April 2025

Alberto Stefanini (Novareckon)

Lorenzo Vandoni (HAL Service)

KPI 10 – Platform Evaluation Results

Funded by the European Union

NGI SARGASSO

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 10 – Platform Evaluation Results
Authors: Alberto Stefanini (Novareckon), Lorenzo Vandoni (HAL Service)
Revised by: Christian Violi (Novareckon)
Date: 17/04/2025
Version: 1
Distribution: Public Report

TABLE OF CONTENTS

Funded by the European Union

NGI SARGASSO

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**

SUMMARY

This report presents a comprehensive evaluation of the GALICIA platform, focusing on its usability and effectiveness based on feedback from six users. The evaluation highlights key strengths and weaknesses of the platform, including its promising functionality but also outlining some limitations. The users emphasized the importance of incorporating benchmark datasets and case studies to assess the platform's performance on a wider variety of tasks. The feedback underscores that GALICIA will require further development over a longer timescale to become fully functional. Users expressed a general sense of optimism for the platform's future, if additional refinement to enhance its usability, consistency, and performance will be performed. GALICIA's potential as a tool for code generation and validation remains promising, provided that future iterations can address the outlined concerns. A strategic focus on improving user guidance, platform consistency, and code validation is critical for its evolution.

INTRODUCTION

The GALICIA platform, aimed at generating and validating code in various programming languages using large language models (LLMs), has undergone usability evaluations by six users. This report summarizes their feedback, highlighting both the strengths and challenges of the platform. While the platform shows promise, several issues were identified. This report provides a detailed analysis of these concerns and offers recommendations for improving the platform. The evaluation suggests that with further development the platform could become a valuable tool for automated code generation and validation.

1. METHODOLOGY

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.

The core of the evaluation methodology relies on the **System Usability Scale (SUS)**, a well-established standard questionnaire developed in the late 1990s. SUS provides a quick and simple means to assess the usability of a system through a small set of questions that can be easily administered and interpreted. Its concise format is a significant advantage, allowing evaluators to complete the task without excessive burden, which is crucial in situations where participation is voluntary.

In determining the size of the evaluation panel, reference is made to **Jakob Nielsen's heuristic** that suggests 5 to 8 users are sufficient for uncovering the majority of usability issues[1]. According to Nielsen, expanding the panel beyond this size often yields diminishing returns and may even complicate the interpretation of results—*tot capita, tot sententiae*: as the number of evaluators increases, so do the opinions, potentially diffusing the clarity of insights.

Nevertheless, several **limitations** must be acknowledged. One inherent drawback of SUS is its focus on usability alone; it does not explicitly capture users' perceptions of functional limitations or

[1] See: Landauer & Nielsen: Why You Only Need to Test with 5 Users).

Funded by the European Union

NGI SARGASSO

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**

strategic shortcomings of the platform itself. As such, while it can indicate whether the platform is easy to use, it may fail to provide insight into whether the platform meets broader user needs or strategic goals.

Furthermore, the evaluation process can be **cumbersome**, especially in contexts where participation is unpaid and voluntary. Responses cannot be guaranteed, and the incentive to contribute useful feedback is often low. This practical constraint underscores the challenges of relying on external evaluators without compensation. The phrase *"beggars can't be choosers"* aptly captures the difficulty of demanding high-quality, comprehensive evaluations under constrained conditions.

In summary, while the chosen methodology for platform evaluation is grounded in established practices and aims for methodological rigor, it remains constrained by practical considerations and the limited scope of usability-focused tools. Despite these challenges, such an approach can still yield valuable insights, particularly when executed with care and in conjunction with other complementary evaluation strategies.

Despite these limitations, for the purposes of this project, this methodology has proven to be valid, as several evaluators involved willingly accepted to fill in the questionnaire.

2. SUS SCORE COMPUTATION

The evaluation of platform usability was conducted using the **System Usability Scale (SUS)**, a widely recognized and validated instrument designed to assess the perceived usability of interactive systems. The SUS consists of **ten standard statements**, each evaluated by users on a **five-point Likert scale** ranging from *Strongly Disagree (1)* to *Strongly Agree (5)*.

First, we include the SUS form itself, which should be taken into account when elaborating the computation methodology:

Questions	**Strongly Disagree**	**Disagree**	**Can't say**	**Agree**	**Strongly Agree**	**Comments**	
I think that I would like to use this product frequently						*Whenever your evaluation goes below average, state briefly why.*	
I found the product unnecessarily complex						*Whenever your evaluation goes above average, state briefly why*	

Funded by the European Union

NGI SARGASSO

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**

I thought this product was easy to use						*Whenever your evaluation goes below average, state briefly why.*	
I think that I would need the support of a technical person to be able to use this product.						*Whenever your evaluation goes above average, state briefly why*	
I thought there was too much inconsistency in this product.						*Whenever your evaluation goes below average, state briefly why.*	
I would imagine that most people would learn to use this product very quickly.						*Whenever your evaluation goes below average, state briefly why.*	
I found this product very awkward to use.						*Whenever your evaluation goes above average, state briefly why*	
I felt very confident using this product						*Whenever your evaluation goes below average,*	

Funded by the European Union

NGI SARGASSO

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**

						state briefly why.	
I needed to learn a lot of things before I could get going with this product.						*Whenever your evaluation goes above average, state briefly why.*	

Computation Methodology

The SUS scoring method involves the following steps:

1. **Numerical Conversion of Responses**
   Each user response is first converted into a numerical value from 1 to 5, corresponding to their level of agreement.

2. **Adjustment Based on Item Polarity**
   To normalize the responses:

   - For **positively worded items** (items 1, 3, 5, 7, 9), the adjusted score is calculated as:

Adjusted score=Response−1\text{Adjusted score} = \text{Response} - 1

   - For **negatively worded items** (items 2, 4, 6, 8, 10), the adjusted score is calculated as:

Adjusted score=5−Response\text{Adjusted score} = 5 - \text{Response}

3. **Summation of Adjusted Scores**
   The adjusted scores for all ten items are summed. The resulting raw score ranges from 0 to 40.

4. **Scaling to SUS Metric**
   To convert the raw score to a standard SUS score on a 0–100 scale, the following formula is applied:

$$S = 2.5 \times \sum_{i=1}^{10} s_i$$

$$s_i = \begin{cases} q_i - 1 & \text{if } i \text{ is odd (positive)} \\ 5 - q_i & \text{if } i \text{ is even (negative)} \end{cases}$$

Figure 1 - Computation of a SUS score

Funded by the European Union

NGI SARGASSO

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**

Interpretation of Results

The SUS score provides a quantitative measure of overall usability. While it is not diagnostic, it enables benchmarking across systems and over time. Common interpretive guidelines are as follows:

- **SUS ≥ 68** is generally interpreted as **above average** usability.
- **SUS ≥ 80** indicates **high usability and user satisfaction**.
- **SUS ≤ 50** may suggest the presence of **significant usability issues**.

Automation Considerations

The computation process is sufficiently structured to allow for **automated analysis** using spreadsheets or scripts. When responses are collected digitally, the adjusted scores and final SUS result can be computed in real time, facilitating timely and consistent evaluations across test participants.

3. CONSULTATION PROCESS AND RESPONDENT OVERVIEW

In order to ensure an informed and independent assessment of the platform's usability, a targeted consultation was carried out involving professionals with relevant backgrounds in user experience (UX), human-machine interaction, and usability evaluation. Invitations were extended to a selected group of individuals, including attendees of the workshop held on **March 4th**, 2025, as well as recognized usability specialists and domain experts.

Participant Selection Criteria

The selection of participants for the usability evaluation was guided by the following key considerations:

- **Externality and Independence**: Wherever possible, individuals were chosen who had no direct involvement in the platform's development. This was intended to minimize bias and ensure that feedback reflected a fresh, independent perspective on user experience.
- **Domain Diversity**: Respondents were drawn from various sectors, including academia, applied research, cybersecurity, industrial automation, and digital services. This helped to capture a broad spectrum of use cases, expectations, and usability assumptions.
- **Proven UX Expertise**: A subset of the invitees consisted of recognized experts in UX, HCI (Human-Computer Interaction), and software usability — individuals who could provide feedback not only as end-users but also from a methodological and design-oriented viewpoint.
- **Prior Exposure to the Platform**: While most participants had previously encountered the platform (e.g., during the March 4 workshop), care was taken to ensure that familiarity did not translate into over-familiarity, allowing room for candid and spontaneous observations.
- **Willingness to Engage with the SUS Format**: Finally, the selection prioritized individuals likely to provide usable responses to the System Usability Scale (SUS), a method requiring not only judgments but also brief justifications, especially for outlier scores.

Funded by the European Union

NGI SARGASSO

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**

The consultation design reflects a pragmatic balance: privileging methodological rigor while accommodating the voluntary nature of participation. The resulting panel of respondents represents a well-calibrated sample aligned with established usability testing guidelines.

Invited Experts

The following individuals were invited to participate in the SUS-based usability assessment:

- **Alessandro Gallina** (HAL Service)
  ICT professional with experience in digital platforms for industrial and mobility applications.
  LinkedIn
- **Franco D'Urso** (Emisfera)
  Senior developer and technical manager, with extensive experience in system integration and front-end development.
  LinkedIn
- **Alberta Bertin** (Novareckon)
  Project manager in innovation and dissemination, with strong expertise in stakeholder engagement.
  LinkedIn
- **Giovanni Guida** (Consultant, Emeritus Professor, University of Brescia)
  Recognized expert in UX design, human-machine interaction, and software ergonomics.
  LinkedIn
- **Franco Alberto Cardillo** (CNR ILC)
  Researcher with focus on digital lexicons, accessibility, and user-centered interface development.
  LinkedIn
- **Felicita Di Giandomenico** (CNR ISTI)
  Senior researcher in dependability, resilience engineering, and safety-critical systems.
  LinkedIn
- **Fernando Garcia Gutierrez** (EDP)
  Technology officer and digital transformation leader with a focus on energy platforms and sustainability.
  LinkedIn
- **Sandro Bologna** (Former ENEA, Founder of AIIC – Italian Association of Critical Infrastructures)
  Leading authority in critical infrastructure protection, cyber-physical systems, and risk analysis.
  LinkedIn
- **Elenio Dursi** (iControl / CLUSIT)
  Cybersecurity expert and UX consultant, involved in various public-private initiatives on ICT security.
  LinkedIn
- **Alberto Servida** (University of Genoa, ANIPLA)
  Academic and industrial automation expert, engaged in standardization and industrial

Funded by the European Union
NGI SARGASSO
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**

process design.
LinkedIn

- **Serge Demeyer** (University of Antwerp)
  Professor and recognized specialist in software evolution and user interface evaluation methodologies.
  LinkedIn

Responses and Follow-up

At the time of writing, valuable and detailed responses to the SUS-based evaluation were received from the following participants:

- **Alessandro Gallina**
- **Franco D'Urso**
- **Elenio Dursi**
- **Giovanni Guida**
- **Franco Alberto Cardillo**

Additional responses may still be forthcoming:

- **Francesca Lonetti** (on behalf of Felicita Di Giandomenico) has expressed interest and may submit feedback shortly.
- **Serge Demeyer** has indicated that he may provide his assessment in late May, once a revised version of the platform is made available.

All other invitees either formally declined or did not respond by the established deadline. Nevertheless, the feedback received so far already meets the suggested minimum threshold for usability testing as outlined in Nielsen's heuristic of **5 to 8 participants**, which is considered sufficient for identifying the majority of usability issues.

4. REPLY ANALYSIS

Alessandro Gallina Evaluation Summary

Alessandro Gallina's responses indicate a generally **positive perception** of the platform's usability, with a final SUS score of **85**, which is considered excellent. This suggests that the platform is viewed as highly usable, with most responses in the "Agree" and "Strongly Agree" range.

Key points:

- Gallina expressed **confidence** in using the platform (Q8) and found it **easy to use** (Q3), with no need for **technical support** (Q4).
- He agreed that users would **quickly learn to use the platform** (Q6) and did not find it **awkward** (Q7).
- However, he did not provide a definitive opinion on whether he would use the product frequently (Q1), indicating **uncertainty** due to insufficient experience.
- In general, the responses were highly positive, but his comments in Q1 suggest a need for **further exploration** before making a firm judgment on repeated use.

Funded by the European Union

NGI SARGASSO

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**

Franco D'Urso Evaluation Summary

Franco D'Urso's responses reflect a **positive assessment** of the platform, with a final SUS score of **82.5**, suggesting a **strong usability rating**. Most responses fall into the "Agree" or "Strongly Agree" categories, showing general satisfaction.

Key points:

- D'Urso found the platform to be **easy to use** (Q3) and expressed **confidence** in using it (Q8).
- He also agreed that **technical support** wouldn't be necessary (Q4) and believed users would **learn quickly** (Q6).
- A slight reservation is noted in his response to Q5 ("too much inconsistency"), suggesting a potential area for improvement regarding **consistency** across the platform's features.
- D'Urso did not explicitly comment on the frequency of use (Q1), but overall, the feedback remains **enthusiastic** with minor reservations about the platform's internal cohesion.

Giovanni Guida Evaluation Summary

Giovanni Guida's evaluation provides a **mixed** view of the platform, with a final SUS score of **80**, which is acceptable but reflects some **concerns**.

Key points:

- While Guida did not find the platform **unnecessarily complex** (Q2) and found it **easy to use** (Q3), he expressed significant concerns regarding the **lack of consistency** across the platform (Q5). This issue could have a considerable impact on the overall user experience and should be addressed.
- Guida also mentioned the **awkwardness** of the platform (Q7), which suggests that certain aspects of the user interface may not be as intuitive or smooth as expected.
- Although he felt **confident** in using the platform (Q8), the response to Q1 ("Can't say") indicates **uncertainty** about its frequent use, which could point to a lack of **engagement** or comfort with the platform after initial exposure.
- The **absence of any positive feedback** about the platform's **integration of functions** or **ease of use in frequent scenarios** signals that the platform may not fully meet the needs of users seeking consistency and smooth interaction.

Elenio Dursi Evaluation Summary

Elenio Dursi's evaluation presents a moderate usability score of 70, highlighting a few key concerns that could affect the platform's overall user experience.

Key points:

- **Inconsistency**: Dursi did not respond to the question about the consistency of the platform (Q5), suggesting a lack of clarity or difficulty in assessing this aspect. This may indicate underlying issues that need to be addressed for a more cohesive user experience.
- **Awkwardness of use**: His response to Q7, "Disagree" with the product being awkward, raises concerns about the platform's ease of use. While not a strong negative, it points to areas where the interface could be smoother and more intuitive.

Funded by the European Union

NGI SARGASSO

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**

- **Learning curve**: Although Dursi felt confident using the platform (Q8), his response to Q9 ("Disagree" with the need for extensive learning) suggests that some users may find the platform somewhat challenging to navigate initially, even though it doesn't require heavy technical support.
- **Limited engagement**: The "Can't say" responses for Q5 and Q10, which ask about the integration of functions and the overall system's use, signal a lack of clear positive impressions regarding the platform's functionality and usability. This absence of strong positive feedback may point to underlying issues with the platform's appeal or consistency.

Additionally, Dursi provided a useful suggestion mentioning that the platform should interface **LLM GPT-3**, which he identifies as the main LLM created by OpenAI for programming. This suggestion has already been considered and the o3-mini LLM (he wrote "Gpt-3" but of course he meant "o3", because Gpt-3 is very old and deprecated) has already been added among the LLM models that can be used in GALICIA.

Sandro Bologna Evaluation Summary

Sandro Bologna's evaluation provides a mixed view of the platform, with a final SUS score of 80, suggesting a generally acceptable usability but highlighting several concerns.

Key points:

- **Uncertainty about result validity**: Bologna's response to Q1 reflects significant doubts about the platform's ability to **validate the suggested code**, a crucial aspect of usability. His inability to check the validity of the results raises concerns about the reliability of the platform, which could be a major limitation for users expecting consistent, verifiable outputs.
- **Potential complexity**: While Bologna did not find the platform unnecessarily complex (Q2), the underlying implication is that **users may struggle to navigate the platform's features effectively**, particularly if they lack expertise in the application domain. This suggests that the platform may not be intuitive enough for a broader user base.
- **Neutral ease of use**: Bologna's response to Q3 ("Can't say") suggests **indifference towards the platform's ease of use**, which indicates that the user interface may not be as intuitive or straightforward as needed. This neutrality points to a lack of engagement or satisfaction with the user experience, suggesting room for improvement.
- **Expert-level user requirement**: Bologna's feedback implies that the platform is primarily suited for **expert users who can understand and check the proposed code**. This highlights a limitation, as the platform's usability may be reduced for those without specialized knowledge in the domain, making it less accessible to a broader audience.
- **Inconsistency**: Bologna's response to Q5 points out **inconsistency in the platform**, which remains a critical issue. While not overly emphasized, the presence of such inconsistencies could undermine the overall user experience, especially for those looking for a more stable and predictable system.
- **Limited confidence and ease of use**: Although Bologna responded positively to Q7 and Q8, noting that the platform is easy to use and that he felt confident using it, these responses still come from a domain expert perspective. **For less experienced users, the platform may not provide the same level of confidence or ease of use**, further emphasizing its limitations for a wider audience.

Funded by the European Union

NGI SARGASSO

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**

- **Learning curve**: The feedback to Q6 suggests that **users may face challenges when first learning the platform**, even though Bologna believes that experts would quickly adapt. This is a negative point as it implies the platform has a higher learning curve than ideal for more casual or first-time users.

Bologna's additional comment about requiring expert-level knowledge and the ability to check code validity highlights a possible concern: **the platform might be inaccessible or difficult to use for users without programming expertise**. This could limit its appeal and usability in broader contexts.

Franco Alberto Cardillo Evaluation Summary

Franco Cardillo's evaluation highlights significant concerns, particularly regarding the platform's usability and the validity of its code generation process.

Key points:

- **Interface Feedback**: While Cardillo acknowledges that the interface is **well-designed and pleasant**, this praise is limited and vague. He suggests **minor improvements** but does not elaborate, implying that the interface may not yet be polished or intuitive enough for users, though the specific issues remain unclear.
- **Lack of Understanding of the Platform's Workflow**: Cardillo's main concern lies in his **inability to understand the underlying pipeline** of the platform. His confusion about how the system ensures the correctness of generated code, especially in formal specifications, indicates a **lack of clarity in explaining the platform's functionality**. This is exacerbated by a prior workshop example where the system's code generation and validation failed. The absence of clear documentation or transparency about the system's validation process may leave users uncertain about the reliability of the platform.
- **Code Generation Failure**: A minor flaw in Cardillo's experience is the **incorrect generation of code**. When selecting Python among options and then requesting Java code, the system generated Python code without notifying the user of the contradiction in his request. This minor issue will be fixed before the project terminates.
- **Suggested Improvements**: While Cardillo does not explicitly suggest many improvements, his experience indicates a **need for better error reporting, clearer feedback on language selection**, and **more transparency about the validity of generated code**. He also suggests experimenting with benchmark datasets to test the platform on more complex tasks, which would allow for better insights into its real-world capabilities.

This evaluation highlights the **need for better user guidance, error handling, and transparency** before the platform can be considered reliable and user-friendly enough for being delivered to a larger public.

Overall Assessment of GALICIA's Usability

Based on the six evaluations provided, the overall usability of the GALICIA platform is characterized by mixed feedback, highlighting both positive aspects and substantial areas for improvement.

Key strengths identified include:

- **Ease of use**: While not universally praised, some users, such as Sandro Bologna, found the platform to be relatively easy to use once they became familiar with it. The interface,

Funded by the European Union

NGI SARGASSO

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**

described as **pleasant and well-designed** by Franco Cardillo, also received some positive comments.

- **Intuitive interface**: Giovanni Guida and Elenio Dursi did not find the platform to be overly complex, and some acknowledged that **basic functions** could be easily accessed, suggesting the potential for **further refinement**.

However, the **negative feedback** are also worth being considered:

- **Inconsistent behavior**: A recurring theme across multiple evaluations is the **lack of consistency** in the platform's behavior, particularly in terms of how it generates code and handles requests. This inconsistency, pointed out by both Giovanni Guida and Franco Cardillo, suggests that the systemcould be improved. For example, the system generated Python code when Pythpn was selected in the options but Java was requested in the prompt, without notifying the user of the contradiction in his request.

- **Lack of transparency in the platform's process**: Some users, including Franco Cardillo and Sandro Bologna, voiced confusion about how the platform validates and generates code. This lack of **clarity and transparency** in the system's underlying processes could lead to a **loss of trust** among users, as it remains unclear how the platform guarantees the correctness of its generated code. This is an important feedback, and will be considered in the last month of the project.

- **User doubts and the need for expertise**: Some users, such as Sandro Bologna and Elenio Dursi, expressed doubts about whether **non-expert users** could fully trust the platform without additional technical support or programming knowledge. This points to a potential barrier to wider adoption, as the platform might not be intuitive or accessible enough for general users, but is not considered a big problem, because the platform is directed to programmers

Evaluation Summary

GALICIA's usability has received good feedback in terms of ease of use and intuitive interface, but some improvements are needed in its code generation accuracy, and clarity of feedback. Some users seem to struggle with trusting the system's outputs, and understanding how the generated source code can be considered correct.

Recommendations:

- Improve **transparency** around the system's validation process and error handling to boost user confidence.

- **Broaden testing** to include a wider variety of users and use cases, to assess whether the platform is truly user-friendly for non-experts.

5. KEY TAKE-UPS AND CONCLUSIONS

The feedback from the usability evaluations of the GALICIA platform reveals some insights into its current state and its potential future development. While the system has demonstrated promising features, the overall usability and reliability of the platform could be improved in some areas. This highlights the **long-term nature of the technology development process** required for GALICIA to reach its full potential.

Funded by the European Union

NGI SARGASSO

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**

1. **Galicia Objectives Cannot Be Fully Achieved**. The feedback suggests that while the GALICIA platform shows promise, the current state of the platform prototype does not yet allow for the full achievement of its objectives. As such, it is clear that achieving the vision behind GALICIA will require more time and development, extending well beyond the initial 10-month timeframe. The feedback indicates that a further year of improvement and testing would likely be necessary for the platform to mature sufficiently and meet the needs of its user base.

2. **Analysis of Replies Needs to Be Manual and Careful**. Another key point raised by the evaluations is the complexity of the tasks at hand, specifically related to code validation and generation. Several users expressed doubts about the platform's ability to handle more complex use cases, particularly when the correctness of the generated code is not easily discernible. This reinforces the idea that, for now, the replies generated by GALICIA cannot be fully automated and must be subject to careful, manual review and analysis by experts. Until the system is capable of guaranteeing code quality consistently and transparently, human oversight will remain essential, particularly for complex or high-risk scenarios.

3. **Benchmarking and Case Studies Are Highly Valuable**. The feedback also strongly points to the importance of incorporating reference benchmarks and case studies into the development of GALICIA. Users such as Franco Cardillo and Sandro Bologna raised concerns about the platform's ability to handle diverse, real-world programming tasks, and suggested that GALICIA be tested against well-established benchmark datasets used in academic research. By doing so, it would be possible to assess how well the platform performs on a variety of tasks, including those involving more complex code generation and validation scenarios. Case studies, in particular, could provide valuable context for understanding the platform's strengths and limitations, enabling further targeted improvements.

4. **The Platform May Prove Highly Useful**, but a longer timescale is required even if the feedback reflects optimism about the platform's potential. With additional refinement and time, GALICIA may become a highly useful tool for generating and validating code across various programming languages. However, the platform's full potential will only be realized over a longer development period. Based on the user feedback, it is likely that an additional year of development, testing, and fine-tuning will be necessary to address the existing usability gaps, enhance consistency, and guarantee the quality of generated code. This extended timeline will also allow for more comprehensive testing with real-world data and will ensure that the platform meets the needs of its intended users.

In conclusion, the GALICIA platform shows promise, but as indicated by the evaluations, it will require further time, resources, and refinement to meet its objectives. A strategic focus on usability, reliable code generation, and clearer user feedback is essential in the coming months, alongside rigorous testing and benchmarking. With these efforts, GALICIA has the potential to evolve into a highly valuable tool, but its full utility will only be realized on a longer timescale. The work that we intend to do in the last months of the project will be divided into three main aspects:

- first, we will consider the feedback received to introduce several improvements to the platform, which will certainly be more stable and reliable than the current prototype at the end of the project;
- secondly, several tests will be carried out, both by the project team and by some external users, invited to use the prototype, which is publicly available;

Funded by the European Union

NGI SARGASSO

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**

- finally, we will elaborate statistics related to all the tests carried out.

The ultimate goal of these statistics is to **verify whether the code produced by Galicia is actually better than that produced by the LLM** to which the same request is made directly, without using our platform. The comparison will be based on the number of test cases passed.

Funded by the European Union

NGI SARGASSO

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 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**