Project Proposal – Responsible AI and the Law

Working Title: Process vs. Outcome Supervision for Trustworthy and Interpretable LLM Reasoning

By : Harshit Ojha(ho2228) , Dhairyasheel Patil(dp3979)

1. Research Question (+ Significance)

Research Question:

How does encouraging and exposing step-by-step reasoning compare to outcome based supervision i.e. only final answers in terms of factual reliability and interpretability/auditability for multi-step reasoning tasks in large language models?

We feel like this question matters for Responsible AI and the Law because many new and upcoming legal frameworks like AI accountability, algorithmic transparency, and safety by design requirements assume that AI systems can provide meaningful explanations of their outputs. If process based supervision produces reasoning that is actually easier to verify and more closer to correct solutions, it could support legal arguments that a model decision is explainable and hence reviewable. If, instead, if explanations look structured but are misleading or unlawful, they may actively undermine accountability and create liability risks.

2. Approach

2.1 Experimental Setup

We would like to propose a controlled evaluation that compares two conditions for the same base models on multi-step reasoning tasks like, arithmetic or math word problems,

- **Outcome-only condition (Final Answer)**
    - **Prompt:** Answer the question. Do not show me your reasoning or any reasons why you computed what you did. Give only the final answer.
    - Simulates outcome-based supervision, where only final outputs are visible and clickable or downloadable.
- **Process condition (Process)**
    - **Prompt:** Think step by step and show me your reasoning. Then give a final answer at the end after showing all the steps one by one.
    - Simulates process-based supervision, where intermediate reasoning is surfaced and could be rewarded or reviewed for further study.
- **Optionally, a stricter variant: Structured Process condition (Process+)**
    - **Prompt:** Write your solution as: Step 1, 2 , 3, followed by the final answer
    - This makes it easier to analyze the correctness of each step

Models / APIs , as addressed in the feedback, by publicly available APIs, we intend to use:

- At least one open-weight model exposed through an inference endpoint like a Llama- or model via Hugging Face or similar.
- Optionally, one hosted model like ChatGPT / GPT-4o-mini or Claude for comparison.

For this project, we will treat any single model as sufficient.

2.2 Data

A subset of multi-step reasoning tasks like math tasks if 150 to 200 word problems or logic problems where:

- The definite final answer is known
- Step by step solution is available and acknowledged

Each problem will be run under both Outcome and Process conditions for the same models, giving us:

- Final answers in both conditions.
- Full reasoning in the Process/Process+ condition.

2.3 Metrics for Factual Reliability

Final Answer Accuracy

- Percentage of problems where the model's final answer matches the definite truth.
- Compare Accuracy of Outcome vs. Accuracy of Process.

Error-Type Breakdown (on a labeled subset)

- Label errors into categories like arithmetic slip, misinterpretation of question, logically incorrect method.
- Compare the distribution of error types between Outcome and Process.

Hallucination Rate in Reasoning (Process only)

- For Process traces, proportion of solutions containing at least one clearly false, unsupported, or irrelevant claim e.g. "3 × 6 = 356" or random unrelated facts.

2.4 Metrics for Interpretability

To address the feedback you gave us on “how would you measure interpretability?”, We would like to propose three operational metrics:

Step Correctness Score (Local Correctness)

For a subset of problems of like 50–70, segment the explanations into steps or use numbered steps in Process+.

Label each step as:

- 0 = incorrect / unjustified
- 1 = partially correct or incomplete
- 2 = correct and well-justified

Step Correctness Score = average step score per problem, then averaged over the problems.Higher scores will indicate that the reasoning is locally valid and thus easier to trust and review.

Faithfulness / Expert Alignment Score

For problems with expert solutions, compare each model step with the corresponding expert step.

A simple rubric:

- 0 = different and incorrect/irrelevant
- 1 = similar operation but sloppy or partially wrong
- 2 = essentially same operation and correct

Optionally we can supplement manual labels with an automatic proxy like text overlap or something similar.This captures wether the model follows a known correct reasoning path, which is important if explanations are to be used as evidence in a legal setting.

Human Auditability (Clarity + Verification Effort)

Conduct a small-scale human evaluation like with myself + 1–2 volunteers as raters

For each explanation:

- **Clarity Score :** A score from 1 to 5 measuring ,how easy was it to understand and the reasoning.
- **Verification Effort:** A score from 1 to 5 measuring, how hard would it be if a reviewer to check correctness without redoing the entire solution from scratch?

Lower verification effort and higher clarity indicate more legally useful explanations.These metrics, together, give a definition of interpretability rooted in auditability and legal review rather than just “does this explanation sound nice to you mister.”

2.5 Success Criteria

The approach will be considered successful if we can:

1. Quantitatively compare Outcome vs. Process conditions on accuracy, step correctness, faithfulness, and human auditability; and
2. Provide a clear qualitative analysis of failure cases (e.g., correct final answers with garbage reasoning vs. wrong answers with good reasoning).

3. Significance

If process-based supervision yields:

- Higher factual reliability, and/or
- More faithful, locally correct, and auditable reasoning steps,

then we can say that this provides empirical support for using process oriented methods like process supervision, self-rewarding models, or deliberative alignment in high-stakes, regulated legal domains. That, in turn, strengthens arguments that:

- Explanations generated by such models can be meaningfully audited by regulators, courts, and affected users.
- Certain classes of alignment techniques are better aligned with legal values such as transparency, non-arbitrariness, and due process.

Conversely, if reasoning traces are often unfaithful or misleading, that’s equally important, why? Well because it suggests that current “explainable” LLM outputs may be legally unreliable, even if they look structured and convincing.

4. Novelty and Relation to Prior Work

This project builds on, but is distinct from, recent work:

- Process vs. outcome feedback for math problems (e.g., Uesato et al., 2022) and process-based self-rewarding models (e.g., Zhang et al., 2025) study how process feedback affects performance and training.
- RLHF and alignment frameworks (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022) focus on making models helpful and harmless via human preferences and constitutions, but often treat explanation quality more informally.

Novel aspects of this project:

- The focus is not on training a new model, but on a clean, controlled evaluation of process vs. outcome prompting or fine-tuning specifically from the viewpoint of interpretability and legal auditability.
- We propose explicit, operational metrics for interpretability, step correctness, faithfulness to expert solutions, and human verification effort that are directly connected to regulatory needs.
- The analysis will explicitly draw implications for Responsible AI and the Law, rather than staying purely in ML benchmarking space.

5. Feasibility

Data & Models:

- Public reasoning datasets like the math benchmarks are readily available and small subsets of about 150–200 items are sufficient.
- Publicly accessible models that are open-weight or hosted via API are also already usable with simple scripts, no custom training is strictly required as such.

Implementation Complexity:

- Querying APIs and storing model outputs is straightforward (Python + standard HTTP/SDK).
- Annotation schemes for step correctness, faithfulness, and clarity are simple and can be applied manually on a subset.

Rough Timeline:

- **Nov 12-20:** Finalize dataset subset, prompts, and models; implement data collection script,run both conditions and store outputs.
- **Nov 21-30:** Design and pilot annotation rubric, manually annotate step correctness, faithfulness, and clarity on a subset, compute accuracy and basic stats.
- **Dec 1-7:** Analyze results (quantitative + qualitative failure cases), draft connections to legal concepts (explainability, accountability).
- **Dec 8-13:** Write final report and prepare presentation.

Given the course timeline and the limited scope , this should be realistically achievable(in theory ofc).

6. Questions / Issues for Feedback

Interpretability Metrics:

Do you feel that the three proposed interpretability metrics the step correctness, expert faithfulness, human auditability sufficiently aligned with what you had in mind when you asked about “interpretable” reasoning? Would you prefer I pay attention to one of them more heavily? If yes, which one and why?

Model Choice & APIs:

Is it okay if we are to use a mix of one open-weight model and, maybe, one closed-source hosted model like ChatGPT or Claude for comparison, or would you prefer we stick to open-weight models only?