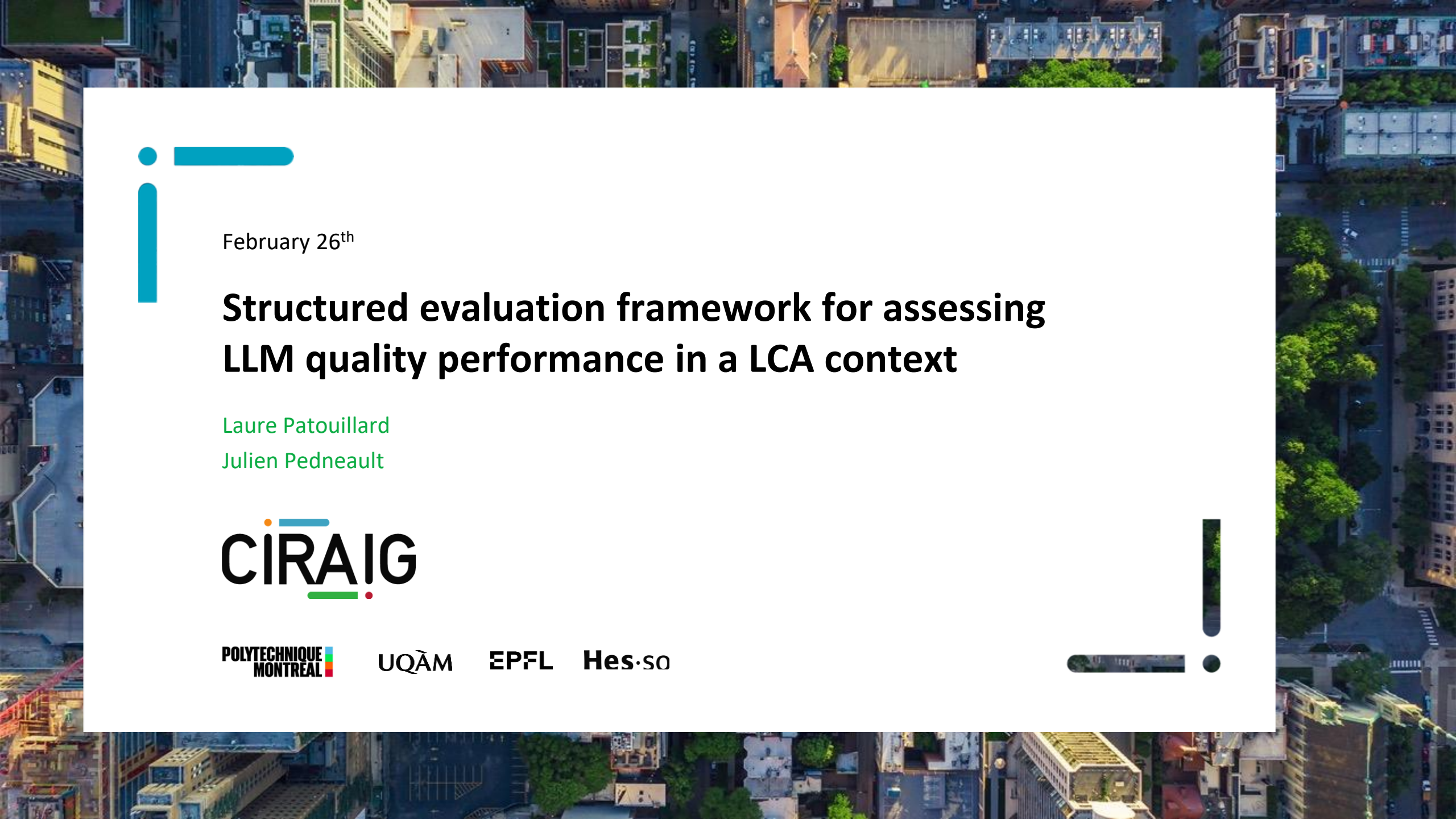


February 26th

Structured evaluation framework for assessing LLM quality performance in a LCA context

Laure Patouillard
Julien Pedneault

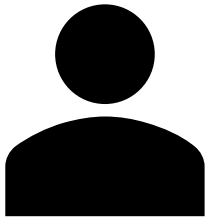
CIRAIG

POLYTECHNIQUE MONTRÉAL UQÀM EPFL Hes·so

Need for a rigorous evaluation method for LCA-related LLM outputs

Due to the probabilistic nature of LLMs, their use in LCA may lead to data hallucinations, methodological inconsistencies, or non-compliance with standards.

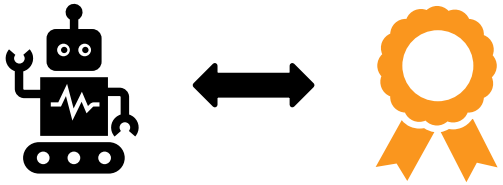
Evaluation approaches:



Human review

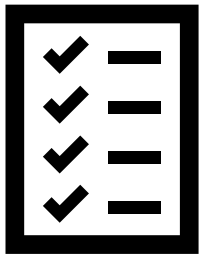
Time consuming

(Donaldson et al, 2025)



Direct comparison with ground truth

Classic approach in ML (classification)
Does not apply to all LCA tasks
No database of groundtruth available
Inherent subjectivity in LCA



(Semi-)Quantative evaluation based on qualitative criteria

Possibility to use a LLM-as-a-judge

Objectives

Objectives of the study

Develop a structured and reproducible evaluation framework for LLM outputs in LCA

Potential applications of the framework

- Enable systematic comparison of the performances of different LLMs
- Define under which settings LLMs can match or complement human expertise
- Determine for which LCA tasks LLM are suitable
- Track performance progress for an LLM during training

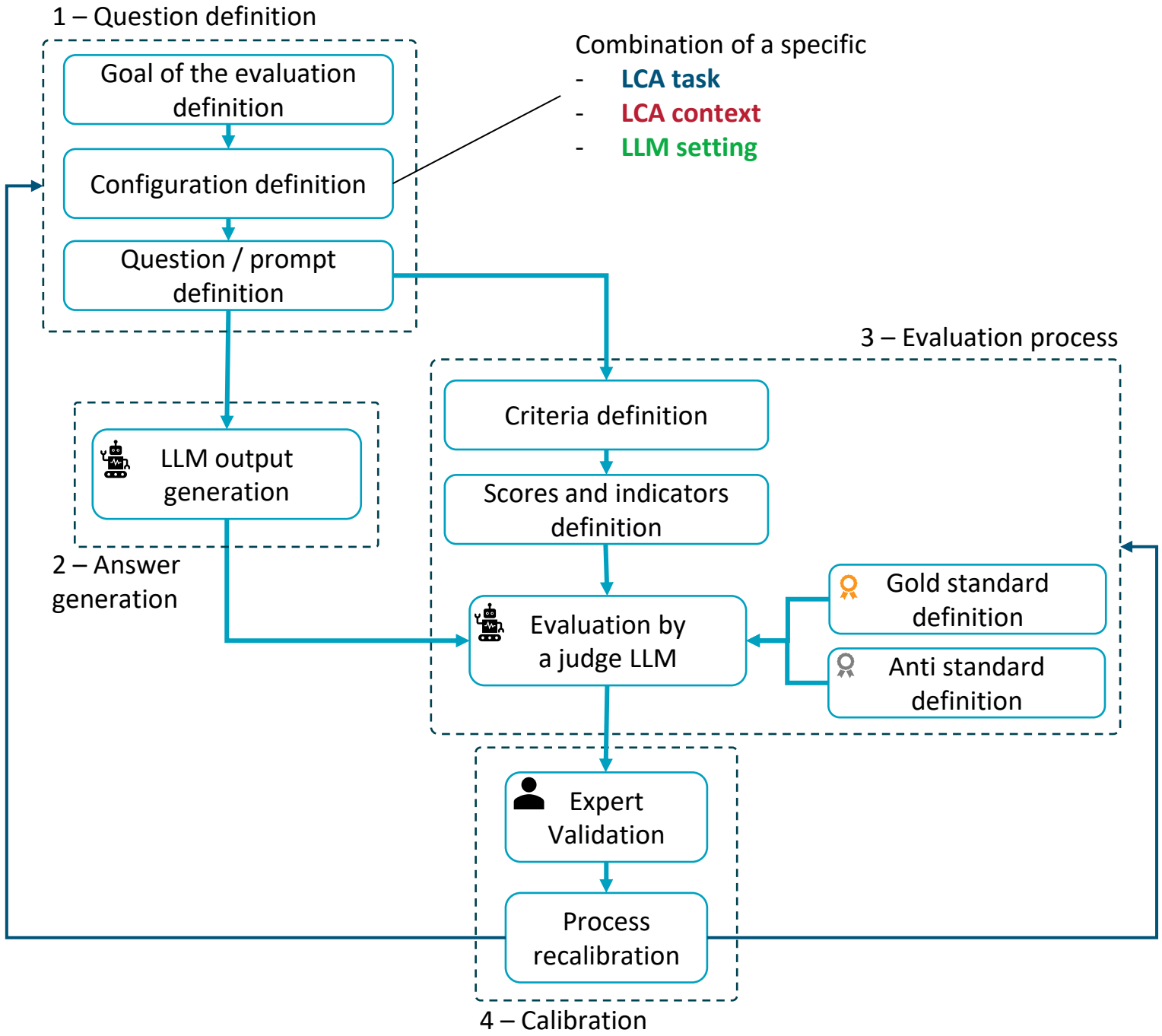
Evaluation framework

Evaluation framework

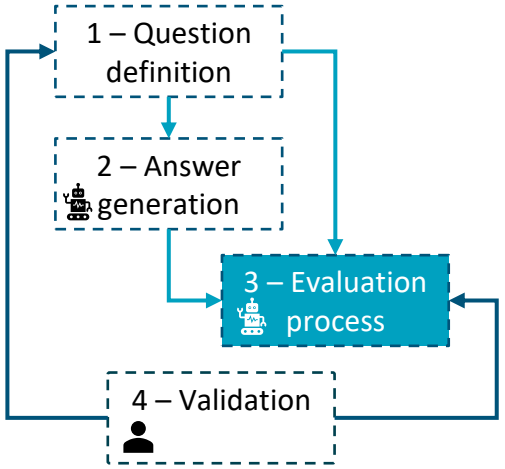
Four-step iterative process:

1. **Question definition**
2. **Answer generation**
3. **Evaluation process**
4. **Validation**

Outputs generated by LLM in a specific context are assessed against predefined generic and task-specific criteria using gold standards as guide



Evaluation framework – Evaluation process

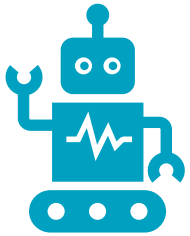


For one configuration to be evaluated (`config1`)
Combination of a specific **LCA task**, different **LCA contexts**, and particular **LLM settings**
Example: Goal & scope definitions for a low-TRL industrial process using a specific prompting strategy

LLM-as-a-judge

Answers

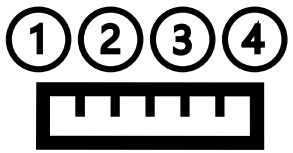
```
config1.Q1 → A1.1: « blabla… »
config1.Q2 → A1.2: « blibli… »
config1.Q3 → A1.3: « bloblo… »
…
```



Scores for each criteria + justification

	C1	C2	C3
A1.1	1	2	4
A1.2	2	1	3
A1.3	1	4	4
…			

Criteria
Generic criteria (for any LCA task): Explainability, Verifiability, Instruction Following
+ Specific criteria tailored to individual LCA task



Scores
Four-point scale providing guidance to the evaluator (in this case, a LLM) for each criteria

Gold standard
Groundtruth data to serve as examples to guide the LLM-as-a-judge evaluation
→ Set of plausible answers to reflect the partial subjectivity inherent to LCA

Interpretation of scores remains an open question

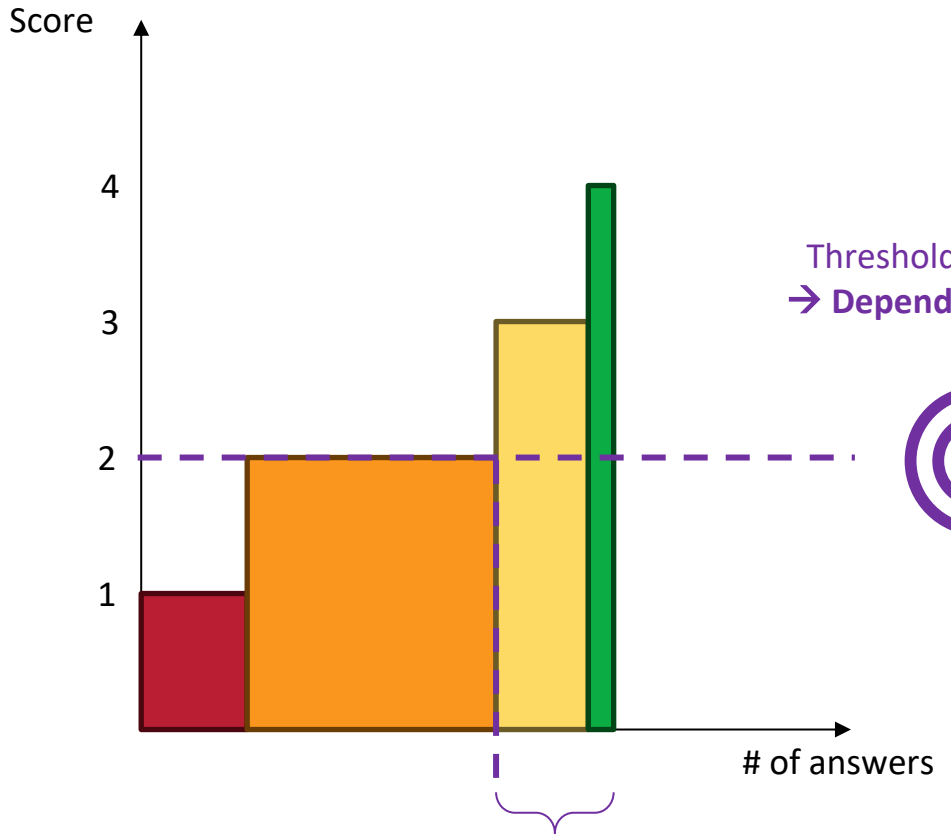
Descriptive statistics

	C1	C2	C3	*Weighting?*
A1.1	1	2	4	*2.3*
A1.2	2	1	3	*2*
A1.3	1	4	4	*3*
…				
Mean	1.3	2.3	3.6	*2.4*
Stdev	0.4	1.1	0.4	*0.3*

« For config1, the LLM perform better on C3 and C1 »

→ **How to conclude?**

Minimum quality threshold



Threshold for min score
→ **Depend on the context**

« 25% of answers > threshold of 2 »
→ **Is it good enough?**

Testing the framework on some LCA tasks

CIRAIG

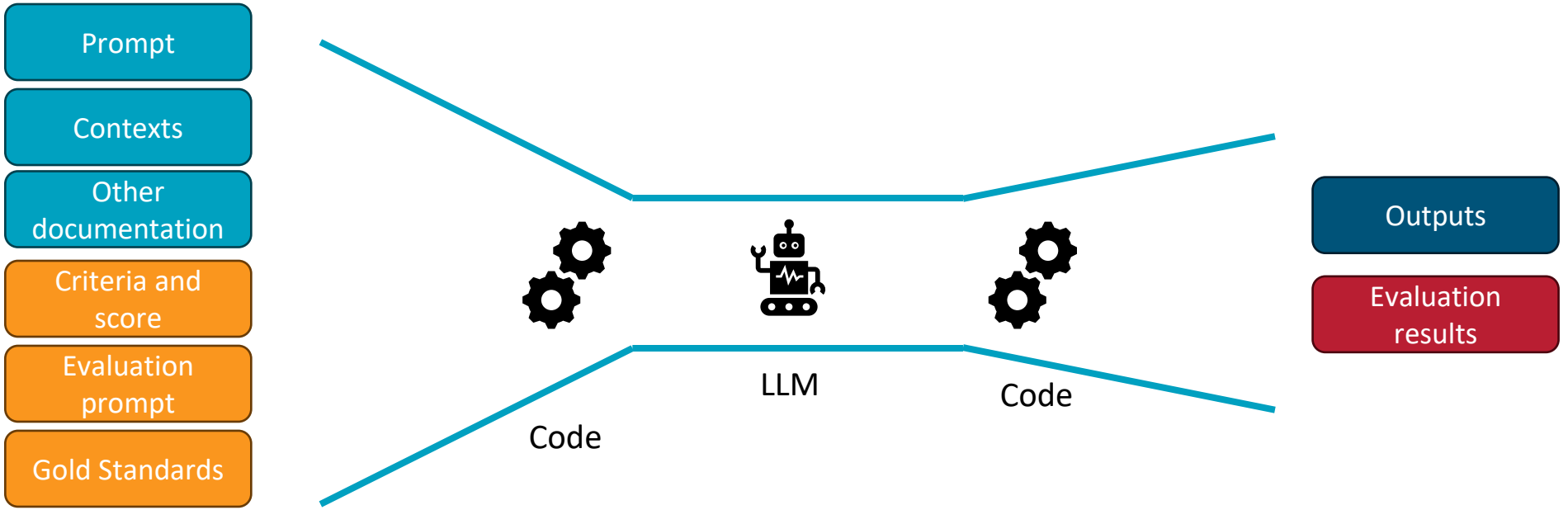
Preliminary insights

- Goal and scope definition
  - Instructional prompting leads to higher scores
  - Human validation is challenging and demands high concentration
- Mapping to LCI database
  - Many hallucinations (made up flow-process-geography) if no process list is provided
- Inventory generation
  - Defining a *gold standard* is difficult, almost impossible, even more for a low TRL technology
  - Human validation needs to be done by an expert of the process
  - The approach is better suited for initial structuring than generating a full LCI

Toward the automation of the framework

Increase the number of evaluations by using an automated pipeline:

- Enable testing the robustness (multiple runs for same question)
- Structure the outputs to facilitate result processing and the validation step
- Allow to easily test different LLMs

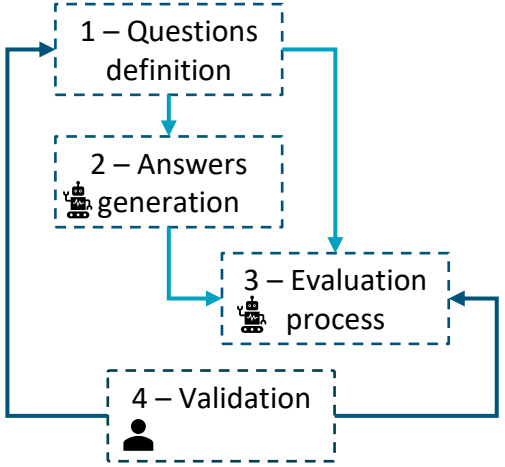


Discussion

Strenghts of our evaluation framework

- Systematically reveal performance trends across different LCA contexts and LLM configurations
- Applicable for any LCA tasks
- Applicable for a human judge or LLM judge
  - Although we recommend LLM-as-a-judge to enable scaling the evaluation process
- Limit the risk of the LLM overfitting the evaluation QA
  - Public groundtruth not used for direct evaluation but for guiding the judge
  - Evaluation questions should be updated regularly
- Automated and Ready-to-use, althought will be improved with future work

Future work



- Test a larger number of configurations
- Strenghten criteria definition, especially for the ones specific to LCA tasks
- No clear minimum quality threshold: defining what constitutes an “acceptable” level of quality
- Develop and test indicators for interpretation of scores
- Compare this LLM-as-a-judge approach to human review to position/validate it
- Create a collaborative groundtruth database to reflect our field subjectivity

Thank you
Questions?

Laure Patouillard

Julien Pedneault

CIRAIG

Bilbiography

Donaldson, A., Balaji, B., Oriekezie, C., Kumar, M., & Patouillard, L. (2025). An Expert-grounded benchmark of General Purpose LLMs in LCA. *ArXiv Preprint ArXiv:2510.19886*. https://doi.org/https://doi.org/10.48550/arXiv.2510.19886