Supplementary Material for "LLMs that Understand Processes: Instruction-tuning for Semantics-Aware Process Mining"

Vira Pyrih
Faculty of Computer Science
University of Vienna
Vienna, Austria
a12228590@unet.univie.ac.at

Adrian Rebmann
SAP Signavio
Berlin, Germany
adrian.rebmann@sap.com

Han van der Aa
Faculty of Computer Science
University of Vienna
Vienna, Austria
han.van.der.aa@univie.ac.at

***Abstract*—This supplementary material investigates the performance variability of instruction-tuned Large Language Models (LLMs) across different process domains. Focusing on anomaly detection, prediction, and process discovery tasks, the analysis shows that model performance is substantially influenced by domain-specific characteristics such as process structure, standardization, and representation in training data. Notably, performance in domains like logistics and education is consistently higher, likely due to their well-defined processes or better data coverage. The findings highlight the importance of domain diversity in instruction-tuning to enhance generalization and robustness across various contexts.**

Domain Performance Analysis of Instruction-Tuned LLMs

This analysis investigates how the performance of instruction-tuned Large Language Models (LLMs) varies across different process domains. We begin by outlining the role and relevance of domains in process mining tasks, followed by an overview of the domain distribution within the evaluation datasets. A detailed analysis of model performance per domain is then presented, focusing on anomaly detection, prediction, and discovery tasks.

Domains like healthcare, finance, or logistics vary in how structured, standardized, or flexible their processes are, factors that can affect both model generalization and the reliability of specific tasks. Moreover, many LLMs are pre-trained on corpora with uneven domain coverage, potentially favoring more common or well-documented domains like IT or finance. For these reasons, it is important to analyze how instruction-tuned models behave across different domains.

A. Domain Distribution in Evaluation Data

To assess domain-specific performance, typical domain labels were first assigned to each process in the evaluation data using a zero-shot classification model applied to the set of activities. Only domain predictions with confidence greater than 30% were considered, and each label represents a unique process model. The distribution of domains across all samples is shown in Figure 1.

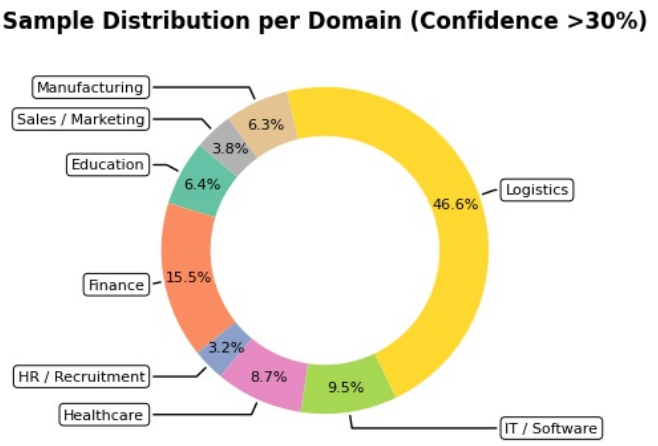

Fig. 1. Distribution of samples considering unique process models. Only samples with confidence score >30% for the assigned label were included.

There is a clear dominance of the logistics domain, which accounts for 46.6% of the data, followed by finance (15.5%), IT/software (9.5%), and healthcare (8.7%). Less represented domains such as education, manufacturing, and sales/marketing make up smaller portions of the dataset. It is important to note that this distribution reflects unique process models; for tasks where individual samples are derived from these models, the actual domain proportions within those specific task datasets may differ slightly.

B. Domain-Specific Performance

In the following, we detail how domain-specific performance varies across different tasks. Each subplot in Figures 2, 4, 5 shows the performance delta per domain relative to the task's overall average score (macro F_1 for anomaly detection and prediction, Fitness for discovery). Positive values indicate

above-average performance, and negative values signal underperformance. The figures also incorporate the percentage of domain samples relative to all samples in the task (indicated by "N: X.X%"). Domains that contribute less than 3% of the total samples have their "N" value highlighted in blue in the figures, cautioning against overinterpreting these scores due to potential instability from small sample sizes.

1) Anomaly Detection and Next Activity Prediction Tasks: Performance in anomaly detection and next activity prediction tasks using the Llama model is detailed in Figure 2 and the Mistral model in Figure 3.

a) Llama Model (Figure 2): For *Activity-level Semantic Anomaly Detection (A-SAD)*, Llama performs best in logistics (macro F_1: 0.59), likely due to the structured nature of these processes. Performance drops notably in IT/software (macro F_1: 0.47), sales/marketing (0.49), and HR/recruitment (0.51), which are more flexible and variable in activity sequences.

In *Trace-level Semantic Anomaly Detection (T-SAD)*, finance and manufacturing show the lowest results (macro F_1: 0.44 and 0.45), while healthcare (0.53) and education (0.50) score above average, with healthcare being an outlier relative to the general trend.

For *Semantic Next Activity Prediction (S-NAP)*, logistics again leads (macro F_1: 0.72), while IT/software and healthcare score lower. Overall, Llama's best domains vary by task, with logistics and education often among the strongest, but performance remains inconsistent across domains.

b) Mistral Model (Figure 3): For *A-SAD*, Mistral achieves top scores in education, logistics, and manufacturing (macro F_1: 0.70 each), with generally even results across domains. Unlike Llama, it avoids steep drops in IT/software and sales/marketing, indicating better robustness in less structured areas.

In *T-SAD*, education (macro F_1: 0.69) and manufacturing (0.66) perform best, while healthcare drops in line with the general domain trend. This contrasts with Llama's unusually high healthcare score for T-SAD. Mistral also avoids Llama's sharp dips in finance and manufacturing.

For *S-NAP*, Mistral scores highest in HR/recruitment (macro F_1: 0.92), manufacturing (0.91), and logistics (0.91), with competitive results elsewhere. Domain fluctuations are present but less pronounced than Llama's.

In summary, Mistral shows greater domain stability in A-SAD and S-NAP, consistently strong results in structured domains, and avoids some of Llama's sharp underperformances, though healthcare in T-SAD remains challenging.

2) Discovery Tasks: Domain-specific analysis for process discovery tasks is evaluated across both Llama (Figure 4) and Mistral (Figure 5) models. Performance is measured by the fitness score delta relative to the overall average per task.

a) Llama Model (Figure 4): For *Semantic Directly-Follows Graph Discovery (S-DFD)*, Llama shows notable variation across domains. The model performs best in education (Fitness: 0.76) and logistics (0.74), both scoring above average. Other domains, especially healthcare and HR/recruitment, are below the average, suggesting challenges in generalizing the DFG structure in less straightforward environments. For instance, healthcare processes often vary considerably between patients, making them less predictable. Similarly, HR workflows are highly dependent on context and often include ad hoc decisions. A similar trend appears in *Semantic Process Tree Discovery (S-PTD)*: performance on logistics tasks is best with a fitness of 0.71, while domains like healthcare (0.63) and sales/marketing (0.66) underperform. Interestingly, HR/recruitment, one of the least performing domains in S-DFD, delivers the highest results in S-PTD with Llama, suggesting a task-dependent sensitivity to the domain.

b) Mistral Model (Figure 5): For Mistral, differences in fitness scores are also visible. In S-DFD, the model performs best on education (Fitness: 0.80) and logistics (0.81), both of which likely benefit from more structured and repeatable process patterns. In contrast, healthcare (0.66) is significantly below the average. In S-PTD, Mistral continues to perform well in logistics (0.78), while scores for healthcare (0.71) and sales/marketing (0.71) remain slightly below average. Notably, Mistral's performance in domains like finance and IT/software is close to the average, indicating that its larger capacity may help generalize across a wider range of process types. Despite this, the model still struggles slightly in domains where implicit decisions and unstructured paths dominate.

C. Overall Observations on Domain Impact

The analysis of discovery tasks confirms that logistics and education regularly appear as the most LLM-compatible domains for process model generation, benefiting from either more structured process patterns or potentially better representation in underlying data. Overall, across all tasks, domain characteristics such as process complexity, structuredness, and frequency in the training data considerably influence model performance. This highlights the importance of considering domain diversity and representation during instruction-tuning to ensure robust generalization across various contexts.

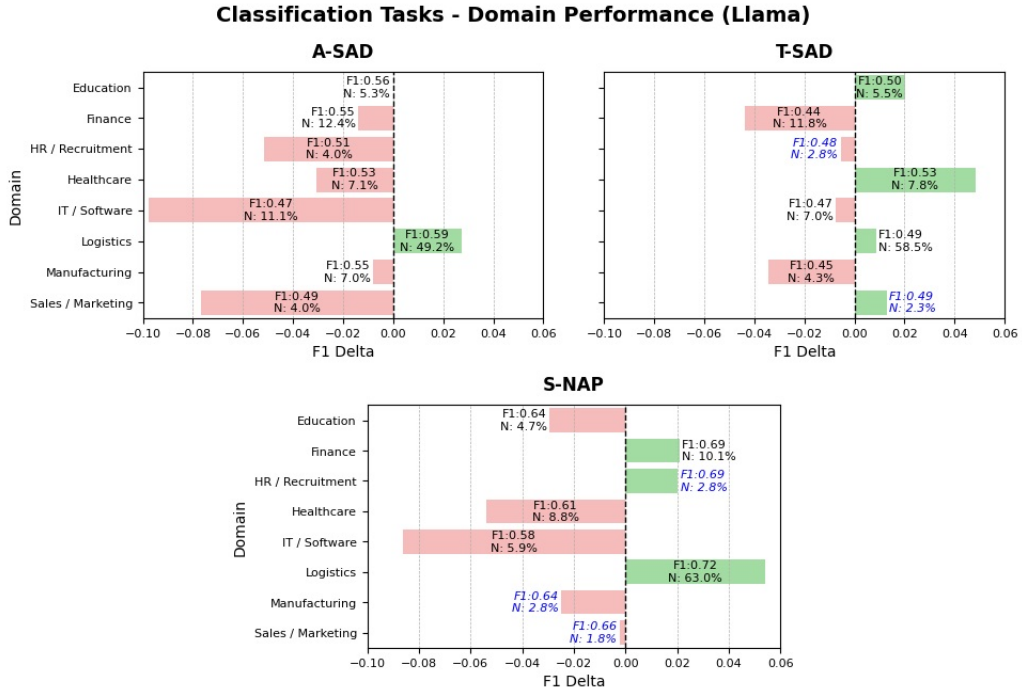

Fig. 2. Domain performance of Llama IT on the A-SAD, T-SAD and S-NAP tasks, relative to average performance. Domains with representation of less than 3% of total samples have their "N" value highlighted in blue.

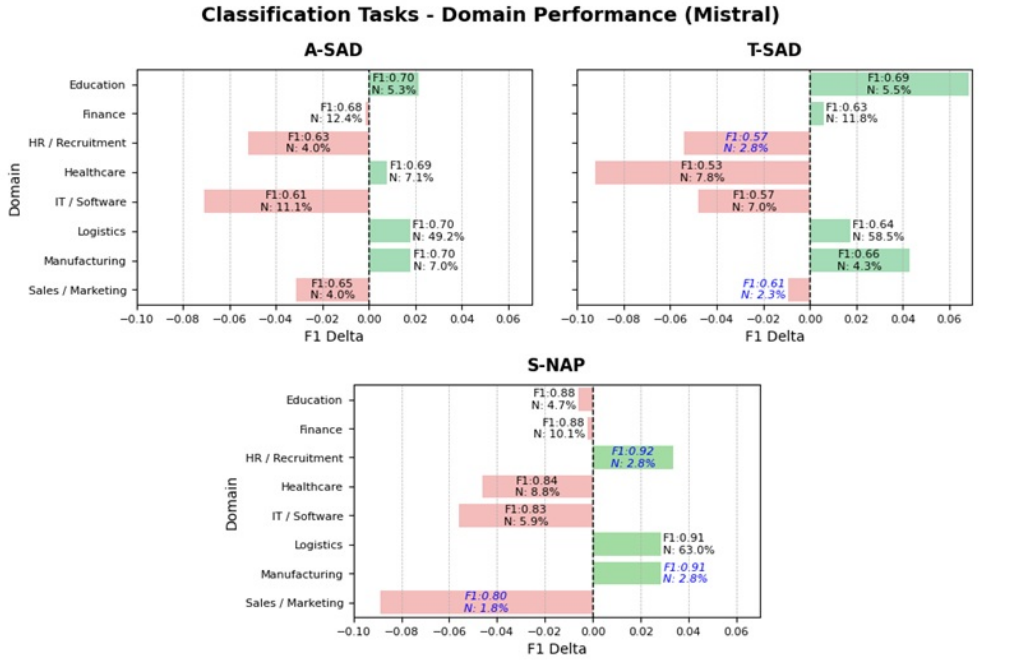

Fig. 3. Domain performance of Mistral IT on the A-SAD, T-SAD and S-NAP tasks, relative to average performance. Domains with representation of less than 3% of total samples have their "N" value highlighted in blue.

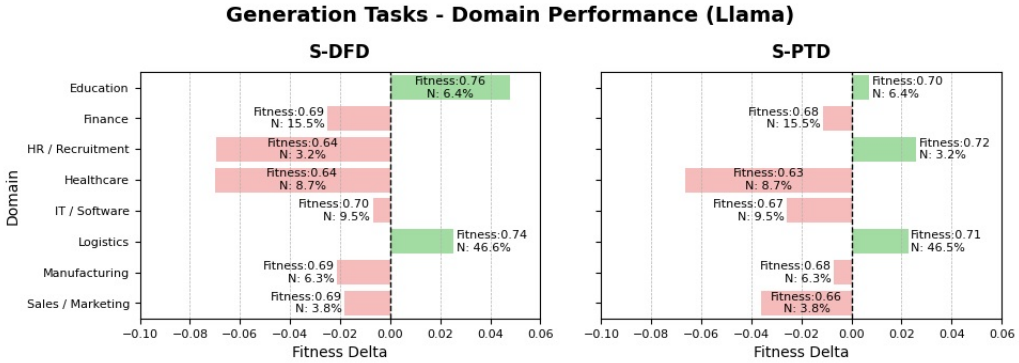

Fig. 4. Domain performance of Llama IT on the S-DFD and S-PTD tasks, relative to average performance.

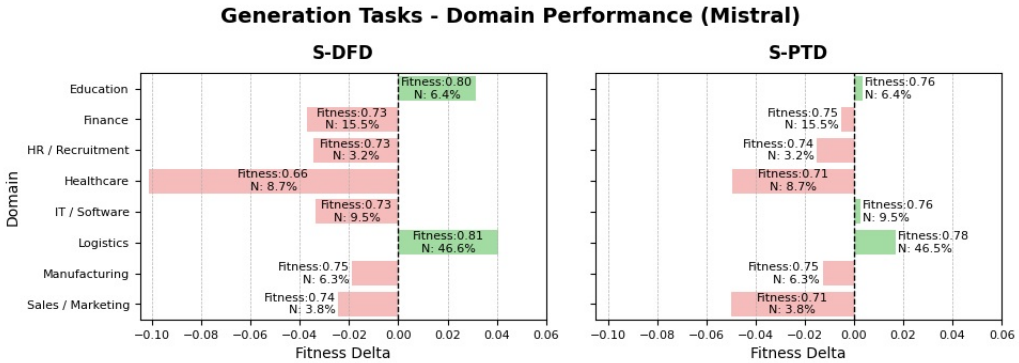

Fig. 5. Domain performance of Mistral IT on the S-DFD and S-PTD tasks, relative to average performance.