UNIVERSITÀ
DEGLI STUDI
DI MILANO

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

Ivan Mazzon

Final project for *Natural Language Processing* course

December 2025

Research objective and hypotheses

Objective: analyzing to what extent the attention mechanisms in a decoder-only model reflect logical reasoning behavior in determining the truth value of a logical proposition given a theory.

Hypotheses:

→ An LLM model allocates a greater attention to the logical theory, and to its proof-relevant statements, when the correct answer is produced

→ Different prompting techniques have an impact on the ability of the model to classify logical propositions, and on the attention distribution

The model

Mistral 7B Instruct

- Instruct fine-tuned
- Decoder-only transformer architecture
- 8-bit quantized version

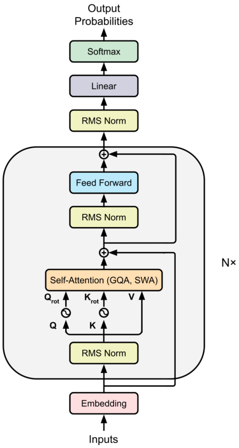


Figure: decoder-only architecture

The dataset

ProofWriter

Instances structure in the dataset

1. a theory
2. a question
3. an answer (*True* or *False*)
4. a proof

Methodology

- Structured prompts: each prompt is subdivided in groups by role

- Average attention over heads in the last layer for each role

- Evaluation of the attention distribution against the actual proof-relevant statements

Prompt segmentation 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. [...] 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. [...]
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. [...] r1: Big things are rough. r2: All white things are rough. r3: If something is white then it is young. [...]
question_label	Question:
question	Anne is kind.
answer_label	Answer:

Results

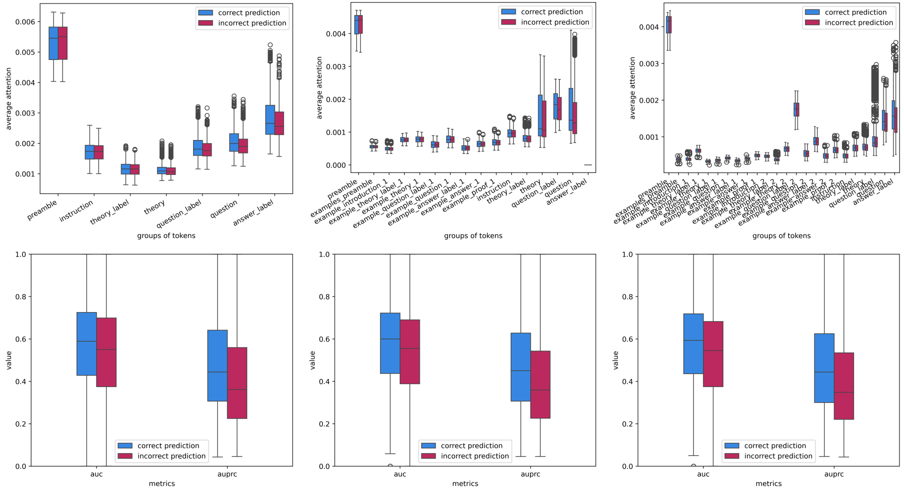
Classification performances

Table: 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

Results

Attention distribution and adherence to proof-relevant terms



Conclusions

The initial hypotheses are not fully supported.

- slight increasing in attention to the theory in correct predictions, but the difference compared to incorrect ones is not marked enough to confirm a reliable pattern

- The metrics AUC and AUPRC do not show a strong increasing adherence in case of correct predictions

- The different prompting strategies influence the distribution of attention and the overall classification performances, but their impact is narrow