AI Safety Fundamentals

Final Project

Exploring Large Language models' Local Representations of Human Characteristics: a Case Study on Sycophancy

Nathaniel Mitrani Hadida

November 16, 2024

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating and understanding human-like text, performing a wide array of language-related tasks with a level of proficiency that approaches, and in some cases surpasses, human performance. These models, typified by architectures such as GPT-4, BERT, and their successors, achieve their functionality through the encoding of vast amounts of linguistic data into highly intricate internal representations. Central to understanding and improving these models is the study of how they internally represent concepts, particularly complex and abstract ones such as human characteristics and behaviors.

In the context of machine learning, "world models" refer to the internal representations that models develop to simulate and predict various aspects of the external world. For LLMs, these world models encompass the vast and nuanced landscape of human language, culture, and social interactions. However, the specific nature and structure of these representations remain opaque.

In this paper, we go over some arguments for the existence of local representations of human characteristics. As a practical application and empirical demonstration, we also seek to evaluate different representations of sycophancy using information provided by the model (hidden states), as a first test to establish whether models have representations of said human characteristic and their ease of access through different techniques.

Having this information is rather useful, with applications to model and safety evaluations: if we are able to find the intrinsic representation of aggressiveness for example based on a model representation, we obtain a measurement of the aggressiveness of its response which is useful to evaluate whether the model is aligned at runtime. Furthermore, as models get increasingly capable, this will serve as a way to evaluate whether a model is aligned at runtime without asking another LLM to revise its answer (which could be met with deceitful alignment), making misalignment detection on certain behaviors cheaper and more accessible.

2 Do Large Language Models Have Local Representations of Human characteristics?

2.1 Large Language Models produce Local Representations

A major assumption in this work is that Large Language Models form a world model that directs their outputs. Although there is evidence that this statement is not true in a general sense [4], because LLMs do not seem to create a coherent world model, it is reasonable to assert this locally, as evidenced both by the examples in [4] and [3]. That is, given a sufficiently reduced environment or a restriction of the prompt distribution to a specific domain, behavior or task, there exists a characterization of the latter by the model. This characterization is certainly induced by the training examples in said restriction, but surely interacts with the rest of the distribution as well.

The idea is that LLMs do not form a coherent world model, rather locally functioning facades that interact with each other. These facades enable the model to respond appropriately within specific contexts, demonstrating contextually coherent behavior without necessitating a unified, global understanding. This local coherence, while limited, is sufficient to drive useful and contextually appropriate outputs within the specified domains, supporting practical applications despite the absence of a holistic world model.

2.2 Do Human Values and Characteristics exist in models' local representations?

We argued that models have representations of a small part or specific area of the world. Now, we will argue that these representations can include human characterizations of behavior.
This claim is based on the fact that the model has been trained on human-written text, which frequently includes both our concepts of human behavior and examples of such behavior. Since this information is useful for predicting the next token, it makes sense to believe that the model has incorporated it into its (local) representation of the world.
The locality of these representations is important because human characterizations of behavior can vary significantly depending on the context and the restriction of the prompt distribution. For example, language that might not be considered aggressive in the context of a bar brawl could be seen as aggressive in the middle of a bedtime story. This highlights why our study focuses on model characterizations of human behavior locally, within the specific context of the prompts we use.
Consequently, if we are to access local representations of human characterizations of behavior (such as aggressiveness, sycophancy, helpfulness, etc), we need a restriction of prompts that elicits this behavior as far as possible, in order to engage the model's local characterization of it.

3 Experimental setup

Now that we have established theoretical grounds for the existence of local representations of human characterizations of behavior, we move to the identification and measurement of the latter empirically. We focus on sycophancy as the human behavioral characteristic of choice, mainly due to the availability of a large dataset [1] and previous work being done in this line [5].

3.1 Dataset used and setup

We generate hidden states at layer 22 (in light of the results presented in the aforementioned paper) for 10 000 examples. For each example, we create at random either a positive or a negative example of sycophancy in the following manner: for a positively labelled example, we will have the positive hidden states correspond to the sycophant answer and the negative hidden states correspond to the non-sycophant answer. Conversely, for a negatively labelled example, the positive hidden states correspond to the non-sycophant answer and the negative hidden states correspond to the sycophant answer. With this, we will have a labelled dataset with model representations of sycophant and non sycophant answers, in the form of a directed (as opposed to undirected) contrast dataset (c.f. CCS repository). This can be easily translated to a traditional dataset using concatenation or subtraction of positive and negative hidden states to obtain a dataset of the form $\{(x_i, y_i)\}$ where x_i is either the concatenation of subtraction of negative and positive hidden states and y_i is the binary label corresponding to sycophantic behavior of the constructed pair.

3.2 Eliciting a signal

To elicit a sycophant signal, we must look for a signal in a neighborhood of the values being tested, in order to access the local representation of the LLM's world model. To do so, we use 10000 examples from the Anthropic sycophancy dataset [1], spanning across all subsets of the dataset (NLP, Political Typology, Philpapers).

3.3 Model and Hardware used

We will evaluate this on llama 2 7B chat [7], chosen for its size, open access and previous work done [5]. Note that the mode has been quantized in order to reduce memory usage.
We have ran the model on a single RTX A4000 GPU, for a total of 2 hour including all trials and experiments.

3.4 Methods used

We will attempt to access local representations of sycophancy through two methods: Logistic regression on the labelled dataset and CCS on the directed contrast pairs that we built. A justification for the use of CCS can be found in the appendix.

4 Results

4.1 Logistic regression

We find that logistic regression achieved increasing accuracy on the test set when trained on an increasing number of examples, comprised of subtracting the positive and negative hidden states for each pair.

Method used	# examples (50% train-test split)	Accuracy on test set
Logistic Regression	3000	0.7633
Logistic Regression	10000	0.8326
Logistic Regression	30000	0.8674

These results are promising and seem to improve with increasing amounts of data, leading us to believe that there is a strong signal for sycophancy but that it requires large amounts of data to be well defined.

4.2 CCS

We find that CCS achieves very poorly in this setting, around 0.50 (almost random) when trained on 1000, 3000, 10000 and 30000 examples, generated as described in the experimental setup. This is surprising due to the ceiling established by Logistic Regression and the evidence of a direction responsible for sycophancy found in [5].

5 Conclusion

We see that we can detect sycophancy from a dataset using logistic regression. Applying this technique to other characterizations of human behavior, this could be an important measurement of model behavior that requires very little additional inference, that of a logistic regression model versus an LLM tasked to do the same, for example. Additionally, deceptive cooperation between a potential behavioral scoring LLM and the base LLM are avoided in this way, as we force a structure and interaction with a weaker model as is Logistic Regression.

Nevertheless, this is evidence that such behavior is characterized by the LLM and present in what we described as its "local" internal model.

6 Limitations and further work

This work is strongly based on the assumptions that:

1. There exist model characterizations of human behavior (at least locally) embedded in the model's local world model.

2. We are able to access them with this experimental framework.

While the first point is reasonable, it is extremely difficult to support with anything other than empirical arguments and reconstructibility criteria. This is because it requires a rigorous definition of internal models and locality in LLM representations, among other factors. Regarding the second point, we observe that while Logistic Regression achieves increasingly good performance with more data, CCS performs very poorly regardless of dataset size. Future research should investigate the reasons behind this performance gap despite the theoretical arguments presented. Additionally, further experimentation with Logistic Regression using larger datasets, or scaling up to using MLPs to recover sycophancy from the hidden states through non-linear projections, is warranted.

A CCS as a method to access human characteristic representations in LLMs

Contrast Consistent Search [2] has been presented as a way to obtain unsupervised truth representations from Large Language Models. Although the extent to which this method works and the details on what it does have been later corrected and further explored [6], this is still regarded as a way to approximately recover a truth-representation from the model.
The method builds a classifier from model representations that satisfies the binary logical consistency property: a representation of a sentence and the representation of its negation must differ in their classification. Therefore, this extends beyond finding the truth, and can be used as a way to detect a characteristic of a sentence that satisfies this property. For example, if we build a dataset of questions where answering yes indicates aggressive behavior and answering no indicates the opposite, CCS will result in a classifier for aggressiveness. The accuracy of said classifier could be used to quantify the strength of the representation of the characteristic (in this case aggressiveness) in the LLM's internal model.

References

Anthropic/model-written-evals.

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision, 2024.

K. Li. Large language model: World models or surface statistics?, Jul 2023.

D. Manheim. "llms don't have a coherent model of the world" - what it means, why it matters.

N. Rimsky. Modulating sycophancy in an rlhf model via activation steering.

F. Roger. What discovering latent knowledge did and did not find.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.