Valencia, 2nd September 2025

15th Model-driven Requirements Engineering Workshop

Assisting Stakeholders in Class Diagram Interpretation with LLMs

a Work in Progress

Chiara Mannari[1,2], Tommaso Turchi[2], Manlio Bacco[1], Alessio Malizia[2]

[1] Institute of Science and Technologies "A. Faedo" ISTI - CNR

[2] Department of Computer Science, University of Pisa

Università di Pisa

Consiglio Nazionale delle Ricerche

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Context and Motivation

- Growing interest in generative AI — LLMs—, including within the MoDRE community
- Strong focus on diagram creation with LLMs
- The opposite direction — **deriving textual explanations from diagrams** — remains less explored

WHY THIS MATTERS

- *Empirical evidence*: application of a MoDRE-based method within interdisciplinary teams; development of an end-user modelling tool
- *Literature*: earlier work [Leopold et al., 2014] — opening space to extend it with LLM-based methods

CODECS
Co-funded by the European Union

1. H. Leopold, J. Mendling, and A. Polyvyanyy, "Supporting process model validation through natural language generation," IEEE TSE, vol. 40, 2014.

Objective - LLM-generated interactive layer

Tooltip

local explanation, interactive

LLM-generated

Class diagram

human-generated

Diagram description

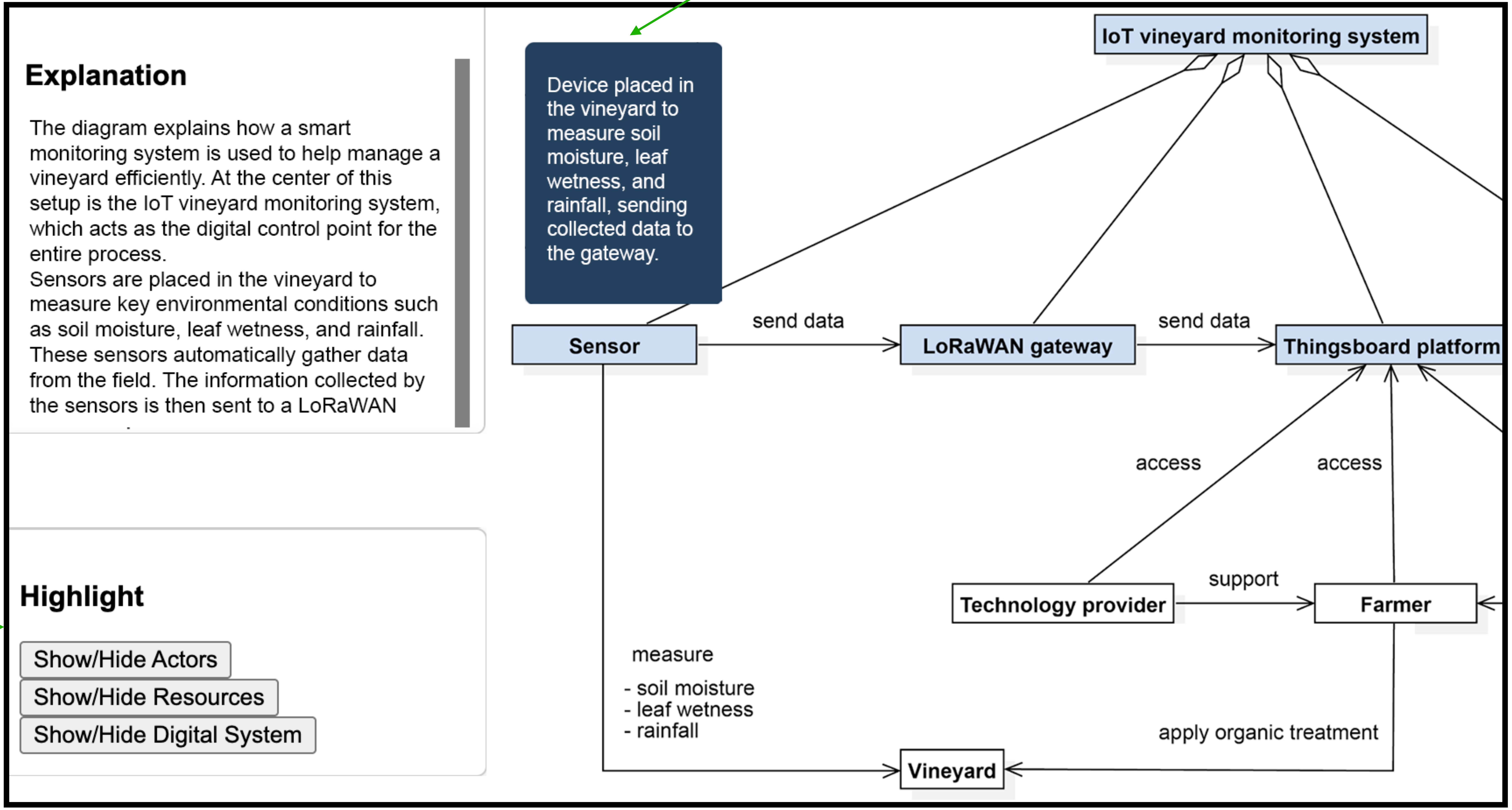
general explanation, static

LLM-generated

Advanced features

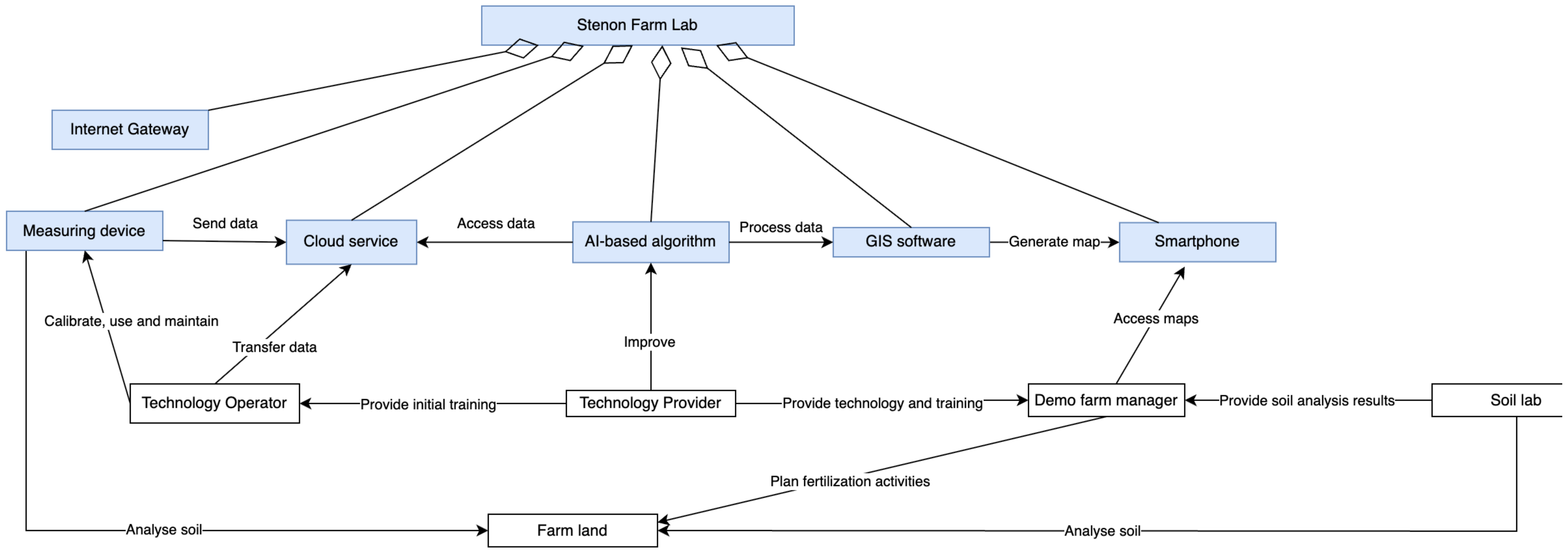
entities classification, interactive

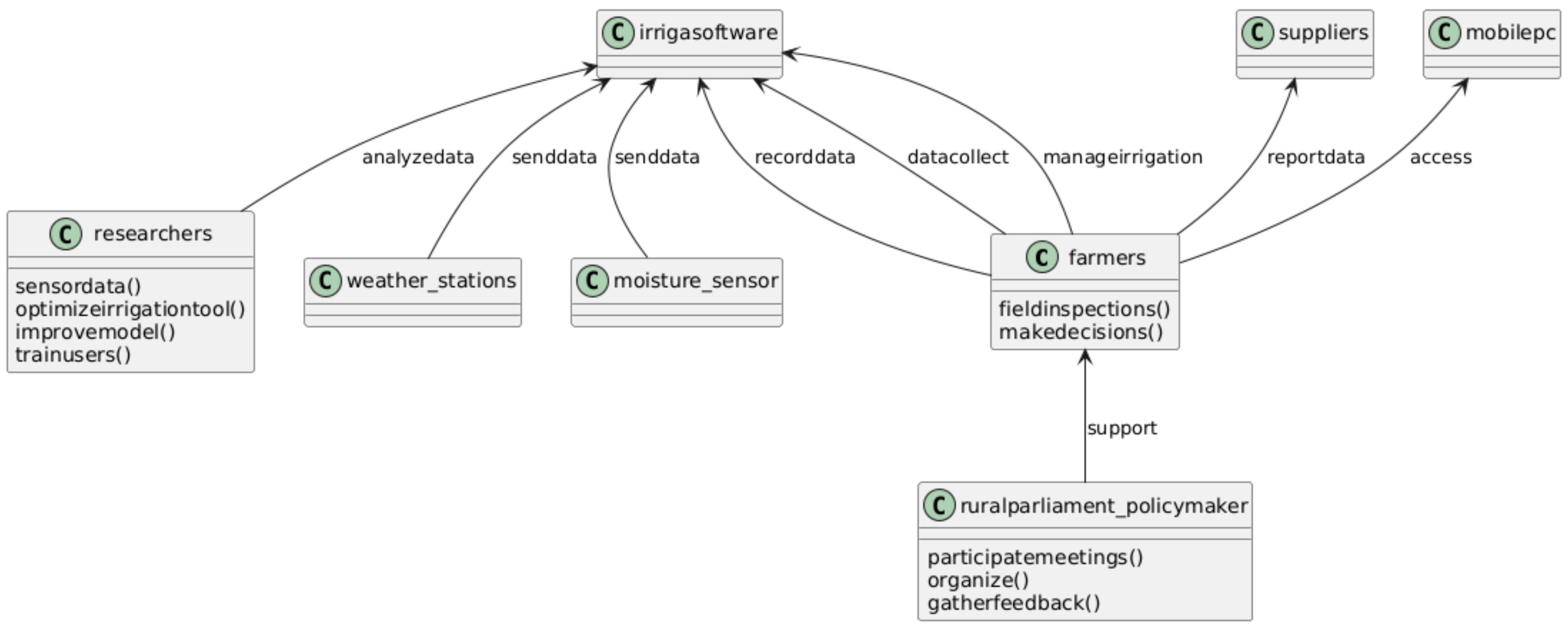
LLM-generated



Technical evaluation - input diagrams

Case		Description	UML	Software	Format
1	IoT vineyard monitoring system	An IoT monitoring system based on sensors installed on-field that measures several parameters in the vineyard to optimize organic treatments.	No: nodes #9; arches #12 Types: class, aggregation, direct association	StarUML	png svg xmi
2	Soil scanner	A soil scanner with sensors and AI-based software that measures soil parameters to optimise fertilisation.	No: nodes #12; arches #20 Types: class, aggregation, direct association	draw.io	png svg drawio
3	Smart irrigation	An AI-based system to monitor field and weather conditions and manage irrigation	No: nodes #8; arches #9 Types: class, direct association	ModeLLer	png PlantUML xmi





Prompts

LLM: GPT4o - GPT4.1

Prompt 1 write a summary that explains the uploaded diagram to a non-technical audience

Prompt 2 (chain-of-prompts) detect UML classes and return a table listing:

- **NAME**
- **DESCRIPTION** (20–30 words, non-technical, explaining role and interactions)
- **TYPE** (digital / actor-organisation / natural resource / other)
- **POSITION** (X, Y, width, height)

Evaluation criteria

Criteria 1-4 inspired by prior work [Ferrari et al., 2024]

1. **Completeness**: the text covers the content of all the (main) entities with a sufficient degree of detail to explain the content of the model to potential stakeholders.
2. **Correctness**: the text describes a system structure that is coherent and consistent with the diagram.
3. **Degree of understandability**: the text is sufficiently clear, given the complexity of the diagram, and does not contain redundancies.
4. **Terminological alignment**: the terminology used in the generated text aligns with the one used in the diagram.

Additional criteria

5. **Acceptability**: the extent to which the positions of the tooltips in the generated interactive layer align with their correct placement as defined in the UML source model.

Likert scale 5: "1– Not fulfilled at all; 2– Fulfilled to a minimal extent; 3– Partially fulfilled; 4– Mainly fulfilled; 5– Completely fulfilled" + comments

2 evaluators, evaluations averaged

2. A. Ferrari, S. Abualhaija, and C. Arora, "Model generation with llms: From requirements to uml sequence diagrams," in 2024 IEEE REW.

Execution and results / 1

PROMPT 1 - diagram summary

- 18 summaries, av. 220 words (range 176-260)

- GPT-4.1 longer outputs: av. 239 words; GPT-4o, av. 202 words

Criteria: ***Completeness, Correctness, Degree of understandability, Terminological alignment***

- The **average output quality is high** (between 4 and 5)

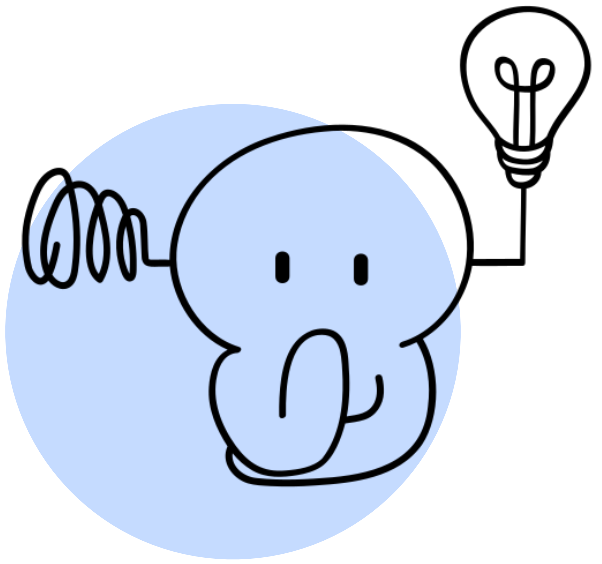
Comments from evaluators: *extra content not present in the original data; commentary and interpretative statements; a few instances of hallucination (unmentioned operations), omissions*

Execution and results / 2

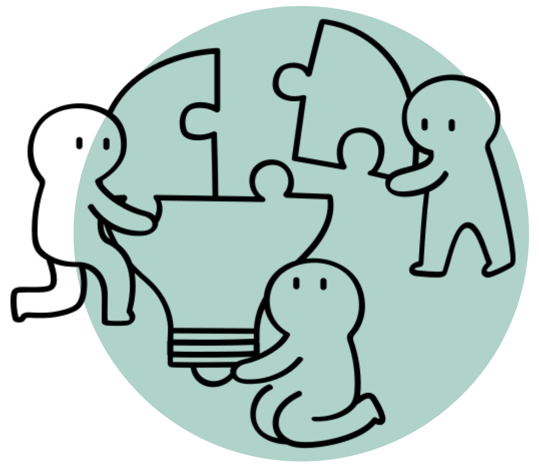
PROMPT 2 - tooltip table

STEP	*CRITERIA*	*APPROACH*	*KEY RESULTS*
Class extraction	Completeness	Precision: TP (nodes correctly detected) / TP + FP (nodes incorrectly identified) Recall: TP / FN (nodes missed)	• GPT-4.1 perfect score • GPT-4o: variability (low on case 1)
Tooltip description	Completeness, Correctness, Degree of understandability	Completeness: accuracy (no. edges mentioned/no. edges) Other: 5-point Likert scale + comments	• Variability (medium-high results) • Notes: *aggregation not recognised; missing info; content additions*
Classification	Correctness	Boolean + comment	• Error rate: 0% GPT-4.1; 18% GPT-4o • Weather station and moisture sensor classified as natural resources
Positioning	Acceptability	Likert scale + comment	• High variability (low-medium results) • GPT-4.1 higher score

Takeaway lessons



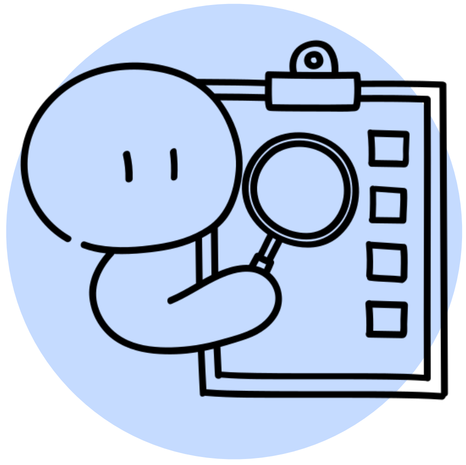
Although based on preliminary findings, results highlight the **technical feasibility of an LLM-generated layer** to support users in diagram reading and validation, *across most features overall*, and **encourage further experimentation.**



Limitations: both models struggle with contextual understanding, fine-grained details, and risk introducing hallucinated content.

Possible solution: **alert users when content is AI-generated** and allow them to **choose between models.**

Future works

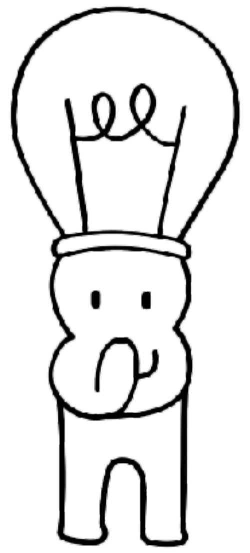


User validation: Test the interactive LLM-generated layer with real users

Extend technical exploration: Experiment with additional models (GPT-5, recently released), LLMs (DeepSeek, LLama, Gemini, and others) , increase the input data even with more complex diagrams, or diagrams containing errors or inconsistencies, focus on specific evaluation criteria, test advanced prompting strategies

Thanks for the attention

Your feedback is much appreciated



icons designed by Freepik

chiara.mannari@isti.cnr.it