Analysis of the role of attention and prompt engineering in logical reasoning with decoder-only transformers

Final project for the course *Natural Language Processing*

Project Prompt: P7 (AY 2024-25)

Ivan Mazzon

1 Introduction

This project examines the ability of a large language model (LLM) to perform logical reasoning by analyzing the attention mechanisms within a decoder-only transformer architecture. The model is provided with instances composed by a theory and a logical proposition, and tasked with determining whether the proposition is true or false. The central hypothesis is that, when the model produces the correct answer, it allocates greater attention to the actual most relevant portions of the input, particularly the logical statements that support the reasoning process. The second aspect is to evaluate the impact of different prompting techniques on the ability of the model to classify logical propositions, as well as on the attention distribution.

2 Related work

Recent studies debate whether LLMs truly perform symbolic reasoning. [1] claims that LLMs act as semantic reasoners, showing better performance in contexts in which semantics are consistent with commonsense, compared to when a symbolic or counter-commonsense scenario is presented. [2] shows that LLMs, when combined with symbolic modules can act as neurosymbolic agents, remarking their potential for integrating language understanding with structured reasoning. [3] argues that LLMs lack true symbolic reasoning, but can approximate it when combined with external modules or advanced prompting strategies.

This project extends these findings by analyzing to what extent attention mechanisms in a decoder-only model reflect reasoning behavior, particularly in relation

to correct predictions and theory-focused attention. It also investigates how different prompting strategies influence both classification performance and attention allocation.

3 Methodology

3.1 The model

To conduct this study it was chosen the model *Mistral 7B Instruct 0.3* [4, 5]. The model is based on the decoder-only transformer architecture [6]. A decoder-only transformer is an artificial neural network architecture that is composed exclusively of the decoder component of the original transformer design. Instead of having an encoder to process the input and then a decoder to generate the output, the model treats the input sequence itself as the beginning of the text to be continued. It uses causal self-attention, which means that at each step the model can only take in account the previous tokens in the sequence, and not to future ones, ensuring that generation is performed left to right.

In order to reduce the usage of GPU memory, the model is loaded in a 8 bit quantized format. Quantization inevitably introduces a trade-off in terms of output quality, but this compromise was considered reasonable given the significant reduction in memory footprint and inference time.

3.2 The dataset

In this project we used ProofWriter (CWA depth-3) [7], a synthetic dataset designed to evaluate logical reasoning capabilities in language models. Each instance in the dataset consists of

- a theory, a set of natural language facts and rules;
- a question, a statement whose truth value is to be determined;
- an answer, labeled as *True* or *False*, that will be used as a ground truth;
- a proof, the symbolic step-by-step derivation of the answer.

Due to technical limitations in device memory and time execution, a subset of 3000 instances from the original dataset is considered. The analysis conducted with this relatively small number of samples should be interpreted as a preliminary indication of the model's general behavior rather than a statistically exhaustive analysis.

3.3 Analysis approach

In order to verify the research hypothesis, we propose the following approach. Each prompt is constructed in a modular way, with every group of words assigned a specific role. This subdivision makes it possible to analyze the mean attention allocated to the tokens within each group, thus identifying how much attention each set of tokens receives. The idea is to compute the average attention of groups of tokens on different heads on the last layer to understand which part of the input has major influence on the output. Calculating the average attention at the last layer is especially interesting because it integrates information from all the preceding layers, reflecting the model's

final representation before output. In the prompt, the model is instructed to provide the simplest possible proof for the question, in order to leverage the advantages of chain-of-thought reasoning.

The token group representing the logical theory is further partitioned into smaller parts associated with single tuples and rules to allow a more fine-grained analysis. For each statement of the theory, the average attention weight assigned to its corresponding tokens is computed, producing a measure of how strongly the model focuses on that element. These values are then compared against the set of statements actually used in the proof, which serves as the ground truth for relevance.

By treating attention scores as predictors of statement importance, the evaluation metrics AUC and AUPRC are computed to quantify how well attention distinguishes between relevant and irrelevant components. There are few cases in which all the statement of the theory are relevant for the proof; in those cases, AUC cannot be computed because we do not have examples of both classes, positives and negatives. For those instances, the AUC metric is neglected.

The described analysis is applied to different types of prompts in order to evaluate what is the impact of prompt engineering in the reasoning. We considered three kinds of prompts to submit to the model: zero-shot, in which only the theory and the instructions are given, one-shot and few-shots, in which one or two examples are given to guide the text generation. Appendix A illustrates an example of role segmentation within a one-shot prompt.

4 Results

The evaluation of logical reasoning performance across different prompting strategies — that are shown in Table 1 — reveals only modest improvements over the random baseline, with the zero-shot setting that achieves an accuracy few greater than 50%. This suggests that, without examples, the predictions of the model are largely heuristic and not strongly guided by logical inference. The one-shot and few-shots settings reach slightly greater performances. The overall improvement remains limited, suggesting that providing examples offer a guidance but does not essentially alter the reasoning capabilities of the model. Precision and recall follow a similar trend, with recall generally higher than precision, showing that the model tends to overpredict the *True* label.

Figures 1-3 illustrate the distribution of average attention across tokens grouped by role and the obtained significance of proof-relevant terms, using AUC and AUPRC. In the zero-shot setting, attention is mainly oriented to preamble, instructions and

Table 1: classification performance using different prompting techniques.

	accuracy	f_1 score	precision	recall
zero-shot	0.520	0.549	0.517	0.585
one-shot	0.555	0.598	0.545	**0.662**
few-shots	**0.571**	**0.599**	**0.563**	0.639

labels tokens, with a weak alignment to reasoning-relevant statements of the presented theory. In contrast, the one-shot prompt is the method that focuses more on the given theory. When there are more examples, in few-shots prompts, the attention is more distributed across the examples and instructions, but the theory is not particularly focused. This behavior suggests that the higher performance using the few-shots technique does not derive from a real understanding of the logical theory, but from an imitation of reasoning patterns proposed in the examples.

Moreover, comparing correct and incorrect predictions, we have that correct predictions tend to exhibit a slightly higher range of mean attention values, especially on reasoning-relevant and question tokens.

Finally, the evaluation of proof-relevant terms metrics shows that attention weights can barely distinguish between relevant and irrelevant statements within the theory section; correct predictions achieve higher values, but the difference is not sufficient to guarantee a distinct separation between statements that are involved in the proof and those that are not.

5 Discussion

The results lead to the following last considerations. The initial hypothesis that the model would allocate more attention to the logical theory when producing correct answers is not fully supported. Although, in general, there is a slight increase in attention toward the theory in correct predictions, the difference compared to incorrect ones is not marked enough to confirm a stable, reliable pattern. The metrics that determine how close the distribution of attention among the terms is to the actual relevant terms for the proofs also do not show a strong increasing adherence in case of correct prediction. The various prompting strategies influence the distribution of attention and the overall classification performances, but their impact remains narrow.

Thus, the analysis shows that, with the configurations considered in this project, attention weights are not a reliable proxy for logical relevance. They provide only a weak signal of the reasoning ability, but they do not offer a robust explanation of the model behavior.

Some directions for new analysis can be the extension to other LLM models and different larger corpora of logical theories and questions; a deeper exploration observing the evolution of attention distribution in intermediate layers; investigate alternative interpretability methods to assess which approaches most effectively explain the reasoning capabilities of LLMs.

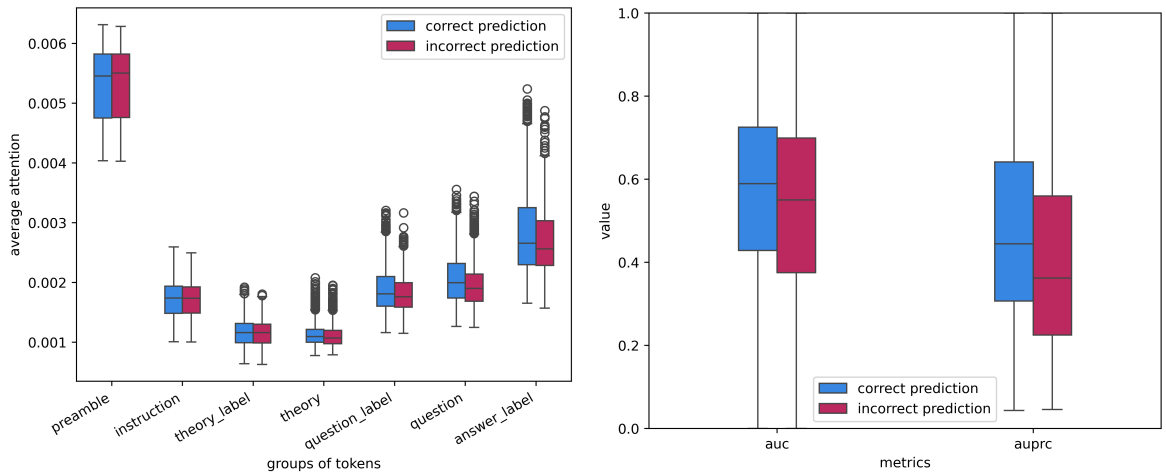

Fig. 1: average attention per group and proof-relevant terms metrics with zero-shot prompting technique.

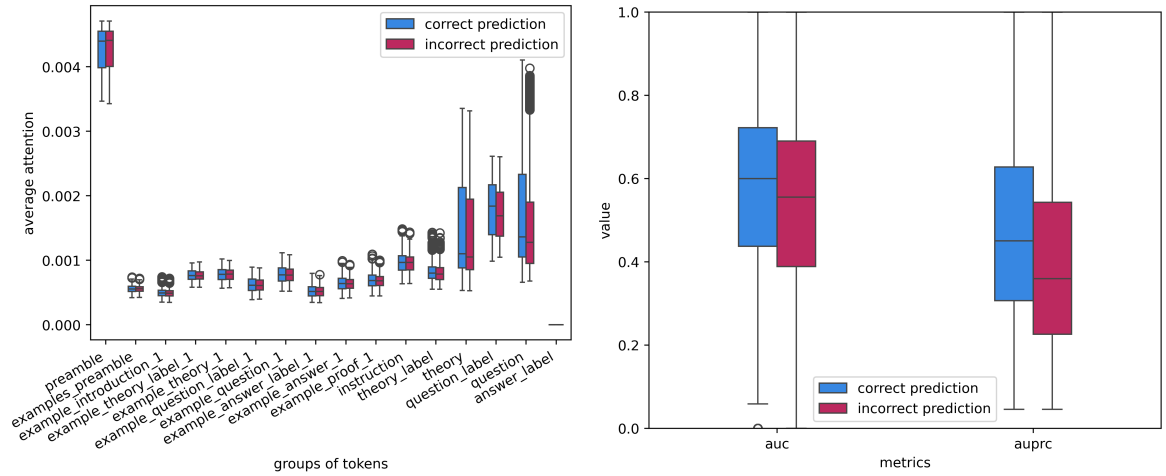

Fig. 2: average attention per group and proof-relevant terms metrics with one-shot prompting technique.

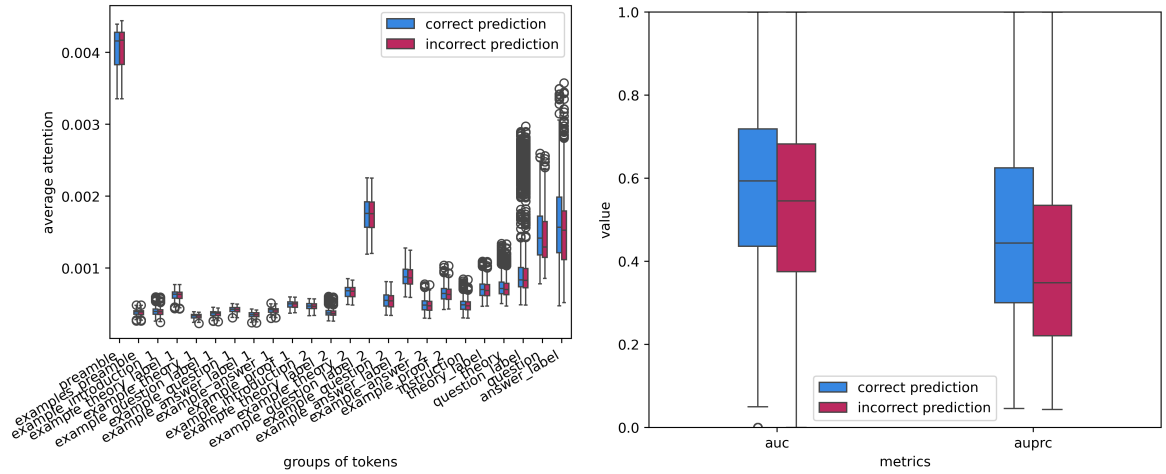

Fig. 3: average attention per group and proof-relevant terms metrics with few-shots prompting technique.

References

[1] Xiaojuan Tang et al. "Large Language Models are In-Context Semantic Reasoners rather than Symbolic Reasoners". eng. In: *arXiv.org* (2023). ISSN: 2331-8422.

[2] Meng Fang et al. "Large language models are neurosymbolic reasoners". In: *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2024. ISBN: 978-1-57735-887-9. DOI: 10.1609/aaai.v38i16.29754. URL: https://doi.org/10.1609/aaai.v38i16.29754.

[3] Rob Sullivan and Nelly Elsayed. *Can Large Language Models Act as Symbolic Reasoners?* 2024. arXiv: 2410.21490 [cs.CL]. URL: https://arxiv.org/abs/2410.21490.

[4] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

[5] *Mistral-7B-Instruct-v0.3*. URL: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. (accessed: 6.10.2025).

[6] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[7] Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. *ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language*. 2021. arXiv: 2012.13048 [cs.CL]. URL: https://arxiv.org/abs/2012.13048.

AI Usage Disclaimer
Parts of this projects have been developed with the assistance of Microsoft Copilot, based on GPT-5. The AI was used to support the development of project ideas, structuring of methodological workflows, draft writing and code prototyping. All content produced with AI assistance has been carefully reviewed, edited, and validated by me. I take full responsibility for the final content and its accuracy, relevance, and academic integrity.

A Role segmentation of a one-shot prompt example

Role	Content
preamble	You will be shown a logic theory and a logic question. Produce two levels of output: 1. Write the final answer (only one word: True or False) inside <final>...</final>. 2. Write the proof, formatted as in the example, inside <proof>...</proof>. When writing the reasoning, always choose the simplest valid proof. Avoid unnecessary steps or complex derivations. If a single-step proof exists, use only that step.
examples_preamble	Here is an example.
example_introduction_1	Example:
example_theory_label	Theory:
example_theory	The bear is cold. t2: The bear is kind. t3: The bear is young. t4: The bear visits the dog. t5: The dog is blue. t6: The dog is young. t7: The dog needs the rabbit. t8: The dog visits the bear. t9: The dog visits the rabbit. t10: The lion is kind. t11: The lion needs the bear. t12: The rabbit sees the bear. r1: If something needs the dog then it sees the lion. r2: If something sees the bear and it needs the rabbit then the rabbit is young. r3: If the dog needs the bear and the bear needs the lion then the bear visits the dog. r4: If something visits the lion then it sees the bear. r5: If something is blue then it visits the lion. r6: If something needs the rabbit and the rabbit visits the bear then it visits the lion.
example_question_label	Question:
example_question_1	The dog is young.
example_answer_label_1	Answer:
example_answer	<final>True</final>
example_proof_1	<reasoning>t6</reasoning>
instruction	Now, evaluate the following.
theory_label	Theory:
theory	t1: Anne is big. t2: Anne is furry. t3: Anne is white. t4: Fiona is big. t5: Fiona is furry. t6: Fiona is kind. t7: Fiona is quiet. t8: Fiona is white. t9: Fiona is young. t10: Harry is big. t11: Harry is furry. t12: Harry is kind. t13: Harry is quiet. t14: Harry is rough. t15: Harry is white. t16: Harry is young. r1: Big things are rough. r2: All white things are rough. r3: If something is white then it is young. r4: If something is quiet and young then it is kind. r5: If something is white and rough then it is quiet. r6: If something is white then it is furry.
question_label	Question:
question	Anne is kind.
answer_label	Answer: