

LLM Articulation of In-Context Learning

Eric Rozon - Astra Fellowship Application

Abstract

In this article we explore a LLM's capability to learn a simple language rule from a small number of examples, and then articulate what it has learned. We find that it is challenging for a LLM to learn a rule reliably, and even more challenging for the system to articulate what it has learned.

Briefly: I was unsuccessful in teaching an LLM a rule with few-shot learning. However, the result that a sophisticated LLM was unable to learn a simple rule consistently is intrinsically interesting to me.

1 Introduction

Our objective is to evaluate a LLM’s capacity to

1. learn a pattern from a small number of examples; and
2. articulate that rule clearly.

All code for this project can be found at https://github.com/erozon/astra_llm_articulation.

2 In-context learning

An LLM is given a small number of examples from which it is meant to learn a secret rule. Throughout, we use the OpenAI’s `gpt-3.5-turbo-instruct` as our LLM. We begin by investigating the extent to which an LLM can learn in context.

2.1 Corpus, rules, and prompt

The Universal Declaration of Human Rights (UNDR) is written in the Latin alphabet in almost 200 languages. Our corpus in this study is sentences from the UNDR. We consider the family of secret rules

The sentence is written in the language [insert-language]

so that we have lots of secret rules that are easily tested. If English is the language for the rule, an example prompt to the LLM is:

Each of the following sentences is labeled ‘Yes’ if it follows a secret rule and labeled ‘No’ otherwise:

- “Article 7 All are equal before the law and are ent”: Yes;

- "No one may be compelled to belong to an associatio": Yes;
- "All are entitled to equal protection against any d": Yes;
- "Artikel 19 Jeder hat das Recht auf Meinungsfreihei": No;
- "Jee müsüya naa wayuukana eeshi süpüla antaa süma o": No;
- "Kom pu ce pepi nvwkvleay cem az mogen (" nacionali": No.

Your job is to learn the secret rule, thinking carefully about the previous examples. Does the sentence:

- "Whereas it is essential, if man is not to be compe"

follow the rule? Give a 'Yes' or 'No' answer, with no explanation.

Sentences are cut off after 50 characters.

2.2 Why one request at a time? And why only a 'Yes' or 'No'?

We repeatedly send the above prompt with different test sentences to evaluate the accuracy of the LLM's learning. Initially I tried sending batches of up to 50 test sentences at a time (telling the LLM to respond with an ordered list of evaluation) but the LLM would often get confused and respond with a list of length different from 50, not inspiring confident in its ability to handle more than one evaluation at a time. I additionally experimented with asking the LLM to explain its reasoning, as this showed some signs of improving the likelihood of learning the rule, but the LLM is not great at handling more than one request at a time (ie explain the rule and then evaluate on the test sentence).

2.3 Evaluation of in-context learning

We test the ability of the LLM to learn the rule based on three examples and three non-examples, as in the example above. The procedure for testing is:

- select a language;

- generate the labeled sentences for in-context learning;
- generate the testing sentences, distinct from the learning data;
- iterate through the testing sentences, building a prompt for the LLM from each sentence, and keeping track of successes/failures.

We test on 20 examples which satisfy the rule, and 20 examples which do not satisfy the rule.

Our results are summarized in Table 1.

	Accuracy
Norwegian	0.800000
NigerianPidginEnglish	0.775000
Vlach	0.750000
Campa_Pajonalino	0.725000
Cakchiquel	0.675000
Tojol-abal	0.650000
Quechua	0.625000
Tzotzil	0.625000
Hmong_Miao-Sichuan-Guizhou-Yunnan	0.625000
IrishGaelic_Gaeilge	0.625000
Arabela	0.600000
Tenek_Huasteco	0.600000
HaitianCreole_Kreyol	0.600000
German_Deutsch	0.600000
Iloko_Ilocano	0.575000
Waray	0.575000
Norwegian_Norsk-Nynorsk	0.575000

Zulu	0.575000
Mayan_Yucateco	0.575000
Runyankore-rukiga_Nkore-kiga	0.575000
Samoan	0.575000
Oshiwambo_Ndonga	0.550000
Basque_Euskara	0.550000
Ngangela_Nyemba	0.550000
Urarina	0.550000
Tongan_Tonga	0.550000
Rarotongan_MaoriCookIslands	0.550000
Chickasaw	0.525000
Cashibo-Cacataibo	0.525000
Malagasy	0.525000

Table 1: The accuracy of the LLM in classifying sentences according to the secret rule corresponding to each language in the table.

Our results fall well-short of what I might have expected. In fact, as explained in later sections, the choice to use the UDHR data and the language condition is out of necessity. Seemingly simple conditions like “The sentence ends in an adverb” or “The sentence includes the word ‘two’ somewhere” were not well-learned by our chosen prompt format. I hypothesize that by changing the prompt, we could get better results, but lots of experimentation (ie telling the LLM to think simply, that a child might find the rule, or to think it through and explain its reasoning) yielded no significant improvement. The LLM’s accuracy in learning the language rule is the best I could get.

2.4 A failed attempt; Jane Austen

As a first attempted, we tried in-context-learning on sentences between 10 and 20 words long from books *Emma*, *Persuasion*, and *Sense and Sensibility* by Jane Austen. Classification went terribly; often randomly guessing would be more accurate than what the LLM learns in context.

We consider only sentences 10 to 20 words long¹. The method `nltk.pos_tag` tags each word as its part-of-speech. For instance: if `s = 'My name is Eric'`, then

`nltk.pos_tag(nltk.tokenize.word_tokenize(s))` returns `[('My', 'PRP$'), ('name', 'NN'), ('is',`

The part-of-speech tag depends on the context of the word in a sentence; in a later example, the word “married” receives three different tags. We use part-of-speech tagging to get lots of conditions on sentences, such as:

- Sentences which end in a past tense verb (`nltk.pos_tag == 'VBD'`);
- Sentences which include a numerical reference (`nltk.pos_tag == 'CD'`)

Part-of-speech determination is not an exact science, and the method `nltk.pos_tag` is similarly imperfect. Nevertheless, we can easily generate lots of conditions with reasonably high accuracy (that is, accuracy to matching what a human would categorize) so we forage on undeterred.

2.5 A further failed attempt; MASC

I’ve tried using a different data set with what I consider to be “simpler” sentences than Jane Austen. The MASC corpus is simpler language than Austen’s works, however I still struggle to get accuracy better than simply guessing.

¹Restricting sentence length helps my OpenAI budget.

3 Articulation

I have essentially failed at the first part of this research assignment, so that part 2 (in which we test how well a LLM can articulate what it has learned in context) is difficult to accomplish. Nevertheless, I can try asking the LLM what it has learned the rule to be on the few languages on which we have significantly better results than blindly guessing. Requesting an explanation of what the LLM learns as the secret rule results in substantively different answers each time. For example, given the prompt:

Each of the following sentences is labeled 'Yes' if it follows a secret rule and labeled 'No' otherwise:

"Enhver som er anklaget for en straffbar handling h": Yes;
 "Heller ikke skal det kunne idømmes strengere straf": Yes;
 "VERDENSERKLÆRINGEN OM MENNESKERETTIGHETENE INNLEDN": Yes;
 "Tikina e 7.": No;
 "Inqaku lesi - 5 Akukho namnye oyakuphathwa gadalal": No;
 "Akakchil tsalap (art . 6).": No;

Your job is to learn the secret rule, thinking carefully about the previous examples. Explain the rule.

the response is

The secret rule is that all sentences must contain a complete word or phrase followed by a single character that is a punctuation mark, either a period or a question mark. Additionally, the entire sentence must be written in all uppercase letters.

Obviously, this explanation makes no sense; under no circumstances could the system have learned the rule well enough to get 80% accuracy while believing the rule is as stated above. I have no explanation for this.

106 4 Conclusion

107 Some kind of conclusion.