

# Bridging LLMs and Symbolic Reasoning in Educational QA Systems: Insights from the XAI Challenge at IJCNN 2025

Long S. T. Nguyen<sup>1,†</sup>, Khang H. N. Vo<sup>1,†</sup>, Thu H. A. Nguyen<sup>1</sup>, Tuan C. Bui<sup>1</sup>, Duc Q. Nguyen<sup>1</sup>, Thanh-Tung Tran<sup>2</sup>, Anh D. Nguyen<sup>3</sup>, Minh L. Nguyen<sup>4</sup>, Fabien Baldacci<sup>5</sup>, Thang H. Bui<sup>1</sup>, Emanuel Di Nardo<sup>6</sup>, Angelo Ciaramella<sup>6</sup>, Son H. Le<sup>7</sup>, Ihsan Ullah<sup>8</sup>, Lorenzo Di Rocco<sup>9</sup> and Tho T. Quan<sup>1,\*</sup>

<sup>1</sup>URA Research Group, Ho Chi Minh City University of Technology (HCMUT), Vietnam

<sup>2</sup>Ho Chi Minh City International University (HCMIU), Vietnam

<sup>3</sup>University of South-Eastern Norway, Norway

<sup>4</sup>Japan Advanced Institute of Science and Technology (JAIST), Japan

<sup>5</sup>Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

<sup>6</sup>University of Naples Parthenope, Italy

<sup>7</sup>VNU Information Technology Institute, Vietnam National University, Vietnam

<sup>8</sup>Visual Intelligence Lab, School of Computer Science & Insight Center for Data Analytics, University of Galway, Ireland

<sup>9</sup>Sapienza University of Rome, Italy

## Abstract

The growing integration of Artificial Intelligence (AI) into education has intensified the need for transparency and interpretability. While hackathons have long served as agile environments for rapid AI prototyping, few have directly addressed eXplainable AI (XAI) in real-world educational contexts. This paper presents a comprehensive analysis of the XAI Challenge 2025, a hackathon-style competition jointly organized by Ho Chi Minh City University of Technology (HCMUT) and the International Workshop on Trustworthiness and Reliability in Neurosymbolic AI (TRNS-AI), held as part of the International Joint Conference on Neural Networks (IJCNN 2025). The challenge tasked participants with building Question-Answering (QA) systems capable of answering student queries about university policies while generating clear, logic-based natural language explanations. To promote transparency and trustworthiness, solutions were required to use lightweight Large Language Models (LLMs) or hybrid LLM-symbolic systems. A high-quality dataset was provided, constructed via logic-based templates with Z3 validation and refined through expert student review to ensure alignment with real-world academic scenarios. We describe the challenge's motivation, structure, dataset construction, and evaluation protocol. Situating the competition within the broader evolution of AI hackathons, we argue that it represents a novel effort to bridge LLMs and symbolic reasoning in service of explainability. Our findings offer actionable insights for future XAI-centered educational systems and competitive research initiatives.

## Keywords

Explainable Question Answering in Education, Logic-Based Natural Language Explanation, Neuro-Symbolic Reasoning Systems, Lightweight Large Language Models, Hackathon-style AI Challenge

## 1. Introduction

Hackathons emerged in the late 1990s as intensive, time-constrained events where developers collaborated to rapidly prototype functional solutions [1]. Initially focused on general-purpose programming, these events gradually evolved into innovation incubators across diverse domains. By the early 2010s,

*ITADATA2025: The 4<sup>th</sup> Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy*

\*Corresponding author.

<sup>†</sup>These authors contributed equally.

✉ long.nguyencse2023@hcmut.edu.vn (L. S. T. Nguyen); khang.vo872003@hcmut.edu.vn (K. H. N. Vo); thu.nguyenhoanganh14@hcmut.edu.vn (T. H. A. Nguyen); tuanbc88@hcmut.edu.vn (T. C. Bui); nqduc@hcmut.edu.vn (D. Q. Nguyen); ttung@hcmiu.edu.vn (T. Tran); anh.nguyen.duc@usn.no (A. D. Nguyen); nguyenml@jaist.ac.jp (M. L. Nguyen); fabien.baldacci@labri.fr (F. Baldacci); bhthang@hcmut.edu.vn (T. H. Bui); emanuel.dinardo@uniparthenope.it (E. D. Nardo); angelo.ciaramella@uniparthenope.it (A. Ciaramella); sonlh@vnu.edu.vn (S. H. Le); ihsan.ullah@universityofgalway.ie (I. Ullah); lorenzo.dirocco@uniroma1.it (L. D. Rocco); qttho@hcmut.edu.vn (T. T. Quan)  
ID 0009-0008-7488-4714 (L. S. T. Nguyen); 0009-0009-7631-2450 (K. H. N. Vo); 0000-0003-0467-6254 (T. T. Quan)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

hackathons had become increasingly integrated into educational contexts, offering informal yet impactful environments for learners to connect theoretical understanding with real-world application [2]. They fostered creativity, collaboration, and technical fluency, which are essential skills in the rapidly advancing field of *Artificial Intelligence* (AI). The mid-2010s marked a turning point, as breakthroughs in machine learning and the rise of open-source frameworks such as TensorFlow<sup>1</sup> and PyTorch<sup>2</sup> enabled hackathons to address more complex and data-driven problems [3]. In the early 2020s, the emergence of *Large Language Models* (LLMs), including ChatGPT<sup>3</sup> and GitHub Copilot<sup>4</sup>, significantly enhanced participants' productivity, code quality, and learning outcomes. At the same time, these tools raised concerns about academic integrity and over-reliance on automation, highlighting the need for clearer ethical guidelines and responsible deployment [4, 5].

As LLMs became increasingly capable but remained difficult to interpret, *eXplainable AI* (XAI) emerged as a crucial paradigm for promoting transparency. This need is especially pronounced in educational settings, where students and educators often seek justifications for automated decisions [6]. In parallel, symbolic AI has regained attention as a complementary approach to purely data-driven methods. A notable milestone was the 2024 release of AlphaGeometry by Google DeepMind<sup>5</sup>, a neuro-symbolic system that achieved gold medal-level performance at the International Mathematical Olympiad. At its core lies a symbolic reasoning engine, which demonstrates the potential of logic-based inference in addressing complex educational problems [7, 8]. In this broader landscape, AI competitions have begun to incorporate educational and explainability-oriented goals. For example, EfficientQA [9] and the Alexa Prize TaskBot<sup>6</sup> focused on answer accuracy and dialogue, but offered limited support for interpretable outputs. Symbolic reasoning benchmarks such as the *Abstraction and Reasoning Challenge* (ARC)<sup>7</sup> and Neuro-Symbolic ARC [10] emphasized structured reasoning but did not require *natural language* (NL) explanations. Similarly, biomedical competitions such as the MIDRC mRALE Challenge<sup>8</sup> targeted explainability in medical diagnostics, rather than educational domains.

To fill this gap, the XAI Challenge 2025<sup>9</sup> was introduced as part of the *International Joint Conference on Neural Networks* (IJCNN 2025), co-organized by *Ho Chi Minh City University of Technology* (HCMUT) and the *International Workshop on Trustworthiness and Reliability in Neurosymbolic AI* (TRNS-AI). Although the challenge ran for three months, it was inspired by the spirit of hackathons, emphasizing rapid iteration, interdisciplinary collaboration, and practical relevance. The competition was structured into multiple phases, including dataset familiarization, system development, and explanation refinement. Participants were asked to build educational *Question-Answering* (QA) systems that could respond to student queries about academic regulations while providing logically grounded natural language justifications. All solutions were required to use lightweight LLMs or hybrid LLM and symbolic reasoning models, with an emphasis on transparency, verifiability, and alignment with human logic. The challenge was supported by a carefully designed dataset grounded in real university policies. This dataset was developed through a two-stage pipeline: (i) synthetic examples were generated using logic-based templates and validated using the Z3 Solver [11], and (ii) these examples were refined and validated by trained university students to ensure clarity, factual accuracy, and educational relevance.

This paper presents a comprehensive overview of the XAI Challenge 2025, including its motivation, structure, dataset design, evaluation methodology, and participant outcomes. By situating the challenge within the broader evolution of AI competitions, we argue that it represents a new class of explainability-focused systems aligned with educational values. In particular, we emphasize how the challenge brings together the strengths of LLMs and symbolic reasoning, establishing a practical benchmark for trustworthy and pedagogically sound AI in education.

---

<sup>1</sup><https://www.tensorflow.org/>

<sup>2</sup><https://pytorch.org/>

<sup>3</sup><https://openai.com/index/chatgpt/>

<sup>4</sup><https://github.com/features/copilot>

<sup>5</sup><https://deepmind.google/>

<sup>6</sup><https://www.amazon.science/alex-prize/taskbot-challenge>

<sup>7</sup><https://www.kaggle.com/competitions/abstraction-and-reasoning-challenge>

<sup>8</sup><https://www.midrc.org/xai-challenge-2024>

<sup>9</sup><https://sites.google.com/view/trns-ai/challenge>

## 2. Positioning Within the Landscape of AI Competitions

In recent years, AI competitions have proliferated not only as benchmarks for measuring technical progress but also as collaborative platforms for exploring emerging paradigms such as LLMs, symbolic reasoning, and explainability. These events cover a wide range of tasks, including open-domain QA, task-oriented dialogue, program synthesis, and interpretable diagnostics. However, few competitions explicitly integrate educational QA, symbolic reasoning, and explainability within a unified framework. Table 1 provides a comparative overview of the XAI Challenge 2025 in relation to a selection of prominent AI competitions.

**Table 1**

Comparison of XAI Challenge 2025 with other notable AI competitions

Competition	Field	QA-focused	Symbolic Reasoning	NL Explanation	Real-world Domain
EfficientQA [9]	Open-domain QA	☑	✗	✗	General knowledge
Alexa Prize TaskBot <sup>6</sup>	Task-based dialogue	☑	☑	✗	Household tasks
ARC <sup>7</sup> / Neuro-Symbolic ARC [10]	Abstract reasoning	✗	☑	✗	Synthetic puzzles
MIDRC XAI Challenge <sup>8</sup>	Medical image diagnosis	✗	✗	✗	X-ray diagnostics
XAI Hackathon Pisa <sup>10</sup>	XAI prototyping	✗	—	—	Mixed datasets
Explainable Fuzzy AI Challenge <sup>11</sup>	Fuzzy logic systems	✗	☑	✗	Agent control
xAI Challenge Austin <sup>12</sup>	Coding + explanation	—	✗	—	General AI tasks
Our XAI Challenge	Educational QA	☑	☑	☑	Academic policy

**Legend:** ☑= Fully addressed; ✗= Not addressed; — = Partial or optional inclusion

Our challenge introduces a unique combination of educational QA, symbolic reasoning, and logic-based natural language explanation. This combination is rarely emphasized simultaneously in other major competitions. Its requirement to use lightweight LLMs together with symbolic inference makes it particularly suitable for student-facing applications where transparency and trust are essential. In addition, the challenge’s structure encourages interdisciplinary collaboration, rapid prototyping, and accountable system behavior. These qualities are often missing in leaderboard-driven competitions that focus narrowly on accuracy. Taken together, these attributes establish the XAI Challenge 2025 as a novel and practical benchmark for building explainable, trustworthy, and educationally aligned AI systems.

## 3. The XAI Challenge 2025

### 3.1. Motivation and Objectives

The XAI Challenge 2025 was motivated by the limitations of conventional LLM-based QA systems in educational settings. These systems often return concise answers with limited explanatory depth, making it difficult for users to trace errors in reasoning or verify the correctness of responses. This lack of transparency is particularly problematic in rule-based, high-stakes scenarios such as those involving university policies. To address this issue, the competition promoted hybrid approaches that combine the fluency of LLMs with the rigor of symbolic reasoning, with the goal of enhancing both interpretability and trustworthiness in educational QA systems. The primary objectives of the challenge are as follows.

- **O1.** Encourage the development of QA systems that generate logic-grounded natural language explanations for policy-related queries.
- **O2.** Promote hybrid architectures that integrate symbolic inference with LLM-based generation.

<sup>10</sup><http://xai-hackathon.isti.cnr.it/>

<sup>11</sup><https://xfuzzycomp.github.io/XFC/>

<sup>12</sup><https://finch-mauve-rk3e.squarespace.com/xai-atx>

- **O3.** Enhance the transparency and verifiability of automated responses in educational contexts.
- **O4.** Showcase real-world applications where explainability improves student comprehension and learning outcomes.
- **O5.** Recognize outstanding solutions through academic dissemination, including presentations at the TRNS-AI workshop and paper submissions to the 4<sup>th</sup> Italian Conference on Big Data and Data Science (ITADATA 2025).

### 3.2. Structure and Timeline

The event was structured to foster rapid prototyping, interdisciplinary collaboration, and real-world impact. While inspired by the fast-paced spirit of traditional hackathons, the competition unfolded across multiple phases over a three-month period, allowing participants to iteratively refine their solutions. The complete timeline of the XAI Challenge 2025 is summarized in Table 2.

The challenge welcomed a diverse pool of participants, including high school and university students as well as early-career researchers with interests in XAI. Teams of up to six members could register between March 2 and April 25, 2025, with individuals without a team having the option to be matched into groups by the organizers. To support effective system development, a virtual kickoff workshop and dataset release were held on April 13. The main competition phase ran from April 14 to May 11, during which participants developed educational QA systems capable of answering university policy-related questions while generating logically grounded natural language explanations. Evaluation was conducted in two stages: (i) Phase 1 results were announced on May 12–13, followed by a brief model refinement period on May 14–15; and (ii) Phase 2 results were released on May 16–17, with final rankings determined on May 18. On June 1, a public test day was held, during which all systems were evaluated on a hidden test set. Each team also delivered a live presentation of their solution, followed by a Q&A session with a panel of challenge chairs comprising renowned professors. Final rankings and awards were officially announced at the end of the day.

To extend the competition’s impact beyond its runtime, the top three teams were invited to present at the TRNS-AI Workshop (held as part of IJCNN 2025) on July 5 and to submit a full paper to the ITADATA Conference by June 30. These post-challenge activities were designed to encourage continued academic dissemination and foster sustained community engagement.

**Table 2**

Timeline of the XAI Challenge 2025

Date(s)	Event
March 2 – April 25	Team registration period
April 13	Kickoff workshop and dataset release
April 14 – May 11	Main competition phase
May 12–13	Phase 1 evaluation results
May 14–15	Model refinement period
May 16–17	Phase 2 evaluation results
May 18	Final ranking announcement
June 1	Public test day, solution presentations, and final result release
June 30	Paper submission (Top 3 teams, ITADATA Conference)
July 5	Presentation at TRNS-AI Workshop (IJCNN 2025)

### 3.3. Dataset

To align with the educational goals of the challenge, the dataset was carefully constructed to reflect realistic, policy-based questions that students often face in academic settings. Each entry consists of a collection of premises presented in both natural language and *First-Order Logic* (FOL), accompanied by one or more questions and their corresponding ground truth answers. The content covers a wide spectrum of university regulations, including course enrollment criteria, graduation requirements,

```

{
  "premises-NL": [
    "Every student enrolled in the course who completes at least 80% of the assignments passes the course.",
    "If a student attends all lectures, then they have a higher chance of passing the final exam.",

    "If a student attends a tutoring session or completes extra practice problems, they are more likely to improve their grades.",
    "No student who fails to submit their research paper passes the course.",
    "There exists a student who is on academic probation and later graduates with honors."
  ],
  "premises-FOL": [
    "ForAll(x, (Student(x) AND Completed80PctAssignments(x)) -> PassCourse(x))",
    "ForAll(x, AttendsAllLectures(x) -> HigherChancePassFinalExam(x))",
    "ForAll(x, (AttendsTutoringSession(x) OR CompletesExtraPractice(x)) -> MoreLikelyImproveGrades(x))",
    "ForAll(x, NOT SubmitsResearchPaper(x) -> NOT PassCourse(x))",
    "Exists(x, OnAcademicProbation(x) AND GraduatesWithHonors(x))"
  ],
  "questions": [
    "Which statement can be inferred?\nA. ... \nB. If a student attends a tutoring session and completes at least 80% of the assignments, they are more likely to improve their grades and will pass the course.\nC. ... \nD. ...",
    "Is this statement true?\nStatement: If a student attends all lectures but does not submit their research paper, they still cannot pass the course even though they have a higher chance of passing the final exam."
  ],
  "answers": [
    "B",
    "Yes"
  ],
  "idx": [
    [1, 3],
    [2, 4]
  ],
  "explanation": [
    "Premise 3. states that attending tutoring (or doing extra practice) increases the likelihood of grade improvement. Premise 1. says completing at least 80% of assignments guarantees passing the course. Together, these two premises support option B.",
    "Premise 2. says attending all lectures raises the chance of passing the final exam, but Premise 4. says any student who fails to submit the research paper cannot pass the course. Both conditions can hold simultaneously, so the statement is true."
  ]
}

```

**Figure 1:** Example record from the XAI Challenge 2025 dataset. Each item includes natural and formal premises, natural language questions, indexed supporting evidence, and human-readable explanations.

credit thresholds, and exceptions for special circumstances. This design ensures that the dataset is both logically rigorous and pedagogically relevant. Questions are divided into three main categories:

- **Yes/No/Uncertain:** Binary evaluations of compound logical conditions (e.g., academic eligibility or rule violations).
- **Multiple-choice:** Selecting the most logically entailed conclusion from a list of candidates.
- **Numerical:** Inferring specific quantities (e.g., credit totals or counts) from constraints embedded in the premises.

Figure 1 illustrates a simple example from the dataset. Each record includes five core components: (i) premises expressed in both natural language (premises-NL) and formal logic (premises-FOL); (ii)

one or more questions designed to test reasoning ability; (iii) ground truth answers; (iv) supporting premise indices (idx) that identify which statements justify each answer; and (v) a natural language explanation that outlines the reasoning steps in a transparent and verifiable manner. This structure encourages systems to generate answers based on explicit logical evidence rather than statistical approximation. Moreover, the data format is directly aligned with the challenge’s evaluation framework, which emphasizes not only answer accuracy but also the clarity and traceability of the accompanying explanations.

The dataset was built through a three-stage pipeline that combines symbolic reasoning, LLMs, and expert validation. In the first stage, a custom logic engine built on top of the Z3 Solver was used to generate logically consistent premises. The engine applied classical inference rules such as Modus Ponens, Hypothetical Syllogism, and De Morgan’s Theorem to construct a diverse set of original, derived, and unrelated statements. The full procedure is described in Algorithm 1, which outlines how premises were sampled, validated, and assembled into complete records. In the second stage, each validated FOL statement was translated into natural language using ChatGPT. These translations aimed to preserve the original logical meaning while rendering the statements in fluent, readable English. In the final stage, trained university students manually reviewed and refined the records to ensure clarity, factual accuracy, and contextual relevance.

---

**Algorithm 1:** Premise Generation for XAI Challenge Dataset

---

**Input** : Number of steps  $s$ , number of chained premises  $c$ , number of derived premises  $d$   
**Output**: Dictionary {original, derived, unrelated} containing valid premises

- 1 Initialize empty sets: original, derived, unrelated;
- 2 Define logic variables (e.g.,  $P, Q, R$ ) and inference rules;
- 3 **Step 1: Generate original premises;**
- 4 **for**  $i \leftarrow 1$  **to**  $s$  **do**
- 5     Generate a premise using a random inference rule;
- 6     Validate with Z3;
- 7     **if** *valid and unique* **then**
- 8         Add to original;
- 9 **Step 2: Derive new premises;**
- 10 **for**  $i \leftarrow 1$  **to**  $d$  **do**
- 11     Select two premises from original or derived;
- 12     Derive a new premise, e.g., using implication or conjunction;
- 13     Validate with Z3;
- 14     **if** *valid and unique* **then**
- 15         Add to derived;
- 16 **Step 3: Add unrelated premises;**
- 17 **for**  $i \leftarrow 1$  **to**  $s - c$  **do**
- 18     Generate an unrelated premise using a random rule;
- 19     Validate with Z3;
- 20     **if** *valid and unique* **then**
- 21         Add to unrelated;
- 22 **return** {original, derived, unrelated};

---

The finalized dataset consists of 481 training records and 50 test records. Each record combines various types of premises and question formats, designed to assess both factual comprehension and multi-step logical reasoning. Table 3 provides an overview of the dataset’s composition, including statistics on premise length, question distribution, and reasoning depth.



**Table 3**

Dataset Statistics for the XAI Challenge 2025

Metric	Training Set	Test Set
Total Records	481	50
Average Premise Count per Record	9.90	6.08
Average Premise Length (Words)	126.60	85.16
Yes/No/Uncertain Records	457	13
Multiple-Choice Records	403	21
Numerical Records	16	16
Maximum Inference Steps	20	6
Maximum Premises per Record	36	10

### 3.4. Rules and Constraints

To ensure transparency, fairness, and alignment with the goals of XAI, the challenge introduced a clear set of modeling and technical constraints. Participants were required to build educational QA systems that could answer university policy questions and generate natural language explanations grounded in specific evidence. Each system received input in the form of a JSON object with two fields:

- **question**: a natural language question about academic policy,
- **premises**: a list of premises written in natural language.

The system was expected to return a JSON object with:

- **answer**: the predicted answer (e.g., Yes, No, Uncertain, a number, or a multiple-choice letter),
- **idx**: indices of the premises that support the answer, using one-based indexing,
- **explanation**: a concise, human-readable justification derived from the cited premises.

To support fair comparison and prevent data leakage, participants were provided with only the training portion of the dataset. The test set was kept private and used exclusively for final evaluation during the competition. All models were required to operate strictly on the given set of premises. External retrieval or lookup mechanisms were not allowed. To encourage transparency and discourage black box behavior, participants were recommended to use interpretable reasoning methods. These included symbolic solvers (such as Z3), lightweight open-source language models, or hybrid combinations. Regardless of approach, each system had to identify which premises were used and explain the reasoning process in a way that was understandable to non-expert users, especially students. In addition to modeling guidelines, each submission had to satisfy technical requirements to ensure reproducibility, robustness, and fairness during evaluation. Each system was deployed as an HTTP API and automatically tested on both private and public test cases. Table 4 outlines the main rules applied to all submissions.

### 3.5. Evaluation Protocol

As outlined in Table 2, our challenge employed a multi-phase evaluation framework to assess participants’ systems through hosted API endpoints, using both private and public datasets. The evaluation focused not only on answer accuracy but also on reasoning transparency and explanatory clarity. Each system was assessed along three primary dimensions:

- **Correctness of Answers ( $P_1$ )** — measured using *Exact Match* (EM) between the system’s output and the ground-truth answer field.
- **Relevance of Premises ( $P_2$ )** — evaluated via EM with the ground-truth **idx** field (one-based indexing), assessing whether the system selects the minimal correct subset of premises that support the answer.

<sup>13</sup><https://openai.com/index/openai-api/>

<sup>14</sup><https://platform.deepseek.com/>

**Table 4**  
Summary of rules and constraints in the XAI Challenge 2025

Aspect	Constraint
Model transparency	Only open-source models with fewer than 8 billion parameters were allowed without penalty. Submissions based entirely on proprietary models (e.g., GPT <sup>13</sup> , DeepSeek <sup>14</sup> ) were ranked lower to encourage reproducibility.
Reasoning method	Systems were required to generate natural language justifications grounded in specific premises. The use of symbolic solvers, lightweight language models, or hybrid systems was encouraged.
External data	Any external data used for training, fine-tuning, or augmentation had to be fully disclosed, including the source and intended use. Non-disclosure resulted in disqualification.
API protocol	Each system had to expose an API accepting POST requests with a JSON input containing a question and list of premises.
API output	The response had to include: (i) the predicted answer, (ii) one-based indices of the supporting premises, and (iii) a concise, human-readable explanation.
Lookup tables	Hardcoded or static responses were prohibited. All outputs had to be computed dynamically.
Rate limit	APIs were limited to 10 requests per second, with a maximum processing time of 60 seconds per request.
Availability	APIs that were offline for more than 30 consecutive minutes or failed more than 10% of test queries were disqualified.

- **Explainability ( $P_3$ )** — evaluates the clarity and logical coherence of the generated natural language explanation, which must be concise, faithful to the selected premises, and comprehensible to human users. In the final round,  $P_3$  was manually scored by the panel of professors based on a rubric.

Each of  $P_1$ ,  $P_2$ , and  $P_3$  is computed per instance and normalized to the range  $[0, 1]$ . Since each question in the dataset has exactly one correct answer and one minimal supporting set of premises (generated via logic-based templates and validated with symbolic solvers), exact match is a reliable metric for  $P_1$  and  $P_2$ . To ensure logical consistency, we enforce the constraint defined in Equation 1.

$$P_1 \cdot P_2 = 0 \Rightarrow \text{score} = 0 \quad (1)$$

**Selection Round.** This round consisted of two sub-phases. In each phase, systems were evaluated on the same set of  $n = 50$  private test cases using the same scoring scheme. For each instance, the score  $s_i$  was computed as shown in Equation 2.

$$s_i = 0.5 \cdot P_1 + 0.5 \cdot P_2 \quad (2)$$

The total score for each phase, denoted  $S^{(1)}$  and  $S^{(2)}$ , was computed by summing over all test cases, as defined in Equation 3.

$$S^{(k)} = \sum_{i=1}^{50} s_i^{(k)} \quad \text{for } k \in \{1, 2\} \quad (3)$$

After Phase 1, participants received feedback and had the opportunity to refine their systems before re-submission in Phase 2. To further encourage a deep understanding of the dataset and its logical structure, we also introduced a dataset feedback incentive. Specifically, a bonus score  $S_{\text{bonus}}$  was awarded based on the number and quality of valid issues (e.g., annotation errors, ambiguous premises) reported by each team. Although the dataset underwent rigorous validation, such community-driven review helped further improve its quality before public release. The final selection score  $S_1$  used to determine advancement was computed as shown in Equation 4.

$$S_1 = 0.6 \cdot \left( 0.7 \cdot S^{(1)} + 0.3 \cdot S_{\text{bonus}}^{(1)} \right) + 0.4 \cdot \left( 0.9 \cdot S^{(2)} + 0.1 \cdot S_{\text{bonus}}^{(2)} \right) \quad (4)$$



The five teams with the highest  $S_1$  scores were invited to the final round.

**Final Round.** In this round, each system was evaluated on  $n = 5$  public test cases and a live presentation session. For each test case, the instance-level score  $s_i$  was computed as shown in Equation 5.

$$s_i = 0.5 \cdot P_1 + 0.3 \cdot P_2 + 0.2 \cdot P_3 \quad (5)$$

The total model score  $S_2$  was computed by summing over all test cases, as shown in Equation 6.

$$S_2 = \sum_{i=1}^5 s_i \quad (6)$$

Each team also gave a live presentation, which was evaluated in two parts: a 7-minute technical presentation and a Q&A session with a panel of professors. Each part was independently scored by the judges using a 5-point Likert scale (1 = very poor, 5 = excellent), based on criteria such as clarity, content depth, delivery quality, and responsiveness. Let  $R_{\text{pres}}$  and  $R_{\text{Q\&A}}$  denote the average rubric scores for the presentation and Q&A portions, respectively. The final presentation score  $S_3$  was computed as shown in Equation 7.

$$S_3 = \frac{R_{\text{pres}} + R_{\text{Q\&A}}}{10} \quad (7)$$

The overall final score  $S$  used for ranking was computed as the average of model and presentation components, defined in Equation 8.

$$S = 0.5 \cdot S_2 + 0.5 \cdot S_3 \quad (8)$$

This composite score  $S$  determined the final team rankings and award decisions.

### 3.6. Participant Overview

The XAI Challenge 2025 attracted a diverse cohort of 107 participants, organized into 28 teams. Team sizes ranged from individual participants to groups of up to six members, enabling a variety of collaboration formats. This flexible structure allowed contributions from individuals across a broad spectrum of backgrounds and experience levels, including undergraduate students, graduate students, and early-career researchers. While most participants were based in Vietnam, the host country, and India, the challenge also welcomed teams from other countries. This international participation underscored the growing global relevance of XAI in education.

### 3.7. Results and Analysis

Table 5 reports the performance of the top five finalists across both sub-phases of the Selection Round. Each system was evaluated on the same set of 50 private test cases per phase, following the scoring procedure described in Section 3.5.

**Table 5**

Performance of top five finalists in the Selection Round (anonymized)

Team	Phase 1 Score	Phase 2 Score	Final Selection Score	Final Round Rank
Team A	19.90	21.00	20.34	2
Team B	19.10	22.05	20.28	1
Team C	13.85	25.80	18.63	4
Team D	17.60	18.45	17.94	5
Team E	18.45	16.65	17.73	3

**Modest Scores Reflect the Task’s Inherent Difficulty.** Although each evaluation phase included only 50 test cases, the absolute scores across all teams remained modest. The highest final selection score

barely surpassed 20, representing less than 50% of the maximum possible. This outcome highlights the intrinsic complexity of the task, which required systems not only to produce correct answers, but also to identify a minimal set of supporting premises and generate logically sound, human-understandable explanations. The results underscore the broader challenge of developing AI systems capable of faithful and interpretable reasoning over real-world educational policies.

**Two-Phase Evaluation Facilitated Iterative Improvement.** The two-phase structure gave teams an opportunity to refine their systems under tight time constraints. Notably, Team C demonstrated substantial improvement, raising its Phase 2 score by nearly 86% over Phase 1. This suggests that meaningful progress was achievable through targeted model updates and deeper engagement with the dataset. In contrast, Team E saw a decline in performance, showing that not all adjustments led to better results and highlighting the risk of overfitting or ineffective tuning during rapid iteration.

**Selection Scores Were Not Always Predictive of Final Rankings.** There was no strict correlation between selection scores and final round outcomes. For instance, Team A, which led the Selection Round, was ultimately overtaken by Team B in the Final Round. Meanwhile, Team E preserved its selection rank but narrowed the gap with top teams. These shifts reflect the multifaceted nature of the Final Round, where systems were evaluated not only on public test cases but also through live presentations assessed by a panel of experts. As a result, final rankings depended not just on model output, but also on a team’s ability to explain its design choices, reasoning strategies, and approach to explainability. This underscores the comprehensive and demanding nature of the challenge.

## 4. Selected Approaches

This section highlights several representative systems developed during the XAI Challenge 2025. While differing in architecture and methodology, these systems shared the common goal of producing accurate answers accompanied by logically grounded and interpretable explanations. We summarize four notable design paradigms that illustrate the diversity of strategies explored by the finalists.

**Multi-Agent Systems with Symbolic Reasoning.** One top-performing team implemented a modular multi-agent system that combined several lightweight open-source LLMs with symbolic reasoning via the Z3 theorem prover. The system divided the QA pipeline into specialized components: one agent parsed natural language premises, another applied logical inference using Z3, and a third synthesized structured explanations. Intermediate outputs were passed between agents to ensure modularity and traceability. This approach proved especially effective in handling complex queries involving conditional rules and policy exceptions, contributing to the team’s first-place result.

**Prompt-Based Learning with Task-Specific Templates.** Several teams adopted prompt-based learning, using carefully designed templates to guide lightweight LLMs in extracting relevant premises and generating step-by-step explanations. Prompts were tailored to specific question types, such as Yes/No/Uncertain, multiple-choice, or numerical answers. Some systems further employed Chain-of-Thought (CoT) [12] prompting to encourage intermediate reasoning steps. This strategy offered a lightweight and interpretable solution but remained constrained by the inherent opacity and prompt sensitivity of black-box models.

**Rule Retrieval with Symbolic Inference.** Another approach focused on rule retrieval and symbolic logic. One team built a structured rulebase of educational policies in Python and used keyword or semantic matching to identify candidate premises. These were then passed to the Z3 solver for formal inference. Explanations were generated by mapping the solver’s logical steps to human-readable justifications. This method performed well on regulation-heavy queries but showed limitations when dealing with ambiguous or loosely structured inputs.

**Multi-Task Fine-Tuning with a Mixture-of-Experts Architecture.** A learning-based system fine-tuned multiple lightweight LLMs on synthetic supervision for three distinct tasks: answer generation, premise selection, and explanation construction. These tasks were routed through a *Mixture-of-Experts*

(MoE) [13] architecture, where each expert model specialized in one task and was activated based on input type. While this approach benefited from task-specific learning and synthetic data control, it lacked the interpretability of symbolic systems and remained reliant on black-box inference.

These approaches illustrate a range of trade-offs between transparency, flexibility, and performance. Systems that integrated symbolic reasoning provided clear, verifiable explanations, while those based on language models offered adaptability and linguistic fluency. The challenge encouraged exploration across this design space, yielding valuable insights into the development of XAI systems for education and policy domains.

## 5. Conclusion

The XAI Challenge 2025 highlighted the feasibility and significance of developing transparent QA systems for educational contexts. By imposing constraints on model size and requiring verifiable reasoning, the competition challenged participants to design solutions that balanced accuracy with interpretability and trustworthiness.

The event drew a variety of approaches, including multi-agent architectures, prompt-based learning pipelines, rule-driven retrieval systems, and multi-task fine-tuning strategies. Despite their differences, these systems shared a common objective: to integrate natural language understanding with logical inference in order to answer policy-related queries with clarity and justification. A key design constraint was explainability, enforced through open-source implementation and the requirement to output explicit reasoning chains. As a result, models were evaluated not only on the correctness of their answers but also on their ability to select supporting premises and articulate coherent explanations.

Taken together, the outcomes of the challenge provide meaningful insights into the design of XAI systems in high-stakes settings. The event demonstrates that *bridging LLMs and symbolic reasoning is not only possible, but can yield practical, interpretable solutions to real-world tasks such as educational QA*. This serves as a foundation for future research at the intersection of language understanding, reasoning, and explainability.

## Acknowledgments

We would like to express our sincere gratitude to the URA Research Group at Ho Chi Minh City University of Technology (HCMUT), Vietnam, especially the undergraduate students who contributed to building and reviewing the dataset. We also acknowledge Bao Gia Quach, Hung Canh Nguyen, Nguyen Bao Le, Hieu Tran Hoang Nguyen, Hoang Huy Vu, Thuong Tran Anh Le, and Quynh Thi Nhu Vo for their support in both technical coordination and team communication throughout the challenge.

Finally, we are grateful to Professor Akka Zemhari and Professor Pascal Desbarats from the University of Bordeaux, France, for their visit to HCMUT and valuable discussions during the early proposal phase, together with Professor Fabien Baldacci, co-author of this paper.

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References

- [1] G. Briscoe, Digital Innovation: The Hackathon Phenomenon, Technical Report, Creative Works, Queen Mary University of London, 2014.
- [2] J. Porras, A. Knutas, J. Ikonen, A. Happonen, J. Khakurel, A. Herala, Code camps and hackathons in education — literature review and lessons learned, in: Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS 2019), 2019, pp. 7750–7759.

- [3] M. Kamariotou, F. Kitsios, Hackathons for Driving Service Innovation Strategies: The Evolution of a Digital Platform-Based Ecosystem, *Journal of Open Innovation: Technology, Market, and Complexity* 8 (2022) 111.
- [4] C. Kooli, Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions, *Sustainability* 15 (2023).
- [5] R. Sajja, C. E. Ramirez, Z. Li, B. Z. Demiray, Y. Sermet, I. Demir, Integrating Generative AI in Hackathons: Opportunities, Challenges, and Educational Implications, *Big Data and Cognitive Computing* 8 (2024).
- [6] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), *IEEE Access* 6 (2018) 52138–52160.
- [7] T. H. Trinh, Y. Wu, Q. V. Le, H. He, T. Luong, Solving olympiad geometry without human demonstrations, *Nature* 625 (2024) 476–482.
- [8] M. Garnelo, M. Shanahan, Reconciling deep learning with symbolic artificial intelligence: representing objects and relations, *Current Opinion in Behavioral Sciences* 29 (2019) 17–23.
- [9] S. Min, J. Boyd-Graber, C. Alberti, D. Chen, E. Choi, M. Collins, K. Guu, H. Hajishirzi, K. Lee, J. Palomaki, C. Raffel, A. Roberts, T. Kwiatkowski, P. Lewis, Y. Wu, H. Küttler, L. Liu, P. Minervini, P. Stenetorp, S. Riedel, S. Yang, M. Seo, G. Izacard, F. Petroni, L. Hosseini, N. D. Cao, E. Grave, I. Yamada, S. Shimaoka, M. Suzuki, S. Miyawaki, S. Sato, R. Takahashi, J. Suzuki, M. Fajcik, M. Docekal, K. Ondrej, P. Smrz, H. Cheng, Y. Shen, X. Liu, P. He, W. Chen, J. Gao, B. Oguz, X. Chen, V. Karpukhin, S. Peshterliev, D. Okhonko, M. Schlichtkrull, S. Gupta, Y. Mehdad, W.-t. Yih, *NeurIPS 2020 EfficientQA Competition: Systems, Analyses and Lessons Learned*, in: *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133, 2021, pp. 86–111.
- [10] P. Batorski, J. Brinkmann, P. Swoboda, NSA: Neuro-symbolic ARC Challenge, 2025.
- [11] L. de Moura, N. Bjørner, Z3: An Efficient SMT Solver, in: C. R. Ramakrishnan, J. Rehof (Eds.), *Tools and Algorithms for the Construction and Analysis of Systems*, 2008, pp. 337–340.
- [12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, in: *Advances in Neural Information Processing Systems* 35 (NeurIPS 2022), volume 35, 2022, pp. 24824–24837.
- [13] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al., *Mixtral of Experts*, 2024.