Configuring Large Reasoning Models using Process Mining: A Benchmark and a Case Study

Alessandro Berti[1,2], Humam Kourani[2,1], Gyunam Park[2,1], Wil M.P. van der Aalst[1,2]

[1] Process and Data Science Chair, RWTH Aachen University, Aachen, Germany
[2] Fraunhofer FIT, Sankt Augustin, Germany
{a.berti; wvdaalst}@pads.rwth-aachen.de;
{humam.kourani; gyunam.park}@fit.fraunhofer.de

Abstract. Large Reasoning Models (LRMs), a subset of Large Language Models (LLMs) trained to articulate their chain-of-thought, have shown promise in tackling complex scientific tasks. However, evaluating and configuring their reasoning processes remains underexplored. This paper leverages a process mining-specific LLM evaluation framework to propose a methodology for analyzing and configuring LRMs. We introduce an approach to extract and classify reasoning steps by type (e.g., Deductive Reasoning, or Hypothesis Generation) and effect (Positive, Indifferent, Negative) on the overall reasoning, enabling a detailed assessment of reasoning quality. From this, we derive a new benchmark, *PMLRM-Bench*, which evaluates not only the correctness of outputs but also the robustness of the reasoning process. A case study on the QwQ-32B LLM demonstrates how targeted adjustments to reasoning type frequencies can boost task-specific performance. Our results reveal distinct reasoning patterns across models and provide actionable insights for LRM configuration. This work bridges process mining and LLM evaluation, offering a scalable framework for reasoning analysis.

Keywords: Large Reasoning Models · Process Mining · Reasoning Analysis

1 Introduction

The rapid rise of Large Language Models (LLMs) has revolutionized natural language processing, excelling in tasks like text generation and question answering. Yet, for scientific reasoning—requiring transparent, logical, and verifiable processes—standard LLMs often falter. Large Reasoning Models (LRMs), a specialized LLM subset, address this by outputting their chain-of-thought, vital for scientific tasks where the reasoning process matters as much as the result. Still, evaluating and optimizing LRMs is insufficiently studied, lacking standard methods to enhance their reasoning abilities.

Analyzing the reasoning traces—textual outputs of an LRM's step-by-step thought process—helps uncover the mechanisms driving LRMs, revealing patterns, strengths, and weaknesses in their reasoning. For instance, Figure 1 illustrates the response of an LRM prompted to identify bottlenecks in a hypothetical event log, employing a diverse array of reasoning types, such as Pattern Recognition, Deductive Reasoning, Hypothesis Generation, and Ethical Reasoning, each

```
<think>
Event log timestamps show Activity A takes 1 minute, B takes 1 minute, C
takes 5 minutes, D takes 1 minute across multiple cases, suggesting C is
significantly slower than others. Pattern Recognition - PE

C consistently takes 5 minutes while other activities take 1 minute, so C is
likely causing a delay in the process flow, as it exceeds the average
activity duration. Deductive Reasoning - PE

In process mining, an activity with consistently longer durations across
cases often indicates a bottleneck, as seen in similar logs where delays
cluster around a single activity. Inductive Reasoning - PE

Delay at C could result from insufficient resources, complex tasks, or
external dependencies, with resource scarcity being the most plausible given
uniform task complexity in the log. Abductive Reasoning - IND

Adding resources to activity C could reduce its duration and improve overall
process throughput, assuming resource constraints are the primary issue.
Hypothesis Generation - PE

Log shows C has fewer assigned resources (e.g., one worker vs. three for A,
B, D), supporting the resource bottleneck hypothesis. Validation - PE

If resource data is unavailable, the conclusion about C's bottleneck may be
premature, so reconsider if C's delay is due to task complexity rather than
resources. Backtracking - IND

Reallocating resources to C could disadvantage other activities or workers
(e.g., overburdening staff), raising ethical concerns about equitable
resource distribution. Ethical or Moral Reasoning - IND

If C had not been identified as the bottleneck, increasing resources for B
or D would likely not prevent delays, as their durations are already
minimal. Counterfactual Reasoning - IND

In process mining, prioritizing resource allocation to the slowest activity
often yields the fastest throughput improvement, so focus on optimizing C
first. Heuristic Reasoning - PE

Cases with high throughput show C with additional resources, reinforcing the
hypothesis that resource allocation mitigates the bottleneck. Validation - PE

If C's delay persists despite added resources, external factors (e.g.,
system downtime) may contribute, requiring further log analysis. Abductive
Reasoning - IND
</think>

Activity C is likely a bottleneck due to resource constraints, supported by
consistent delays and validated by resource allocation patterns. Further
analysis of external factors is recommended. Conclusion - C
```

Fig. 1: Example reasoning trace from an LRM analyzing an event log for bottlenecks, showcasing a diverse set of reasoning types (Section 2.2) classified using the LLMs-as-a-Judge evaluation strategy (the methodology and choice of the judge is explained in Section 2.3). Each step, as explained in Section 3.1, is labeled as PE (positive), IND (indifferent), or NE (negative) based on its effect on the overall reasoning correctness.

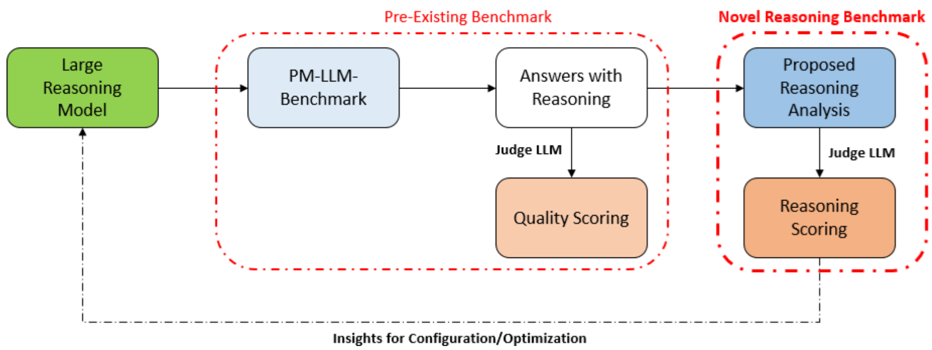

Fig. 2: Schematic overview of the reasoning analysis pipeline proposed in the paper.

labeled with its effect on correctness (Positive, Indifferent, or Negative) of the overall reasoning.

This paper leverages the PM-LLM-Benchmark v2.0, a process mining-specific LLM evaluation framework providing a rich dataset of LLM responses to complex process mining prompts `https://github.com/fit-alessandro-berti/pm-llm-benchmark` [2], to propose a novel methodology for evaluating and enhancing Large Reasoning Models (LRMs).

We introduce a reasoning analysis pipeline, schematically overviewed in Figure 2. This pipeline starts with the raw reasoning traces (textual outputs) collected from the PM-LLM-Benchmark v2.0 dataset as input. An extraction step then parses these unstructured traces, transforming them into structured objects. During this extraction, each identified reasoning step within a trace is classified by its type (e.g., Deductive Reasoning, Hypothesis Generation) and its effect on the overall correctness (Positive, Indifferent, or Negative), as demonstrated in Figure 1. Subsequently, an analysis stage processes these structured JSON traces to compute various metrics and systematically evaluate the quality and patterns of the LRM's reasoning process, going beyond mere output accuracy.

Following the reasoning analysis pipeline, we propose two key contributions to advance the evaluation and configuration of LRMs for process mining tasks:

- *Benchmark Introduction*: We present *PMLRM-Bench*, an extension of the PM-LLM-Benchmark, designed to evaluate both the correctness of LRM outputs and the robustness of their reasoning processes, providing a comprehensive assessment of reasoning quality.
- *Case Study Insights*: A case study on the `qwen-qwq-32b` model illustrates how adjusting the frequency of specific reasoning types can address weaknesses, enhancing performance on process mining tasks and offering actionable strategies for optimizing LRMs for scientific applications.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 presents our framework and the proposed benchmark. Section 4 evaluates LRM performance, while Section 5 showcases reasoning adjustments via a case study. Eventually, Section 6 summarizes findings and future steps.

2 Related Work

2.1 LLMs in Process Mining

Large Language Models (LLMs) have demonstrated potential in Business Process Management (BPM) and Process Mining (PM). Research shows LLMs can automate BPM tasks like process documentation, though challenges persist with complex structures [2, 11, 25]. Their application in process mining highlights the need for domain-specific fine-tuning [24]. Encoding process mining data into textual prompts is crucial due to LLMs' input limitations. Abstractions for representing event logs have been proposed, enabling tasks like conformance checking, though prompt engineering remains a hurdle [4]. Transforming logs into narratives for bottleneck detection emphasizes the role of structured prompts for accurate outputs [3].

Evaluation frameworks are vital for assessing LLMs in PM. Studies focusing on semantics-aware tasks find LLMs excel in contextual inference but struggle with complex tasks [24]. A benchmark for causal process reasoning notes LLMs' potential in decision-making but limitations in complex causal analysis [8]. Evaluations of LLMs on process modeling and querying show they generate correct models and answer queries effectively with structured prompts, though performance drops with ambiguous data [19, 20]. The PM-LLM-Benchmark assesses LLMs across seven PM task categories, revealing strengths in pattern recognition but weaknesses in speculative reasoning [2].

This work builds on these efforts by introducing a reasoning analysis pipeline that classifies LRM reasoning steps by type and effect, extending the PM-LLM-Benchmark to optimize LRMs for scientific tasks through detailed evaluation.

2.2 Reasoning Types

A taxonomy of ten reasoning types is adopted to evaluate LRMs for scientific tasks like process mining: Pattern Recognition (PR), Deductive Reasoning (DR), Inductive Reasoning (IR), Abductive Reasoning (AR), Hypothesis Generation (HG), Validation (V), Backtracking (B), Ethical or Moral Reasoning (ER), Counterfactual Reasoning (CR), and Heuristic Reasoning (HR). These types draw from cognitive and computational theories, including deduction, induction, and abduction [7,12,22], meta-cognitive strategies like hypothesis generation and validation [15,23], and socio-technical considerations such as heuristics and ethical checks [5, 10, 21]. This taxonomy supports comprehensive analysis of LRM reasoning processes.

Each reasoning type is defined with its theoretical basis and relevance to process mining: PR detects patterns and anomalies [5]; DR derives logical conclusions [14]; IR generalizes from data [13]; AR infers plausible explanations [9]; HG formulates testable hypotheses [16]; V verifies reasoning steps [23]; B revises prior steps [17]; ER ensures ethical fairness [18]; CR explores causal scenarios [21]; and HR applies efficient rules [10]. This taxonomy enables the dissection of LRM reasoning traces, as shown in Figure 1, where steps are classified by type and effect (Positive, Indifferent, or Negative) to assess reasoning quality.

2.3 LLMs-as-a-Judge Evaluation Strategy

The LLMs-as-a-Judge paradigm is employed to evaluate LRM reasoning processes, excelling in classifying reasoning traces by type and effect, surpassing traditional metrics. LLMs leverage their natural language understanding to assess response quality beyond correctness. Studies show LLMs align closely with human judgments in reasoning evaluation, enabling rapid processing of large datasets like the PM-LLM-Benchmark v2.0 [27]. Their ability to evaluate coherence and logical consistency is crucial for classifying reasoning types like Deductive Reasoning or Hypothesis Generation [26]. With appropriate prompting, LLMs handle domain-specific tasks effectively, achieving high-precision evaluations using detailed rubrics [6]. LLMs also outperform rule-based systems in evaluating multi-step reasoning [1].

3 Methodology

This section presents the core framework for analyzing and benchmarking Large Reasoning Models (LRMs) using the PM-LLM-Benchmark v2.0 dataset, culminating in the *PMLRM-Bench* benchmark available at
`https://github.com/fit-alessandro-berti/pmllmbench-lrms-reasoning-analysis`.
The approach leverages reasoning traces—textual outputs from LRMs that articulate their chain-of-thought—to assess not only the correctness of answers but also the quality and structure of the reasoning process. By extracting, classifying, and analyzing these traces, we establish a systematic method to evaluate and configure LRMs for scientific tasks, particularly within the process mining domain.

3.1 Reasoning Trace Extraction

The foundation of our framework lies in extracting structured data from the unstructured textual responses of LRMs, as provided by the PM-LLM-Benchmark v2.0 dataset (`https://github.com/fit-alessandro-berti/pm-llm-benchmark/`). This dataset contains a diverse set of process mining prompts spanning seven categories, such as contextual understanding and conformance checking, along with corresponding LRM outputs. Each response, referred to as a reasoning trace, is parsed into a sequence of individual reasoning steps, which are then stored as JSON objects. These steps are classified by *Gemini-2.5-Pro-Preview-03-25*, currently the latest model of the Google Gemini family, serving as the judge LLM to determine their reasoning type and effect. Every step is assigned a name that combines its reasoning type—such as PR, DR, IR, AR, HG, V, B, ER, CR, or HR—and its effect on correctness, labeled as **PE** for positive effect, **IND** for indifferent or neutral effect, or **NE** for negative effect. Alongside the name, each JSON object includes a text field containing the specific excerpt from the reasoning trace that corresponds to that step. At the end of each trace, a special conclusion entry is appended to indicate the overall correctness of the reasoning process, marked as Conclusion - **C** for correct, Conclusion - **PC** for partially correct, or Conclusion - **W** for wrong, also judged by *Gemini-2.5-Pro-Preview-03-25*. This conclusion entry does not include a text snippet, serving solely as a summary of the trace's outcome. Through this extraction process, we transform raw LRM outputs into a standardized format suitable for detailed analysis.

3.2 Metrics

The analysis pipeline processes structured JSON logs extracted in the previous subsection by aggregating reasoning steps and conclusions. The goal of the composite score S is to reward correct conclusions (**C**) and positive-effect steps (**PE**), which advance reasoning accuracy, while penalizing partially correct (**PC**) and wrong (**W**) conclusions, as well as indifferent (**IND**) and negative-effect (**NE**) steps, which either do not contribute to or detract from correctness.

The pipeline first counts conclusion outcomes (**C**, **PC**, **W**) to assess overall accuracy. It then aggregates step effects (**PE**, **IND**, **NE**) to evaluate reasoning quality. Frequencies of reasoning types (e.g., Pattern Recognition, Deductive Reasoning) are calculated, including the proportion of **PE** steps per type. The composite score is defined as:

$$S = (100 \cdot \mathbf{C} - 100 \cdot \mathbf{PC} - 200 \cdot \mathbf{W}) + (10 \cdot \mathbf{PE} - 1 \cdot \mathbf{IND} - 20 \cdot \mathbf{NE}),$$

where weights reflect the relative impact of each component. Correct conclusions (**C**) receive a high positive weight (+100) to emphasize accurate outcomes. Wrong conclusions (**W**) are heavily penalized (-200) due to their detrimental effect on reliability. Partially correct conclusions (**PC**) receive a moderate penalty (-100) to account for partial accuracy. For steps, positive-effect steps (**PE**) are rewarded (+10) for advancing reasoning, negative-effect steps (**NE**) are penalized (-20) for introducing errors, and indifferent or redundant steps (**IND**) receive a small penalty (-1) for their neutral or inefficient contribution. These weights

Table 1: Model performance scores and reasoning step statistics for evaluated LRMs.

Model	**S**	**C**	**PC**	**W**	**PE**	**IND**	**NE**	**PM-LLM-B. Score**
Grok-3-thinking-20250221	**14459**	45	1	0	1039	131	10	**39.8**
qwen-qwq-32b-nostepbystep	**13281**	44	2	0	945	109	13	**36.9**
exaone-deep7.8b-fp16	**12381**	44	2	0	851	109	11	**30.2**
DeepSeek-R1-671B-HB	**12303**	45	1	0	830	97	15	**36.8**
Perplexity-R1-1776	**12095**	45	1	0	779	75	1	**32.1**
qwen-qwq-32b-stepbystep	**12027**	44	2	0	848	113	27	**35.8**
nvidia-nemotron-super-49b	**11856**	46	0	0	747	74	7	**36.4**
QwenQwQ-32B-Preview	**11725**	44	1	1	783	65	7	**28.9**
exaone-deep32b-fp16	**11187**	43	3	0	779	103	25	**31.6**
exaone-deep2.4b-fp16	**10481**	40	3	3	806	139	27	**23.5**
R1-Distill-Qwen-14B	**10433**	42	4	0	697	97	12	**27.8**
R1-Distill-Llama-70B	**10153**	40	5	1	717	57	13	**28.6**
R1-Distill-Qwen-32B	**9970**	42	4	0	650	50	14	**30.7**
DeepSeek-R1-Zero	**9765**	46	0	0	523	45	1	**29.9**
R1-Distill-Llama-8B	**7307**	33	10	3	726	153	75	**20.6**
R1-Distill-Qwen-7B	**3811**	26	16	4	578	149	101	**16.8**
R1-Distill-Qwen-1.5B	**-7724**	15	12	19	472	184	438	**9.9**

were tuned to balance the emphasis on correct outcomes and robust reasoning processes.

For example, consider the reasoning trace in Figure 1 with 12 steps and a conclusion: 6 **PE** steps (PR, DR, IR, HG, V, HR), 5 **IND** steps (two AR, V, ER, CR), and no **NE** steps. The conclusion is $\mathbf{C} = 1$, $\mathbf{PC} = 0$, $\mathbf{W} = 0$. Limited to that answer, the score is:

$$S = (100 \cdot 1 - 100 \cdot 0 - 200 \cdot 0) + (10 \cdot 6 - 1 \cdot 5 - 20 \cdot 0) = 100 + (60 - 5) = 155.$$

4 Benchmark

This section proposes a benchmark evaluating Large Reasoning Models (LRMs) that do not artificially obscure their chain-of-thought. Models that obscure their reasoning process (such as OpenAI `o1-2024-12-17`) are excluded. The analysis, summarized in Tables 1 through 4, assesses model performance across overall scores, reasoning type distributions, correctness rates, and task-specific patterns.

Results: Table 1 reports composite scores (S) and reasoning step effects for the proposed *PMLRM-Bench*, alongside PM-LLM-Benchmark scores. Leading models achieve high S with predominantly positive-effect steps, excelling in structured reasoning tasks. Weaker models show more errors and negative-effect steps, reflecting challenges in logical coherence. Intermediate models balance accuracy and efficiency, indicating effective reasoning strategies. A strong positive correlation (Pearson's $r = 0.89$) between S and PM-LLM-Benchmark scores suggests that robust reasoning processes align closely with overall task performance.

Table 2 highlights how models allocate reasoning effort. Top performers favor deductive reasoning for structured tasks while balancing hypothesis generation and validation. Less effective models over-rely on speculative reasoning, often lacking sufficient validation, which impacts performance.

Correctness rates per reasoning type, shown in Table 3, reveal strengths in foundational reasoning like pattern recognition and deduction among strong models. Weaker models struggle across reasoning types, particularly in logical and exploratory tasks, reflecting inconsistent accuracy.

Category-Reasoning Correlation: Table 4 details reasoning steps by PM-LLM-Benchmark task category. DR and HG dominate, aligning with process mining challenges like contextual understanding and model generation. Fairness-related tasks show increased ER, while CR remains minimal, indicating limited "what-if" analysis.

Validation: To validate the robustness of reasoning step classifications by *Gemini-2.5-Pro-Preview-03-25*, we used *ChatGPT-4o-latest-2025-03-26* as a second judge to evaluate their correctness. This review assessed the alignment of each step's text with its assigned reasoning type and effect in the PM-LLM-Benchmark v2.0 dataset. Results showed 82.40% full agreement (Y), 13.01% partial agreement (P), and 4.59% no agreement (N), indicating high consistency in the primary classifications.

Table 2: Percentage distribution of reasoning types over total steps for each model.

Model	PR	DR	IR	AR	HG	V	B	ER	CR	HR
Grok-3-thinking-20250221	13.5	31.9	2.8	2.1	16.8	17.9	3.9	1.3	1.4	**8.5**
qwen-qwq-32b-nostepbystep	13.5	33.5	1.5	1.6	20.1	17.5	5.0	**2.2**	0.9	4.2
exaone-deep7.8b-fp16	14.8	33.2	**3.6**	2.9	19.4	14.4	5.3	0.8	0.9	4.7
DeepSeek-R1-671B-HB	13.0	32.0	2.2	**4.5**	16.9	**19.2**	2.1	1.0	1.1	8.2
Perplexity-R1-1776	16.5	28.2	2.9	2.7	20.4	16.6	3.6	1.3	1.5	6.3
qwen-qwq-32b-stepbystep	15.9	33.1	2.4	1.0	20.3	16.2	4.5	1.3	0.6	4.7
nvidia-nemotron-super-49b	16.2	35.4	1.0	1.9	17.4	18.1	2.9	1.0	0.7	5.4
QwenQwQ-32B-Preview	14.6	34.6	3.3	2.9	15.3	17.5	1.9	1.6	**1.8**	6.4
exaone-deep32b-fp16	12.8	32.1	2.5	2.6	17.3	17.9	5.6	1.4	0.8	6.9
exaone-deep2.4b-fp16	15.5	30.8	1.6	2.2	20.1	15.4	6.5	1.2	1.2	5.5
R1-Distill-Qwen-14B	18.6	30.5	2.6	1.6	19.7	15.9	3.3	0.7	0.7	6.2
R1-Distill-Llama-70B	16.1	31.4	1.9	2.9	21.6	14.2	3.3	0.9	0.9	6.7
R1-Distill-Qwen-32B	15.5	34.2	3.5	2.4	17.8	14.8	4.5	0.6	0.4	6.3
DeepSeek-R1-Zero	**18.8**	**40.1**	3.5	3.0	11.6	15.1	0.5	1.2	0.7	5.4
R1-Distill-Llama-8B	14.7	28.9	2.0	1.2	23.9	15.0	6.0	1.4	0.8	6.2
R1-Distill-Qwen-7B	15.8	27.1	2.5	0.6	21.1	18.7	5.6	1.1	1.0	6.5
R1-Distill-Qwen-1.5B	14.6	25.3	0.9	1.8	**29.2**	10.2	**9.5**	0.7	0.7	6.9

Table 3: Percentage of correctness per reasoning type for each model.

Model	PR	DR	IR	AR	HG	V	B	ER	CR	HR
Grok-3-thinking-20250221	97.5	96.3	90.9	68.0	72.2	94.3	76.1	**100.0**	47.1	75.0
qwen-qwq-32b-nostepbystep	95.1	96.9	**100.0**	88.2	77.2	89.3	60.4	**100.0**	60.0	82.2
exaone-deep7.8b-fp16	93.8	96.9	85.7	82.1	74.5	95.0	74.5	**100.0**	66.7	56.5
DeepSeek-R1-671B-HB	95.9	97.3	85.7	90.5	69.8	95.6	70.0	88.9	40.0	70.1
Perplexity-R1-1776	96.5	97.5	96.0	78.3	**88.5**	93.0	67.7	90.9	30.8	83.3
qwen-qwq-32b-stepbystep	93.6	93.6	**100.0**	60.0	66.7	91.9	52.3	**100.0**	**83.3**	**93.5**
nvidia-nemotron-super-49b	97.0	96.2	87.5	62.5	79.2	92.0	75.0	**100.0**	**83.3**	77.8
QwenQwQ-32B-Preview	97.6	97.6	**100.0**	76.0	85.5	90.7	87.5	92.9	40.0	80.0
exaone-deep32b-fp16	91.4	95.5	91.3	91.7	73.9	87.7	66.7	92.3	57.1	69.8
exaone-deep2.4b-fp16	91.4	91.6	**100.0**	76.2	68.7	88.7	69.8	**100.0**	25.0	67.9
R1-Distill-Qwen-14B	96.0	93.5	95.2	**100.0**	70.4	88.3	77.8	**100.0**	66.7	68.0
R1-Distill-Llama-70B	98.4	95.1	**100.0**	82.6	84.1	92.0	**88.5**	**100.0**	71.4	79.2
R1-Distill-Qwen-32B	96.4	95.9	96.0	82.4	87.4	**96.2**	50.0	**100.0**	66.7	80.0
DeepSeek-R1-Zero	**99.1**	**98.7**	**100.0**	88.2	78.8	83.7	0.0	**100.0**	25.0	80.6
R1-Distill-Llama-8B	92.9	88.8	**100.0**	81.8	61.0	77.6	42.1	**100.0**	12.5	59.3
R1-Distill-Qwen-7B	84.7	79.0	61.9	**100.0**	56.0	67.7	45.7	**100.0**	25.0	68.5
R1-Distill-Qwen-1.5B	72.5	46.6	60.0	65.0	21.6	50.0	34.6	87.5	25.0	50.0

Table 4: Percentage distribution of reasoning steps by category across all models.

PM-LLM-B. Category	PR	DR	IR	AR	HG	V	B	ER	CR	HR
cat01 (Context Understanding)	18.6	28.8	2.9	2.7	17.5	18.1	4.2	0.0	1.1	6.1
cat02 (Conformance Checking)	17.2	**38.1**	2.4	3.6	12.2	15.7	3.8	0.0	0.8	6.2
cat03 (Process Modeling)	12.7	30.1	2.2	1.0	**24.1**	15.9	7.6	0.0	0.4	6.0
cat04 (Process Querying)	8.3	32.9	1.7	2.9	20.4	19.9	4.1	0.0	1.5	**8.1**
cat05 (Hypotheses Generation)	18.0	28.9	1.6	0.7	**26.5**	15.2	2.2	0.1	0.3	6.6
cat06 (Fairness)	18.4	29.7	4.1	1.9	18.8	9.9	3.1	**9.5**	1.9	2.7

5 Case Study

In this case study, we explore how adjusting the frequency of specific reasoning types in Large Reasoning Models (LRMs) can enhance their performance, using the QwQ-32B model—a 32-billion parameter LRM from the Qwen Team at Alibaba Cloud, refined via reinforcement learning for complex problem-solving (`https://huggingface.co/Qwen/QwQ-32B`). We leverage the PM-LLM-Benchmark v2.0 benchmark to evaluate variants of the QwQ-32B model, modified through additional system prompts to emphasize or suppress reasoning types such as HG and ER The baseline model and its variants were assessed using the *PMLRM-Bench* framework, with the scores shown in Table 5.

As shown in Table 6, the baseline `qwq-32b` model exhibits a balanced reasoning profile, with notable strengths in DR and HG Adjustments to this baseline reveal varying impacts: increasing ER boosts the reasoning quality, particularly in fairness-related tasks, while enhancing HG aids exploratory reasoning. Conversely, reducing HG or ER can streamline reasoning but may compromise depth in specific contexts, as shown in Table 5.

To further validate these insights, we executed the PM-LLM-Benchmark on these custom variants, with results presented in Table 7. The benchmark evaluates performance across seven categories: Contextual Understanding, Conformance Checking, Process Modeling, Process Querying, Hypotheses Generation, and Fairness.

Table 5: Performance scores and reasoning step statistics for QwQ-32B variants with adjusted system prompts.

Model	S	C	PC	W	PE	IND	NE	PM-LLM-B. Score
qwq-32b-moremoral	**14218**	44	2	0	1067	112	27	**35.4**
qwq-32b-morehypgen	**13372**	44	2	0	980	108	26	**37.1**
qwq-32b-lessvalidbacktr	**13310**	43	3	0	956	130	6	**36.4**
qwq-32b	**13281**	44	2	0	945	109	13	**36.9**
qwq-32b-lessmoral	**12584**	45	1	0	846	76	10	**35.7**
qwq-32b-lesshypgen	**11856**	39	6	0	913	94	24	**36.3**

Table 6: Percentage distribution of reasoning types over total steps for QwQ-32B variants, with percentage change relative to the baseline *qwq-32b* on the next line.

Model	PR	DR	IR	AR	HG	V	B	ER	CR	HR
-moremoral	13.3 (-1.5%)	34.0 (+1.5%)	1.5 (0.0%)	1.6 (0.0%)	18.0 (-10.4%)	18.2 (+4.0%)	3.9 (-22.0%)	3.2 (+45.5%)	0.4 (-55.6%)	6.0 (+42.9%)
-morehypgen	14.7 (+8.9%)	32.1 (-4.2%)	1.1 (-26.7%)	1.4 (-12.5%)	23.8 (+18.4%)	15.7 (-10.3%)	5.4 (+8.0%)	1.5 (-31.8%)	0.5 (-44.4%)	3.7 (-11.9%)
-lessvalidbacktr	14.7 (+8.9%)	33.1 (-1.2%)	2.3 (+53.3%)	1.6 (0.0%)	21.5 (+7.0%)	16.6 (-5.1%)	3.8 (-24.0%)	0.8 (-63.6%)	1.1 (+22.2%)	4.5 (+7.1%)
qwq-32b	13.5 (0.0%)	33.5 (0.0%)	1.5 (0.0%)	1.6 (0.0%)	20.1 (0.0%)	17.5 (0.0%)	5.0 (0.0%)	2.2 (0.0%)	0.9 (0.0%)	4.2 (0.0%)
-lessmoral	13.6 (+0.7%)	33.0 (-1.5%)	1.2 (-20.0%)	1.6 (0.0%)	22.1 (+10.0%)	17.7 (+1.1%)	5.5 (+10.0%)	1.0 (-54.5%)	0.4 (-55.6%)	3.9 (-7.1%)
-lesshypgen	14.0 (+3.7%)	34.7 (+3.6%)	1.6 (+6.7%)	1.6 (0.0%)	17.0 (-15.4%)	19.7 (+12.6%)	4.8 (-4.0%)	1.2 (-45.5%)	0.9 (0.0%)	4.7 (+11.9%)

Table 7: PM-LLM-Benchmark results for QwQ-32B variants, evaluated with 1-shot prompting.

Model	PM-LLM-B. Score	Contextual Underst.	Conf. Checking	Process Modeling	Process Querying	Hypotheses Generation	Fairness
qwq-32b-morehypgen	**37.1**	**6.3**	**7.1**	5.4	4.4	4.9	5.1
qwq-32b	36.9	5.7	6.4	**6.0**	4.5	4.9	5.7
qwq-32b-lessvalidbacktr	36.4	5.5	6.6	5.1	4.6	4.8	**6.0**
qwq-32b-lesshypgen	36.3	5.5	6.6	5.9	**4.8**	4.3	5.2
qwq-32b-lessmoral	35.7	5.9	6.7	5.0	4.3	**5.0**	5.1
qwq-32b-moremoral	35.4	5.0	6.5	5.7	**4.8**	4.3	5.3

Key observations emerge from these results: the variant with more HG achieves the highest score in the PM-LLM-Benchmark (37.1), maintaining a high score in the Hypotheses Generation category (4.9), outperforming the baseline in Contextual Understanding (6.3) and Conformance Checking (7.1), suggesting broad benefits in exploratory tasks. However, the variant with less HG, displaying reduced performance in Hypotheses Generation (4.3), overall underperforms, indicating that limiting HG weakens performance on tasks requiring creative exploration. Similarly, the variant with less ER fares worse than the baseline (5.1 versus 5.7) in Fairness, underscoring that suppressing ER diminishes effectiveness in fairness-related tasks. The variant with more ER also does not improve Fairness (5.3 versus 5.7) over the baseline, suggesting that excessive moral reasoning may introduce complexity without proportional gains.

These findings highlight trade-offs in LRM configuration. Enhancing HG improves performance, aligning with exploratory needs, while reducing it hampers performance. Adjusting ER shows mixed results: both variants fail to yield improvements in the PM-LLM-Benchmark score, possibly due to overcomplication in the variant with more ER

This case study[3][4] demonstrates the *PMLRM-Bench* framework's ability to guide targeted LRM adjustments, optimizing performance for specific process mining challenges.

6 Conclusion

This paper introduced a novel framework for configuring and evaluating LRMs through a structured analysis of reasoning traces. While demonstrated using the PM-LLM-Benchmark v2.0 dataset, our proposed approach is generalizable and not inherently bound to this benchmark. The reasoning analysis pipeline can readily be initiated from any reasoning trace, facilitating applicability across various domains requiring structured and transparent reasoning evaluation.

Through a detailed analysis, distinct reasoning patterns emerged among different LRMs, emphasizing the importance of tailored configurations. The case

[3]Results of the modified variants in *PMLRM-Bench*: `https://github.com/fit-alessandro-berti/pmllmbench-lrms-reasoning-analysis/tree/ft-case-study`

[4]Results of the modified variants on the PM-LLM-Benchmark: `https://github.com/fit-alessandro-berti/pm-llm-benchmark/tree/ft-case-study`

study on `qwen-qwq-32b` variants further showcased the efficacy of targeted adjustments in reasoning types, such as boosting hypothesis generation to enhance exploratory reasoning or fine-tuning ethical considerations, highlighting both improvements and potential trade-offs.

However, our methodology is subject to certain limitations. Primarily, its effectiveness depends significantly on the accuracy of the classification by the judge LLM, introducing potential biases or errors in the reasoning step categorization. Additionally, while the structured reasoning approach enhances transparency, it may not fully capture nuanced or implicit reasoning processes inherent in LLM processing (i.e., the latent space).

Future research could focus on refining the classification accuracy through ensemble methods or human-in-the-loop approaches and extending the pipeline's applicability to broader sets of reasoning-intensive tasks, beyond process mining, further validating its robustness and general utility.

References

1. Bai, Y., Ying, J., Cao, Y., et al., X.L.: Benchmarking foundation models with language-model-as-an-examiner. Advances in Neural Information Processing Systems **36**, 78142–78167 (2023)
2. Berti, A., Kourani, H., van der Aalst, W.: PM-LLM-Benchmark: Evaluating large language models on process mining tasks. In: International Conference on Process Mining. pp. 610–623. Springer (2024)
3. Berti, A., Qafari, M.: Leveraging large language models (LLMs) for process mining (technical report). arXiv preprint arXiv:2307.12701 (2023)
4. Berti, A., Schuster, D., van der Aalst, W.: Abstractions, scenarios, and prompt definitions for process mining with LLMs: A case study. In: International Conference on Business Process Management. pp. 427–439. Springer (2023)
5. Bishop, C., Nasrabadi, N.: Pattern recognition and machine learning, vol. 4. Springer (2006)
6. Chiang, C., Chen, W., Kuan, C., Yang, C., Lee, H.: Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. arXiv preprint arXiv:2407.05216 (2024)
7. Flach, P.: Abduction and induction: Syllogistic and inferential perspectives. In: Abductive and Inductive Reasoning Workshop Notes. pp. 31–35. University of Bristol (1996)
8. Fournier, F., Limonad, L., Skarbovsky, I.: Towards a Benchmark for Causal Business Process Reasoning with LLMs. In: International Conference on Business Process Management. pp. 233–246. Springer (2024)
9. Frankfurt, H.: Peirce's notion of abduction. The Journal of Philosophy **55**(14), 593–597 (1958)
10. Gigerenzer, G., Todd, P.: Fast and frugal heuristics: The adaptive toolbox. In: Simple heuristics that make us smart, pp. 3–34. Oxford University Press (1999)
11. Grohs, M., Abb, L., Elsayed, N., Rehse, J.: Large language models can accomplish business process management tasks. In: International Conference on Business Process Management. pp. 453–465. Springer (2023)
12. Holland, J.: Induction: Processes of inference, learning, and discovery. MIT press (1986)
13. Holland, P., Bowskill, I., Bailey, A.: Adaptors and innovators: Selection versus induction. Psychological Reports **68**(3_suppl), 1283–1290 (1991)

14. Johnson-Laird, P.: Deductive reasoning. Annual review of psychology **50**(1), 109–135 (1999)
15. Klahr, D., Dunbar, K.: Dual space search during scientific reasoning. Cognitive science **12**(1), 1–48 (1988)
16. Klahr, D., Fay, A., Dunbar, K.: Heuristics for scientific experimentation: A developmental study. Cognitive psychology **25**(1), 111–146 (1993)
17. Knuth, D.: The art of computer programming, vol. 3. Pearson Education (1997)
18. Kohlberg, L.: Stages of moral development as a basis for moral education. Center for Moral Education, Harvard University Cambridge (1971)
19. Kourani, H., Berti, A., Hennrich, J., Kratsch, W., Weidlich, R., Li, C., Arslan, A., Schuster, D., van der Aalst, W.: Leveraging Large Language Models for Enhanced Process Model Comprehension. arXiv preprint arXiv:2408.08892 (2024)
20. Kourani, H., Berti, A., Schuster, D., van der Aalst, W.: Evaluating Large Language Models on Business Process Modeling: Framework, Benchmark, and Self-Improvement Analysis. arXiv preprint arXiv:2412.00023 (2024)
21. Pearl, J.: Causality. Cambridge university press (2009)
22. Peirce, C.: The collected papers of charles s. Peirce **8**, 1–6 (1931)
23. Prasad, A., Saha, S., Zhou, X., Bansal, M.: Receval: Evaluating reasoning chains via correctness and informativeness. arXiv preprint arXiv:2304.10703 (2023)
24. Rebmann, A., Schmidt, F., Glavaš, G., van Der Aa, H.: Evaluating the ability of LLMs to solve semantics-aware process mining tasks. In: 2024 6th International Conference on Process Mining (ICPM). pp. 9–16. IEEE (2024)
25. Vidgof, M., Bachhofner, S., Mendling, J.: Large language models for business process management: Opportunities and challenges. In: International Conference on Business Process Management. pp. 107–123. Springer (2023)
26. Wang, Z., Pang, Y., Lin, Y.: Large language models are zero-shot text classifiers. arXiv preprint arXiv:2312.01044 (2023)
27. Zheng, L., Chiang, W., Sheng, Y., et al., S.Z.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems **36**, 46595–46623 (2023)