A Certification Framework for Large Language Models

Alex Della Bruna

July 14, 2026

Supervisor: Prof. Marco Anisetti
Co-supervisors: Prof. Claudio A. Ardagna, Dr. Nicola Bena

Context

Rise of **Large Language Models** (LLMs)

- **Ubiquity:** becoming standard across many applications
- **Capabilities:** text generation and connection to external tools (e.g., Model Context Protocol)
- **Lifecycle integration:** used in the software development lifecycle
- **Security gap:** their opacity raises security and privacy concerns

$\Longrightarrow$ **Assessment and certification-based assurance techniques are needed**, yet existing techniques fall short

Motivations

Gaps of **traditional assurance** techniques

- **Transparency:** training data and architectures are often hidden
- **Non-functional assessment:** accuracy is the only focus, while non-functional properties (e.g., safety, robustness) are neglected
- **Ambiguous properties:** non-deterministic behavior lacks clear metrics
- **End-to-end assessment:** assessing the inference outcome only does not cover the entire LLM life cycle

Objectives

A **multi-dimensional certification scheme** for LLM-based applications

- **Assess LLM-based applications against a set of compatible behaviors:** definition of the set of behaviors that are compatible with the required property p, to address non-determinism
- **Multi-dimensional analysis:** in-depth assessment of LLMs by considering multiple facets of their behavior
- **Case studies:** apply the scheme to real-world scenarios

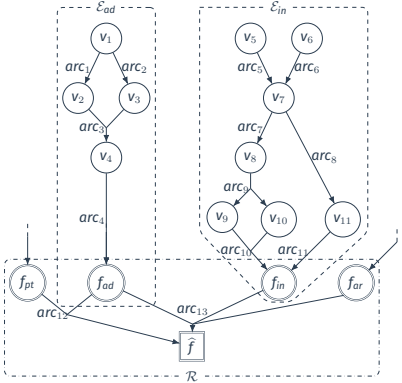
Nicola Bena, Marco Anisetti, Ernesto Damiani, **Alex Della Bruna**, Chan Yeob Yeun, Claudio A. Ardagna. "A Certification Scheme for Large Language Models-Based Applications." *ACM Transactions on Intelligent Systems and Technology*, 2026. DOI: https://doi.org/10.1145/3819077

Certification Model

The certification process is driven by a **certification model** $\mathcal{CM}$

Certification Model $\mathcal{CM}$

- a textual label p describing the non-functional property
- a hypergraph $\mathcal{B}$ modeling the set of behaviors compatible with p

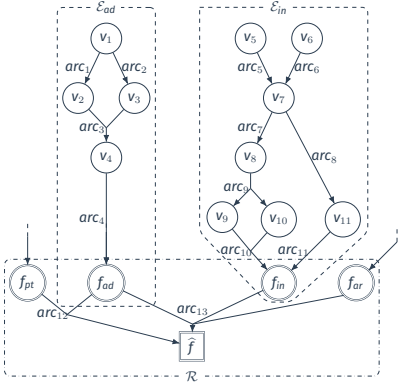


Certification Model

The certification process walks through different **dimensions**

Dimension *d*

A specific facet of the behavior to be verified

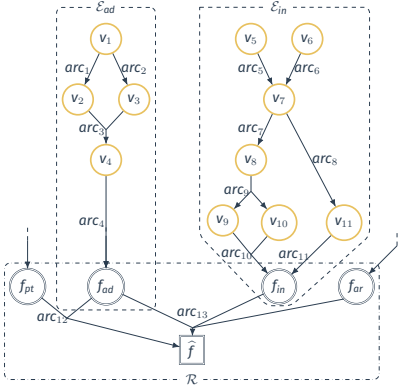


Considered: pre-training (*pt*), adaptation (*ad*), inference (*in*), and artifact (*ar*)

Certification Model

The hypergraph $\mathcal{B}$ is composed of multiple **sub-hypergraphs** $\{\mathcal{E}\}$ and **evaluation nodes** $\{f\}$

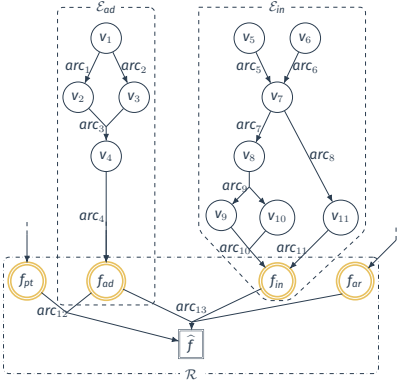
- **a sub-hypergraph $\mathcal{E}_i$ is a subset of $\mathcal{B}$ that models the behaviors compatible with a specific dimension d_i**
- each $\mathcal{E}_i$ is associated with an evaluation node f_i
- all the evaluation nodes $\{f_i\}$ are aggregated into a global evaluation node $\widehat{f}$



Certification Model

The hypergraph $\mathcal{B}$ is composed of multiple **sub-hypergraphs** $\{\mathcal{E}\}$ and **evaluation nodes** $\{f\}$

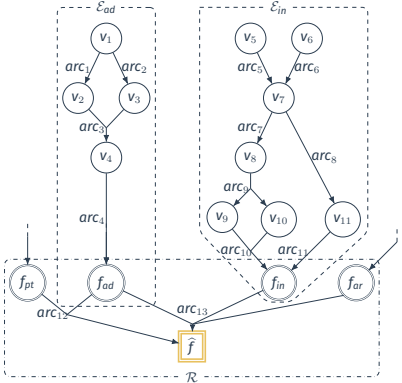
- a sub-hypergraph $\mathcal{E}_i$ is a subset of $\mathcal{B}$ that models the behaviors compatible with a specific dimension d_i
- **each $\mathcal{E}_i$ is associated with an evaluation node f_i**
- all the evaluation nodes $\{f_i\}$ are aggregated into a global evaluation node $\widehat{f}$



Certification Model

The hypergraph $\mathcal{B}$ is composed of multiple **sub-hypergraphs** $\{\mathcal{E}\}$ and **evaluation nodes** $\{f\}$

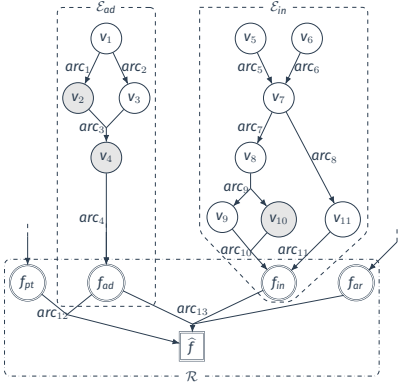
- a sub-hypergraph $\mathcal{E}_i$ is a subset of $\mathcal{B}$ that models the behaviors compatible with a specific dimension d_i
- each $\mathcal{E}_i$ is associated with an evaluation node f_i
- **all the evaluation nodes $\{f_i\}$ are aggregated into a global evaluation node $\widehat{f}$**



Certification Process

The **certification process checks** which of the **admissible behaviors** are actually **supported** by the application

- Hypergraph exploration algorithm that visits the nodes representing specific behavior facets
- If at least one compatible behavior is found, then the application supports the property *p*



Certification Process

Finally a **certificate** is issued, which contains the **certification model** $\mathcal{CM}$ and the **evidence** of the behaviors

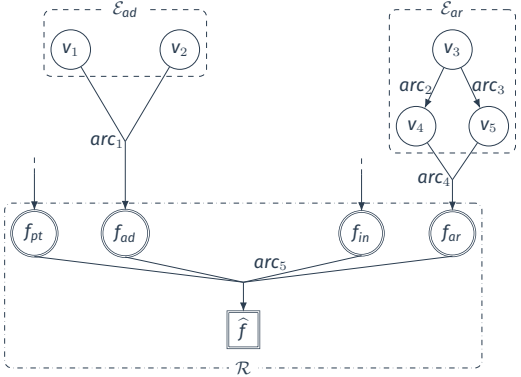
- Evidence shows which behaviors are supported and which not

Collected evidence	
$EV(v_1)$	`repositories-inspection=✓`
$EV(v_2)$	`linting=✗`
$EV(v_3)$	`incremental-difficulty-inspection=✓`
$EV(v_4)$	`diversity-check=✗`
$EV(v_5)$	`jailbreak-injection=✓`
$EV(v_6)$	`inference-time=✓`
$EV(v_7)$	`ha-inclusion=✓`
$EV(v_8)$	`inbound-traffic-inclusion=✓`
$EV(v_9)$	`health-check-inclusion=✓`
$EV(v_{10})$	`recovery-inclusion=✗`
$EV(v_{11})$	`load-balancer-threshold-check=✓`

Case Study: LLM Integrity

We applied the certification scheme to assess the **integrity** of an LLM after a fine-tuning process

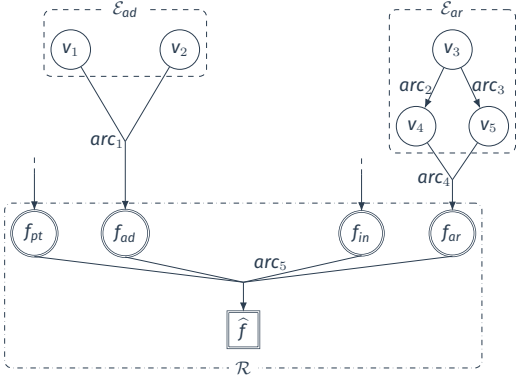
- **Compared LLMs:**
  - Base model: GPT-OSS:20b
  - Fine-tuned model: GPT-OSS:20b fine-tuned with qLoRA
- **Integrity:** the property p of an LLM to maintain its original output after a fine-tuning process



Case Study: LLM Integrity

The assessment focuses on:

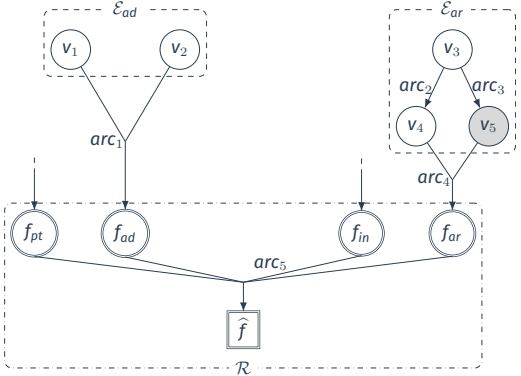
- **General reasoning (v_4):** reasoning capabilities satisfy minimum requirements, assessed through a "Golden Dataset"
- **Accordance (v_5):** responses remain semantically similar, assessed using G-Eval (LLM-as-a-Judge)



Case Study: LLM Integrity

The assessment focuses on:

- **General reasoning (v_4):** reasoning capabilities satisfy minimum requirements, assessed through a "Golden Dataset" $\Longrightarrow$ passed ($EV(v_4)$)
- **Accordance (v_5):** responses remain semantically similar, assessed using G-Eval (LLM-as-a-Judge) $\Longrightarrow$ failed ($EV(v_5)$)



Case Study: LLM Integrity

The assessment focuses on:

- **General reasoning (v_4):** reasoning capabilities satisfy minimum requirements, assessed through a "Golden Dataset" $\Longrightarrow$ passed ($EV(v_4)$)
- **Accordance (v_5):** responses remain semantically similar, assessed using G-Eval (LLM-as-a-Judge) $\Longrightarrow$ failed ($EV(v_5)$)

Collected evidence	
$EV(v_1)$	`download=`✓
$EV(v_2)$	`fine-tuning=`✓
$EV(v_3)$	`golden-dataset-creation=`✓
$EV(v_4)$	`behavior=`✓
$EV(v_5)$	`g-eval=`✗

Case Study: LLM Integrity

The assessment focuses on:

- **General reasoning (v_4):** reasoning capabilities satisfy minimum requirements, assessed through a "Golden Dataset" $\Longrightarrow$ passed ($EV(v_4)$)
- **Accordance (v_5):** responses remain semantically similar, assessed using G-Eval (LLM-as-a-Judge) $\Longrightarrow$ failed ($EV(v_5)$)
- thus, the fine-tuned model is **not certified** for the property integrity

Collected evidence	
$EV(v_1)$	`download=`✓
$EV(v_2)$	`fine-tuning=`✓
$EV(v_3)$	`golden-dataset-creation=`✓
$EV(v_4)$	`behavior=`✓
$EV(v_5)$	`g-eval=`✗

Conclusions

A **multi-dimensional certification scheme for LLM-based applications**

- based on the definition of the set of **compatible application behaviors**
- supports a **multi-dimensional assessment** covering different aspects of the LLM lifecycle

Future Work

- expand the framework to assess agentic AI applications
- strengthen the integration within enterprise MLOps architectures