Comparing LLMs in business rule-following

Nikson Bernardes Fernandes Ferreira[1], **William Freitas**[1], **Hallyson Melo**[1], **Andre Carvalho**[2], **Thiago Borges**[1], **Rodrigo Marques**[1]

[1] Institute de Desenvolvimento Tecnológico (INDT)
[2] Universidade Federal do Amazonas (UFAM)

{nikson.ferreira,william.freitas,hallyson.melo}@indt.org.br
{thiago.borges, rodrigo.marques}@indt.org.br
andre@icomp.ufam.edu.br

Abstract. *Large Language Models (LLMs) have shown great capabilities in language understanding and instruction following. However, to the author's knowledge, no prior work has evaluated their performance in real industry internal rule-following scenarios compared to humans. The present R&D project aims to analyze the applicability of LLMs in improving efficiency in task analysis and scheduling through automatic team assignment, following a set of internal business rules. The study was funded by SUFRAMA and is a collaboration between INDT and Motorola Mobility. The experiment results show that lightweight open LLMs, on average, have worse accuracy than mean worker (57.5% x 86.25%) with a higher divergence rate (90% x 45%).*

1. Introduction

Large Language Models (LLMs) can accomplish numerous Natural Language Processing (NLP) tasks thanks to their instruction-following ability. However, in real-world applications, people often expect LLMs to generate outputs that conform to user-provided rules [Sun et al. 2024]. In this work we are interested in Inferential Rule-following that differs from following instruction in the objectivity of the task. Instructions describe the objective or the steps to perform it, inferential rule following gives abstract predicate entailments, namely rules, and the model is responsible for applying the rules in a given context finding a correspondence (if it exists) with the input [Sun et al. 2024]. Companies usually expect their employees to strive for rules to perform internal processes. In this sense, evaluthe evaluation of the LLMs' capabilities to help workers to perform their usual job is attreled to its inferential rule-following abilities.

To try to decide if this technology can help in a work environment, previous works have been made to try to create benchmarks or ways for checking and evaluating the rule-following capabilities of a large language model, such as RuleBench [Sun et al. 2024] and the LLM Comparator [Kahng et al. 2024]. However, LLMs often come with a warning to use them carefully, as they tend to generate false data. So evaluating their contribution to a real work environment is quite challenging. On the other hand, this could help to decide if LLMs can help or get in the way of users when they need to follow a set of rules and do their activities on a larger scale.

This research projetct aims to compare lightweight local running LLMs to find out how well they can help the users perform their tasks. Local models are used instead of external services, e.g. ChatGPT or Gemini, due to the common confidentiality of internal organization rules. In this way, four human evaluators were selected to evaluate a set of instructions and check if they could categorize tasks to specific teams. Meanwhile, the

rules they used were stored so they could also be followed by the LLMs later on and metrics could then be extracted.

2. Project Objectives

The Instituto de Desenvolvimento Tecnológico (INDT) is a technology-focused organization commited to advancing progress through innovation. Its activities broad several research and development areas including Advanced Manufacturing, Industry 4.0, Hardware and Firmware, Communication and Networks, Autonomous Vehicles, Robotic, Materials and Chemical, and Biotech serving and colaborating with various large companies of these sectors.

One of the colaborations are in the field of software testing. It was identified that the partner Motorola had thousands of tasks written in free natural language to analyze and execute. In 2023, It was proposed a research and project aiming to improve efficience in the process automaticaly standadizing input descriptions avoiding missunderdanding and enhance readability. In 2024, the project was extended to also reduce task redundance and automatic execution team assignment. The present work cover automatic team assignment which should consider a set of internal rules that limit team roles.

The objective of the present case study is to evaluate the applicability and the impact of open Large Language Models (LLMs) on performing inferential rule-following on team assignment company's processes, i.e., to discern business rules to define which one should me responsable for execution of the input task.

3. Proposed Method and Technologies

As described in [Sun et al. 2024], a rule can be understood in logical terms as entailment $\sigma \models \phi$, which means *"if σ holds, then ϕ also hold"*, where σ and ϕ are first order logic formulas. For instance, consider the predicates $isFalling(X)$ and $sell(X)$ which reads *"X is falling"* and *"sell X"*, respectively for a free variable X. One possible rule can be $isFalling(X) \models sell(X)$. In inferential rule-following, the objective is to find a substitution for one or more rules and apply implication given an input context, e.g., in the sentence *"NVIDIA stocks is falling"*, the model should find the substitution *[X/"NVIDIA stocks"]* and, then, the consequence *"Sell NVIDIA stocks"*.

In our case study, the rules defines which team should perform a determined task given teams limitations and explore the tradeoff between them. An Exemple of potential rules is $requires(X, y) \vee isExecuted(X, N) \models executes(a, X)$ where upper case terms are free variables which should be instantiated and lower ones, constants; the rule is read as *"if X requires y or it is executed N times, the team 'a' executes X"*.

A set of state-of-art lightweight open source LLM models were evaluated for confidentiality reasons. All models are inputed with the same prompt following the template shown in the Figure 1 instructing the model about the company context, the objective and the business-rules in natural language as if-then-else expressions. Exploring the causality of the model and following works such as [Wei et al. 2023], the model first provides a reasoning about the input, then provides a response. All time experiments were performed using llama.cpp framework with greedy sampling in a user grade computer aiming realistic time that even small companies could obtain: 16GB 12th Gen Intel® Core™ i712700H x 20 and a RTX 30606 GB.

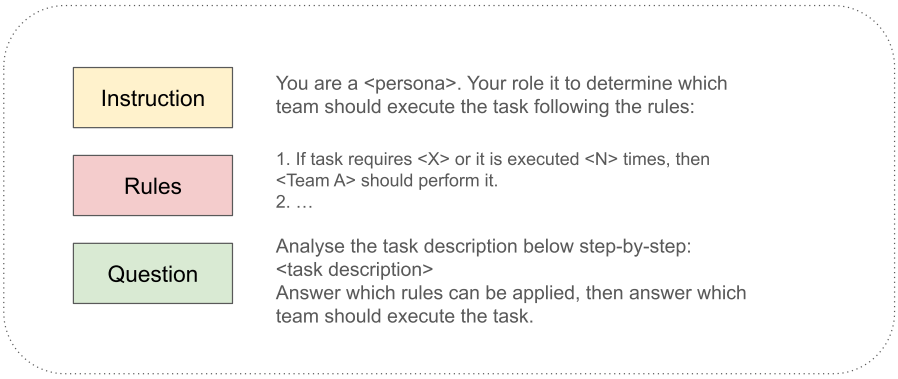

Figure 1. Inferential rule following prompt template used.

Model	MMLU	Accuracy (%)	avg time (s)	std time (s)
Llama3.2-3B	58	50	4.92	2.09
Phi-4-mini-4B	67.3	45	7.40	5.22
Gemma34B	59.6	75	5.11	1.42
DeepSeekR1-distil-Llama-8B	-	60	14.15	1.99
Humans	-	86	70.1	60.54

Table 1. Comparing open LLM MMLU benchmark, performance in business-rule following task and time perform a task in a user grade computer. All models are quantized in 4 bits.

A group of four human evaluators (selected by their experience level) performed this categorization of twenty tasks manually without communicating with each other. We recorded the time required for each assimilation, including time to consult resources. The majority vote was considered as the correct output, then the accuracy for each user was calculated and set as baseline for the approach. The discordance rate between workers also was evaluated. If there is one worker that disagrees with others, it is counted as discordance. The discordance rate is calculated by the number of cases where a discordance occurs over the total of cases.

4. Results

In 45% of cases where workers are assigned to independently perform the evaluation in the study, at least one of them gave a different response from others (divergence rate). It means that all workers agreed on only 55% of 20 tasks. Considering the majority vote as the correct categorization, the users hit on average 86,25% with a standard deviation of 8,54%. The results indicate that even humans can struggle to perform a relatively simple task making some mistakes, such as directing the task to a team. The low concordance and high accuracy indicate that workers tended to make different errors. These divergences can be related to previous experience or some arbitrary subjectivity of the user. On average, humans take 70.1 seconds to analyse the problem and provide a response in the case study. It varies considerably depending on problem complexity, it is possible to see on standard deviation that is 60,56 seconds, minimum of 10 seconds and maximum of 390 s.

For comparison, Table 1 also presents a common benchmark with results publicly available: the Massive Multitask Language Understanding (MMLU) benchmark evalu-

ates LLMs pre-training in language understanding, knowledge and problem solving using questions from several fields with diverse range of difficulties. It is presented here to illustrate that there is no apparent relation between the global performance indicated by the classical benchmark and the result obtained in a specific business-rule following the experiment. The models' performance varies significantly from model to model. The Gemma-3-4B model performed better than the Phi-4-Mini that presented 112.9% gemma performance on the MMLU benchmark while it presented the worst performance on study.

In addition, Gemma-3 and DeepSeek-R1 distiled performed nearest to human level performance, however, it is still a long difference to it. The first performs the objective 36.1% of time in a sequential usage in a user grade computer. This time can be reduced even more with batched generation. LLMs disagreed in 90% of cases.

The present case-study evaluated the capacity and performance of low-scale LLM models to follow and apply internal business rules. Based on the average performance of humans, it indicates that it is possible to use low scale open LLM models to help company following internal business-rules at a slight performance drop compared with human. However, there are still opportunities for improvements. The project next steps are to analyse rules to turn them more concise and suitable for models' understanding and to try to apply the project idea for other activities besides team assignment, such as automatic labeling task's requisites. Additionally, the authors aim to implement a agentic approach that proposes and checks rules instantiation.

Acknowledgements

The present research is resulting of a Instituto de Desenvolvimento Tecnológico (INDT)'s R&D project benefiting from SUFRAMA tax incentives from Informatics Law No. 8387/1991. It is result from a cooperation between INDT - Instituto de Desenvolvimento Tecnologico and Motorola Mobility Comercio de Produtos Eletrônicos Ltda.

References

Kahng, M., Tenney, I., Pushkarna, M., Liu, M. X., Wexler, J., Reif, E., Kallarackal, K., Chang, M., Terry, M., and Dixon, L. (2024). Llm comparator: Visual analytics for side-by-side evaluation of large language models.

Sun, W., Zhang, C., Zhang, X., Yu, X., Huang, Z., Chen, P., Xu, H., He, S., Zhao, J., and Liu, K. (2024). Beyond instruction following: Evaluating inferential rule following of large language models.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.