to allow for quantitative post-processing of the results. Most of the questions were expressed in Yes/No format, and were numerically codified by assigning a score of 0 to "No" answers and 1 to "Yes" answers. Likewise, questions using an ordinal scale such as No, Sometimes, Possibly, Frequently, Usually were mapped to values 0, 0.25, 0.5, 0.75, 1 respectively.

This approach permits the construction of a summatory indicator expressing the overall degree of UX satisfaction, in analogy to scoring methods used in established instruments such as the System Usability Scale (SUS). Such a metric allows for comparative insights and the tracking of platform improvements over time.

Computability of UX Score

While qualitative in form, the structured questionnaire was expressly designed to allow for quantitative post-processing. This was achieved by assigning numerical scores to the answers as follows:

For Yes/No questions, responses were mapped to binary values:

Yes = 1, No = 0

For scaled questions (e.g., No, Sometimes, Possibly, Frequently, Usually), values were assigned on a linear scale:

No = 0, Sometimes = 0.25, Possibly = 0.5, Frequently = 0.75, Usually = 1

The total UX score for each evaluator, denoted UX_e, can be computed as:

$$UX_e = (1/N) \sum_1^n s_i$$

where:

N is the number of scored questions (excluding open comments),

s_i is the score assigned to the i-th response.

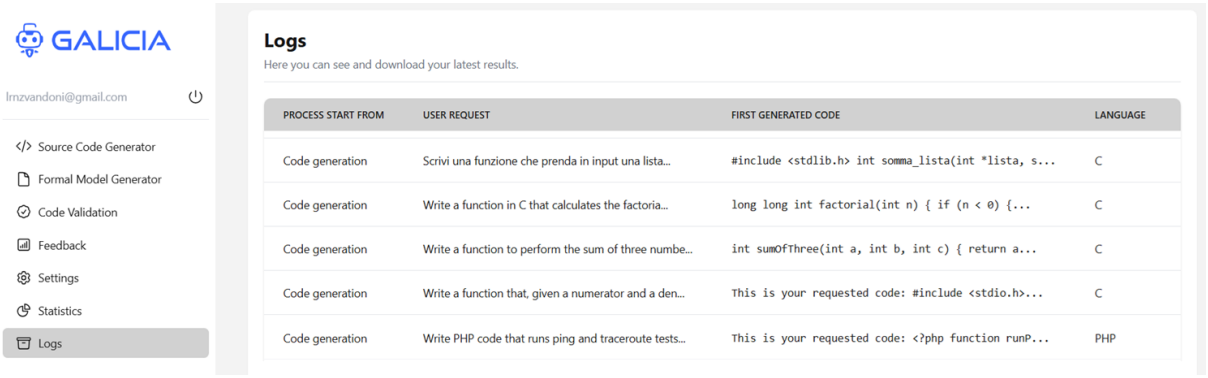
This approach yields a UX satisfaction indicator ranging from 0 (no satisfaction) to 1 (full satisfaction), analogous in spirit to the System Usability Scale (SUS). Aggregate statistics (mean, standard deviation) over all evaluators can then be used to assess the platform's perceived usability and identify key areas for improvement.

2. USAGE STATISTICS

This section presents usage statistics and a comparative assessment of code quality produced by the GALICIA platform. The objective is to empirically demonstrate that the code generated through GALICIA is, while not flawless, generally “slightly better” than the code initially produced by the selected large language model (LLM) when tackling the same task.

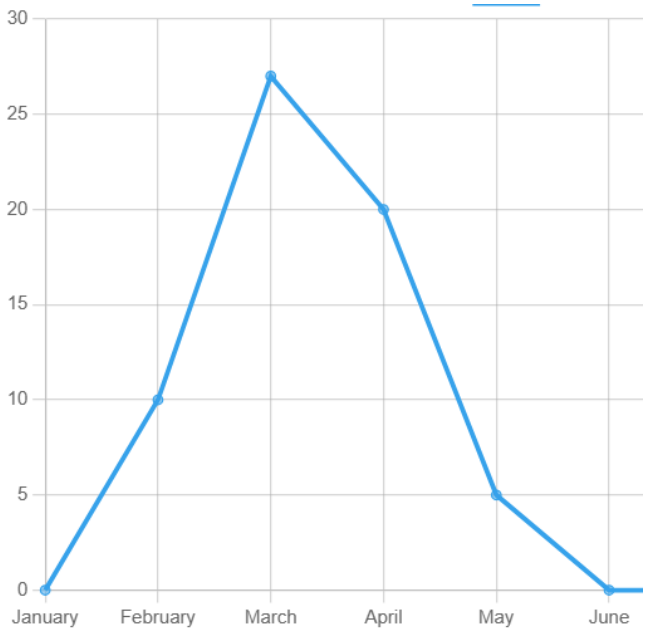
The Galicia platform **records all the work performed by the registered users**.

Each user is then allowed to view all the sessions performed, in section *Logs*, and can also export them in CSV format using button *Export*.

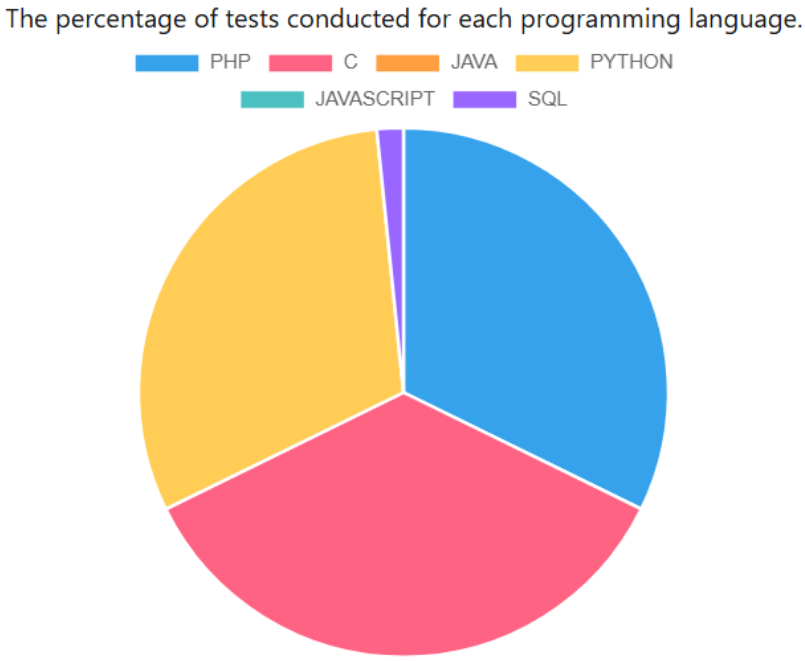


Galicia also contains a section called *Statistics*, which is only available to administrators, that helps to **extract relevant information from the user sessions** by all registered users.

Statistics show that Galicia, starting from February 2025 what it has been first released, until the beginning of May when this report has been prepared, has been used to **perform 62 sessions**:



The sessions have been conducted **in different programming languages**, mainly C, Python and PHP:



Statistics are also used for **measuring the affordability and correctness of the generated source code**, and comparing the source code generated by the selected LLM in response to the original user prompt with the source code obtained from Galicia after the different iterations that Galicia performs to verify and correct this source code.

The measure used is the number of test cases passed. Galicia, in fact, after generating the source code and the corresponding formal model, creates a set of test cases based on them, and then verifies, performing a static source code analysis, if the source code passes these tests. In case the source code does not pass all the tests, Galicia performs a new iteration, creating an updated version of it, and then performing again all verifications.

In the user sessions performed at the moment of writing this document, we have collected the following results:

- Number of sessions where the source code passed all tests on the first iteration: **19 (31%)**. This means that only in 31% of the user sessions, the source code generated by the selected LLM in response to the original user prompt has passed all the tests

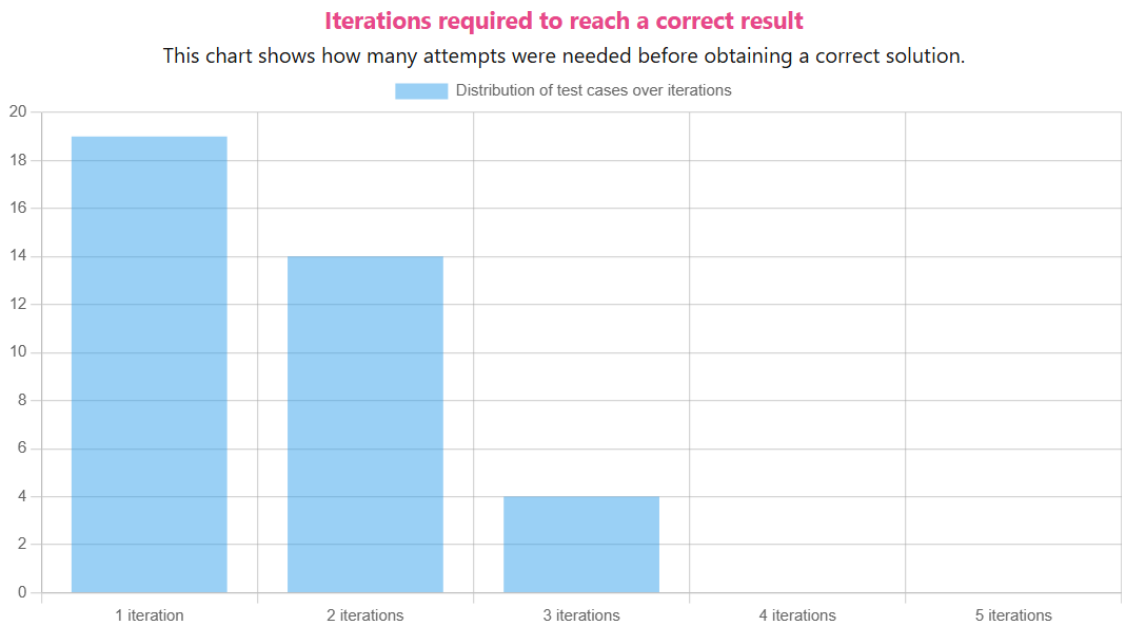
- Number of sessions where the source code passed all tests after the last iteration: **18 (29%)**. This means that in 29% of the user sessions, the source code generated by the selected LLM in response to the original user prompt did not pass all the tests, but the source code obtained from Galicia after the different iterations has passed all the tests

- Number of sessions where the source code did not pass all tests after the last iteration: **25 (40%)**. This means that in 40% of the user sessions, the source code generated by the selected LLM in response to the original user prompt did not pass all the tests, and also the source code obtained from Galicia after the different iterations did not pass all the tests. This might also depend on the maximum number of allowed iterations, that can be set by each user among application options, and that has a default value of 3.

- Percentage of passed tests after the first iteration: **89%**. This is the percentage of tests passed by the all-source codes generated by the selected LLM in response to the original user prompt

- Percentage of tests passed after the first iteration: **93%**. This is the percentage of tests passed by the all-source codes obtained from Galicia after the different iterations

The number of iterations needed to get a correct result varies from 1 to 3. Again, this might also depend on the maximum number of allowed iterations, that can be set by each user among application options, and that has a default value of 3.



The evaluation of the platform is a critical component in assessing its effectiveness, usability, and overall user satisfaction. To ensure objectivity and credibility, the proposed methodology places strong emphasis on involving external evaluators who are not directly linked to the project. This strategic choice reduces the risk of bias and allows for more impartial feedback, which is especially important when attempting to draw reliable conclusions about the platform's real-world performance and acceptance.

3. CONSULTATION PROCESS AND RESPONDENT OVERVIEW

Participant Selection Criteria

The invitation to take part in the evaluation was addressed, on one hand, to the panel previously involved in the KPI10 assessment phase, thus ensuring continuity and comparability. On the other hand, a complementary group was selected from among the nominatives registered in the GALICIA project mailing list, prioritizing individuals who had previously expressed interest in being involved in testing and feedback activities.

This resulted in a reasonably representative sample of the project's broader target audience, comprising:

- Representatives of small and medium enterprises (PMIs), particularly active in cybersecurity or software development;
- Experts in user experience (UX), natural language (NL), and formal methods (FM);
- Exponents from two national organizations with a prominent role in cybersecurity awareness and practice in Italy:
  - AIIC (Associazione Italiana Infrastrutture Critiche), focused on the protection and

resilience of critical infrastructures;
- CLUSIT (Associazione Italiana per la Sicurezza Informatica), a leading Italian association committed to promoting the culture of information security across public and private sectors.

Additionally, a small number of individuals from HAL and Mind in a Box not directly involved in the project, were invited and accepted to share their observations, further enriching the evaluation with external perspectives.

The Evaluation Panel

The following experts contributed to the consultation through interviews or by answering a structured questionnaire. Their profiles reflect a wide spectrum of expertise, from technical standardization to cybersecurity risk governance, and from SME-level implementation to policy-level insight.

- **Glauco Bertocchi** – Co-Founder and CEO, Doing Next s.r.l. (Italy)
  A digital entrepreneur based in Rome, Bertocchi leads Doing Next, a cybersecurity-oriented SME offering advanced IT services. He provided insights into the challenges faced by small businesses in adopting and maintaining compliance with ISO/IEC standards, including the impact of resource constraints and audit complexity.
- **Sandro Bologna** – Independent expert, formerly ENEA and CIEM
  With over 40 years of experience in critical infrastructure protection and cybersecurity, Bologna has contributed to national and European initiatives on resilience and risk management. He is the author of key studies on cyber risk in operational technologies and has been involved in defining methodological frameworks for risk assessment in both private and public sectors.
- **Elenio Dursi** *(Dfree/ CLUSIT, Italy)* - Cybersecurity expert and UX consultant, involved in various public-private initiatives on ICT security. LinkedIn
- **Franco D'Urso** (Emisfera, Italy) - Senior developer and technical manager, with extensive experience in system integration and front-end development.
- **Alessandro Gallina** *(HAL Service, Italy)* - ICT professional with experience in digital platforms for industrial and mobility applications.
- **Roberto Mascheroni (***HMS IT S.r.l., Italy***)** - Representing H.M.S. S.r.l. Health Medicine Services, a company operating in the medical sector, exclusively focused on homeopathy. Since 1991, it has been developing software specifically designed for use in homeopathic medicine.
- **Giovanna Dondossola (***RSE S.p.a. (Italy)* Leading Scientist at the Transmission and Distribution Technologies Department of the RSE SpA where she technically manages and leads National and European projects on the evaluation of cyber risks in energy system communications.

- **Serge Demeyer (***Professor at Universiteit Antwerpen, Belgium)* - Serge Demeyer is a professor at the University of Antwerp (Department of Mathematics and Computer Science) and the spokesperson for the ANSYMO (Antwerp System Modelling) research group. He directs a research lab investigating the theme of "Software Reengineering". In 2007 he received a "Best teacher" award from the Faculty of Sciences at the University of Antwerp. As a consequence, he remains very active in all matters related to teaching quality.
- **Jerin, Nitish, Jean-Christophe (Canada)** – These Canada-based data-science and software specialists contributed perspectives on LLM applicative requirements in North America, especially concerning cross-jurisdictional recognition, AI-supported certification, and public incentives for SME compliance. They requested anonymity due to professional confidentiality agreements.

4. REPLY ANALYSIS

Glauco Bertocchi:

Platform Functionality & Performance:

The example function provided did not pass all verification checks and did not effectively handle edge cases. The verification/validation feedback was not understandable or reliable. GALICIA's feedback was clear and aligned with expected results. Errors were identified in GALICIA's results. The initial omission of input validation was identified and corrected in a subsequent iteration. The system clearly communicated the programming language and formal method used. Overall, there is no trust in GALICIA's output.

Workflow and transparency:

The pipeline logic of GALICIA was understandable. Additional documentation on verification mechanisms would be useful. Including detailed explanations for each verification step would improve understanding. Suggestions included providing "online help" and "additional documentation" as explanations.

Potential use and value:

GALICIA would not be used for real-world cases. The platform shows promise for applications requiring formal verification, pending further validation. Compared to traditional verification tools, GALICIA offers a more streamlined and intuitive interface. No improvements are needed to match the depth of specialized tools. It was suggested that code validation with some explanation would make GALICIA more useful.

Engagement and contribution:

GALICIA is more streamlined and intuitive than traditional verification tools. The respondent is not interested in future testing.

Open feedback:

The generated code was formally correct but behaved unexpectedly due to a logical error.

Final UX Score for Bertocchi = 0.529 → 53%

Sandro Bologna:

Functionality and performance of the platform:

The example function provided passed all verification checks and effectively handled edge cases. The verification/validation feedback was understandable and reliable. GALICIA's feedback was unclear and not aligned with expected results. No errors were identified in GALICIA's results. The initial omission of input validation was identified and corrected in a subsequent iteration. The system clearly communicated the programming language and formal method used. Overall, trust in GALICIA's output is low.

Workflow and transparency:

GALICIA's pipeline logic was understandable. Additional documentation on verification mechanisms would be useful. Including detailed explanations for each verification step would improve understanding. There was a call for more transparency on how the verification was conducted (VERIFICATION vs. WHAT).

Potential use and value:

GALICIA would not be used for real-world cases. The platform shows promises for formal verification applications, pending further validation. Compared to traditional verification tools, GALICIA does not offer a more streamlined and intuitive interface. Improvements are needed to match the depth of specialized tools. There was confusion about what would make GALICIA more useful. More transparency was requested, especially on what exactly is being verified.

Engagement and contribution:

GALICIA is not more streamlined and intuitive than traditional tools. The respondent is interested in future testing.

Open feedback:

It was clarified that if “code verification” means “verification against specific test cases,” the answer is yes.

Final UX Score for Bologna= 0,661 →66.1%

Alessandro Gallina:

Functionality and performance of the platform:

The example function passed all verification checks and handled edge cases effectively. The verification/validation feedback was understandable and reliable. GALICIA’s feedback was clear and aligned with expected results. No errors were identified. The initial omission of input validation was identified and corrected. The system clearly communicated the programming language and formal method used. The code and its formal model were transparent and well-documented. Overall, there is high confidence in GALICIA’s output.

Workflow and transparency:

The pipeline logic was understandable. Additional documentation would be useful. Detailed explanations for each verification step would improve understanding. No suggestions were made to improve the experience.

Potential use and value:

GALICIA would be used occasionally for real-world cases. The platform shows promise for formal verification applications, pending further validation. It is unclear whether GALICIA offers a more streamlined and intuitive interface than traditional tools or if improvements are needed to match specialized tools. Being integrated into larger projects would make GALICIA more useful.

Engagement and contribution:

Unclear if GALICIA is more streamlined than traditional tools. The respondent is interested in future testing.

Open feedback:

GALICIA shows significant potential for automating code verification processes. Further integration with established standards and comprehensive documentation is recommended to fully realize its capabilities.

Final UX Score for Gallina = 0.838 → 83.8%

Roberto Mascheroni:

Functionality and performance of the platform:

The example function passed all verification checks and effectively handled edge cases. The verification/validation feedback was understandable and reliable. GALICIA's feedback was clear and aligned with expected results. No errors were found. The input validation omission was corrected in later iterations. The system clearly communicated the programming language and formal method. The code and its formal model were transparent and well-documented. Overall, there is high confidence in GALICIA's output.

Workflow and transparency:

GALICIA's pipeline logic was understandable. The overall process was clear. It was not considered necessary to include detailed explanations for each verification step.

Potential use and value:

GALICIA would be used for real-world cases. The platform shows promise for formal verification applications, pending further validation. It offers a more streamlined and intuitive interface than traditional tools and requires no improvements to match their depth.

Engagement and contribution:

GALICIA is more streamlined and intuitive than traditional tools. The respondent is interested in future testing.

Open feedback:

GALICIA demonstrates significant potential in automating code verification processes. Full realization of its capabilities requires further integration with established standards and complete documentation.

Final UX Score for Mascheroni = 0.882 → 82.2%

Elenio Dursi:

Functionality and performance of the platform:

The example function did not pass all verification checks and did not handle edge cases effectively. The verification/validation feedback was understandable and reliable. GALICIA's feedback was unclear and not aligned with expectations. Errors were identified in GALICIA's results. The omission of input validation was corrected in subsequent iterations. The system clearly communicated the

coding language and formal method. The code and its formal model were transparent and well-documented. Overall, there is high confidence in GALICIA's output.

Workflow and transparency:

The pipeline logic was understandable. The process was clear. No additional documentation or detailed explanations for each step were deemed necessary.

Potential use and value:

GALICIA would be used frequently for real-world cases. The platform does not show promise for formal verification applications, pending further validation. GALICIA offers a more streamlined and intuitive interface than traditional tools and does not require improvements to match their depth.

Engagement and contribution:

GALICIA is more streamlined and intuitive than traditional tools. The respondent is interested in future testing.

Open feedback:

GALICIA demonstrates significant potential in automating code verification processes. Full realization of its capabilities requires further integration with established standards and complete documentation.

Final UX Score for Dursi = 0.617 → 61.7%

Franco D'Urso:

Platform Functionality & Performance:

The sample function passed verification checks, and the feedback was understandable and trustworthy. GALICIA's feedback was clear and aligned with expected outcomes, and no outcome errors were identified. However, there was an initial omission of input validation that was corrected. The system clearly communicated the coding language and formal method used. The code and its formal model were transparent but lacked documentation. Overall, there's high confidence in GALICIA's output.

Workflow and Transparency:

GALICIA's pipeline logic was understood. Additional documentation on verification mechanisms and detailed explanations for each step would be beneficial, although the process is already mostly

clear. A suggestion for improvement is the possibility to request modifications to the generated source code.

Potential Use and Value:

GALICIA shows promise for applications requiring formal verification and might be used for real-world cases, pending further validation. Several respondents found questions about GALICIA's interface compared to traditional tools unclear. Integration into the development environment was suggested to make GALICIA more useful.

Engagement & Contribution:

There is interest in future testing of GALICIA.

Open Feedback:

GALICIA demonstrates significant potential in automating code verification processes. Recommendations include further integration with established standards and comprehensive documentation. Some feedback mentioned that the tool is nice but limited by the possibility of asking only one question at a time.

Final UX Score for D'urso = 0.789 → 79%

Serge Demeyer:

Functionality and Performance of the Platform:

The example function passed all verification checks and effectively handled edge cases. The verification/validation feedback was clear and reliable. The feedback from GALICIA was consistent with the expected outcomes, and no errors were identified in GALICIA's results. The initial omission of input validation was identified and corrected in a subsequent iteration, and the system clearly communicated the coding language and the formal method used. Overall, the respondent expressed high confidence in GALICIA's output.

Workflow and Transparency:

The respondent understood the logic of GALICIA's pipeline and found the overall process to be clear, though suggested that additional documentation on the verification mechanisms would be helpful. Including detailed explanations for each verification step was not considered necessary.

Potential Use and Value:

The respondent would not use GALICIA for real-world cases at this stage but acknowledged that the platform shows promise for applications requiring formal verification, pending further validation. Compared to traditional verification tools, GALICIA offers a more streamlined and intuitive interface, although improvements may be needed to match the depth of specialized tools.

Suggestions for Improving GALICIA:

The respondent provided several suggestions for enhancing GALICIA, including the visualization of changes made during "Code Validation," conducting usability studies, and fixing issues related to the "Save" button, test case display, and the loss of previous prompts. Additional suggestions included showing the actual test code, improving the report, and providing a markdown file with code snippets and execution steps.

Engagement and Contribution:

GALICIA is considered more streamlined and intuitive than traditional verification tools. The respondent is not interested in participating in future testing.

Open Feedback:

The respondent noted that GALICIA shows significant potential in automating code verification processes but recommended further integration with established standards and the development of comprehensive documentation.

Final UX Score for Demeyer = 0,691 →69.1%

Giovanna Dondossola:

Functionality and Performance of the Platform:

The example function passed all verification checks and effectively handled edge cases. The verification/validation feedback was clear and reliable. However, GALICIA's feedback was unclear, not aligned with the expected results, and errors were identified in GALICIA's output. The initial omission of input validation was identified and corrected in a subsequent iteration, and the system clearly communicated the coding language and formal method used. Overall, the respondent expressed low confidence in GALICIA's output.

Workflow and Transparency:

The respondent understood the logic of GALICIA's pipeline and found the overall process to be clear, but suggested that additional documentation on the verification mechanisms would be helpful. Including detailed explanations for each verification step would improve understanding, and the

respondent recommended providing further documentation on the supported programming languages and configuration options.

Potential Use and Value:

The respondent would occasionally use GALICIA for real-world cases and acknowledged that the platform shows promise for applications requiring formal verification, pending further validation. Compared to traditional verification tools, GALICIA offers a more streamlined and intuitive interface, though improvements may be needed to match the depth of specialized tools.

Engagement and Contribution:

GALICIA is considered more streamlined and intuitive than traditional verification tools. The respondent is interested in participating in future testing.

Open Feedback:

The respondent believes that comprehensive evaluations of GALICIA by experienced developers are necessary to assess the overall benefits and costs of migrating to an AI-based coding paradigm.

Final UX Score for Dondossola = 0,794 → 79.4%

Jerin:

Functionality and Performance of the Platform:

The example function passed all verification checks and effectively handled edge cases. The verification/validation feedback was clear and reliable. GALICIA's feedback was clear and aligned with the expected results, although errors were identified in GALICIA's output. The initial omission of input validation was identified and corrected in a subsequent iteration, and the system clearly communicated the coding language and formal method used. The provided code and its formal model were transparent and well documented. Overall, the respondent expressed strong confidence in GALICIA's output.

Workflow and Transparency:

The respondent "somewhat" understood the logic of GALICIA's pipeline and found the overall process to be clear, but suggested that additional documentation on the verification mechanisms would be helpful. Including detailed explanations for each verification step would improve understanding. The respondent recommended providing a clear explanation of what the error was and what was changed in each iteration, maintaining a history of previous modifications for context, and possibly adopting an agent-based approach with tools such as web search. It 18a salso suggested to consider whether the formal model should evolve in sync with changes to the code

across iterations, and to implement a mechanism to detect and address situations where the number of passed test cases decreases in later iterations.

Potential Use and Value:

The respondent would “possibly” use GALICIA for real-world cases and recognized that the platform shows promise for applications requiring formal verification, pending further validation. Compared to traditional verification tools, GALICIA offers a more streamlined and intuitive interface, although improvements may be needed to match the depth of specialized tools.

Suggestions for Improving GALICIA:

The respondent provided several suggestions to improve GALICIA, including: enabling conversational clarification before code generation, interactive control of correctness, support for user-defined test scenarios, integration with version control systems, and a built-in environment to run and test the generated code—facilitating iterative improvements based on real-time feedback.

Engagement and Contribution:

GALICIA is considered more streamlined and intuitive than traditional verification tools. The respondent is interested in participating in future testing.

Open Feedback:

The respondent noted that GALICIA is a promising tool for automatic code generation and formal model validation, and would be even more effective with clearer explanations of errors and changes between iterations, as well as an integrated environment for testing the generated code. These improvements, according to the respondent, could enhance transparency and help users refine their solutions more efficiently.

Final UX Score for Jerin = 92.6%

Nitish:

Functionality and Performance of the Platform:

The example function passed all verification checks and effectively handled edge cases. The verification/validation feedback was clear and reliable. GALICIA’s feedback was clear and aligned with expected results, although errors were identified in GALICIA’s output. The initial omission of input validation was identified and corrected in a subsequent iteration, and the provided code and its formal model were transparent and well documented. Overall, the respondent expressed strong confidence in GALICIA’s output. However, the system did not clearly communicate the coding language and formal method used.

Workflow and Transparency:

The respondent understood the logic of GALICIA's pipeline and found the overall process to be clear but suggested that additional documentation on the verification mechanisms would be helpful. Including detailed explanations for each verification step would improve understanding. The respondent also recommended adding features such as one-click code copying, an explanation of the formal model, and the ability to switch models. Additionally, suggestions were made to include a history of previous inputs and to enhance model functionality through web search and agent-based capabilities.

Potential Use and Value:

The respondent would "possibly" use GALICIA for real-world cases and acknowledged that the platform shows promise for applications requiring formal verification, pending further validation. Compared to traditional verification tools, GALICIA offers a more streamlined and intuitive interface, though improvements may be needed to match the depth of specialized tools.

Suggestions for Improving GALICIA:

The respondent proposed several enhancements to GALICIA, including the ability to select models directly within the source code generation tab, support for user-defined edge cases during testing, and more detailed explanations of the approach used to generate test cases.

Engagement and Contribution:

GALICIA is considered more streamlined and intuitive than traditional verification tools. The respondent is interested in participating in future testing.

Open Feedback:

The respondent noted that GALICIA is promising for code generation and formal verification, offering an intuitive interface. Suggestions included clearer communication of the coding languages used, detailed explanations of the verification steps, and features such as one-click code copying, model switching, and custom edge case testing to improve transparency and usability.

Final UX Score for Nitish= 89.7%

Jean-Christophe:

Functionality and Performance of the Platform:

The example function passed all verification checks and effectively handled edge cases. The verification/validation feedback was clear and reliable. GALICIA's feedback was clear and aligned with expected results, and no errors were identified in GALICIA's output. The initial omission of input

validation was identified and corrected in a subsequent iteration, and the system clearly communicated the coding language and formal method used. The provided code and its formal model were transparent and well documented. Overall, the respondent expressed strong confidence in GALICIA's output.

Workflow and Transparency:

The respondent understood the logic of GALICIA's pipeline and found the overall process to be clear. Additional documentation on the verification mechanisms was not deemed necessary. However, the respondent felt that including detailed explanations for each verification step would improve understanding and suggested the addition of inline hints or guidance.

Potential Use and Value:

The respondent would "possibly" use GALICIA for real-world cases and recognized that the platform shows promise for applications requiring formal verification, pending further validation. Compared to traditional verification tools, GALICIA offers a more streamlined and intuitive interface.

Engagement and Contribution:

GALICIA is considered more streamlined and intuitive than traditional verification tools. The respondent is interested in participating in future testing.

Open Feedback:

The respondent suggested adding a Haversine distance filter in the SQL query or applying it after the query (e.g., PHP filtering for distance < $radius).

Final UX Score for Jean-Christophe = 89.7

Evaluation Summary

The evaluation of the GALICIA tool highlighted several strengths, along with areas that require improvement and clear directions for future development. Overall, the platform was appreciated for its clarity in communicating the programming language and the formal verification methods it uses. Many evaluators noted that the code and its formal models were transparent and, in most cases, well documented. GALICIA's interface was frequently described as intuitive and more user-friendly than traditional verification tools, and the system showed an ability to adapt—for instance, by correcting missing input validation in subsequent iterations.

Moreover, GALICIA demonstrated strong potential for use in formal verification tasks, especially if further validation is conducted. The interest expressed by many respondents in participating in

future tests also reflects a positive level of engagement and confidence in the platform's development potential.

However, several weaknesses emerged from the feedback. While some users found GALICIA's output reliable, others experienced logical errors or unclear verification feedback, which reduced overall trust in the results. A recurring issue was the lack of detailed documentation and explanations for the verification steps, which made it harder to fully understand how the tool operates. Additionally, the platform's workflow was seen as somewhat rigid, limiting the flexibility needed for iterative or exploratory development processes. Users also pointed out the absence of contextual help or in-platform guidance, making the tool less accessible to newcomers.

To address these concerns, the panel suggested a number of directions for future improvement. These include enabling runtime execution of generated code and tests—especially important for interpreted languages like Python—and providing clearer documentation and detailed explanations of verification steps. Making the workflow more flexible, for example by supporting conversational refinements of user input or allowing partial specifications, was also recommended. The addition of contextual help, short tutorial videos, and the ability to track changes and test regressions across iterations would significantly enhance usability. Finally, features such as user-defined test cases, the ability to switch formal models, and integration with common development tools and environments were also proposed to support real-world usage.

In summary, GALICIA is seen as a promising platform for formal verification and automated code validation, but realizing its full potential will require a concerted effort to improve transparency, usability, and integration with standard development practices.

5. KEY TAKE-UPS AND CONCLUSIONS

1. **Limited Flexibility in Application Development**

GALICIA's current workflow can be overly rigid, offering limited support for the exploratory, conversational process typical of LLM-assisted software development. Developers often rely on iterative prompts to refine both models and implementations. GALICIA should evolve to support modular prototyping, partial specifications, and dialogical co-design via LLM guidance.

This comment has already been considered by adding the possibility for Galicia to ask for clarifications when the initial prompt is not clear or complete.

In the follow up actions that will be performed after the end of the project to refine the Galicia prototype, the possibility for the user to add additional comments and information to the initial prompt will be implemented,

2. **Absence of Runtime Help and Experiment Guidance**

Users currently receive little support while navigating the platform. Adding contextual help, guided experiment modes, would greatly enhance usability and support self-directed experimentation.

In the follow up actions that will be performed after the end of the project to refine the Galicia prototype, we will consider adding contextual help and short video explanations on the web site.

3. **"The generated code was formally correct but behaved unexpectedly due to a logical error."**

In the follow up actions that will be performed after the end of the project to refine the Galicia prototype, we plan to introduce, at least for some interpreted languages like Python, the possibility to run test cases and generated source code, in order to better verify if the tests are passed,

compared against current approach which involves a static analysis of the source code performed by the LLM.

The feedback collected on the GALICIA platform highlights significant potential, along with a clear need for targeted improvements in reliability and usability. Most respondents appreciated the clarity of the programming language and formal methods used, as well as the transparency of the generated code. However, issues such as limited documentation, initial omissions of input validation, and occasional inconsistencies in output were noted.

While the verification workflow was generally understood, many users requested more detailed explanations and the ability to modify the generated code. Opinions on the user interface and real-world applicability were mixed: some considered GALICIA ready for practical use, while others emphasized the need for further development—particularly integration with existing development environments.

Overall, GALICIA is seen as a promising platform for formal code verification with strong growth potential. To fully realize its capabilities, key recommendations include improving documentation, aligning with established standards, and enabling more flexible interactions beyond single-question limitations. The expressed interest in future testing reflects an engaged user base willing to support the platform's evolution.

PROJECT CONSORTIUM:

Funded by
the European Union

Funded by the **European Union**. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. Funded within the framework of the **NGI Sargasso** project under grant agreement No **101092887**