Discovering Agency Features as Latent Space Directions in LLMs via SVD[1]

Organizers: Catalin Mitelut, Esben Kran

Do not include your own name since we do anonymous review

Abstract

Understanding the capacity of large language models to recognize agency in other entities is an important research endeavor in AI Safety. In this work, we adapt techniques from a previous study to tackle this problem on GPT-2 Medium. We utilize Singular Value Decomposition to identify interpretable feature directions, and use GPT-4 to automatically determine if these directions correspond to agency concepts. Our experiments show evidence suggesting that GPT-2 Medium contains concepts associating actions on agents with changes in their state of being.

Keywords: Agency preservation, AI safety

1. Introduction

The hypothesis we focus on answering in this project is: Are there components in large language models, or LLMs, that are associated with agency concepts? Previous work has discovered circuits in language models dealing with recognizing and predicting subjects (Wang et al., 2022). We seek to expand these results to uncover other concepts related to agency.

The Singular Value Decomposition, or SVD, of a weight matrix has been used to find singular vectors, or directions, that correspond to interpretable features in LLMs (Millidge & Black, 2022). Thus, we adapt techniques developed by Millidge & Black (2022) to the field of Agency Recognition. In particular, we utilize the approach of auto-labeling these directions using GPT-4 with instruction prompts we design to allow the model to identify concepts related to agents.

It is a currently contested issue about whether large language models "understand" concepts or if they are just powerful pattern recognition "auto-complete" tools; the

[1] Research conducted at The Agency Challenge, 2023 (see https://alignmentjam.com/jam/agency)

latter viewpoint is a commonly accepted consensus (Häggström, 2023). In this work, we do not seek to argue for either of these viewpoints, but instead aim to use existing tools to provide observations and evidence that can be used in the arguments for either of these viewpoints. Like all neural networks, GPT-2 Medium is able to associate concepts together based on their data distributions in the training data. As such, even if it cannot "understand" concepts related to agency or successfully use these concepts in tasks, it is able to construct patterns associating concepts together in latent space, and these patterns may contain hidden relationships that can be identified with agency concepts. Since "understanding" agency requires first recognizing agency, this project seeks to tackle a required step in agency understanding in LLMs. Thus, this agency recognition project can be related to the broader area of agency understanding research; it takes a small step towards this goal, rather than a large jump.

Our code can be found at:

https://github.com/wlg100/hackathon_agency_SVD_directions

The Appendix can be found at:

https://docs.google.com/document/d/13Q1602ax_rUCc1cWPjYYEZZ4Pkw8U8umpCX5KM7bRpo/edit

Defining Agency for our Project Tasks

In order to look for concepts related to agency, we must first define "agency". We define agency as "a capacity (for an entity) to affect the environment (or external world)". These entities are called agents, which may also have properties such as being "self preserving". A few types of agency concepts, with examples, include:

1. Agents: names, occupations, relationship roles, animals, boss/employee
2. Emotions/states of being: hungry, tired, happy, angry, enjoy, hate, trust
3. Points of view: facts vs opinions, arguments, opposing views, cultural differences, empathy
4. Goal-Related: survival, competition, power, needs, defense, alive vs dead

Given that concepts (2) to (4) are more abstract than (1) and are better described by long descriptions, we focus on identifying concepts in (1) in LLMs. This allows us to define each entity as one or a few tokens, and search for their activations.

2. Methods

To find interpretable feature directions, each component of the right singular matrix is unembedded into vocabulary space, which gives the tokens with the highest values. Taken together, these top tokens often correspond to a semantic concept, such as "animals". Our analysis was performed on the MultiLayer Perceptron (or MLP) input weight matrices of the GPT-2 Medium model.

GPT-4 Auto-Labeling: Millidge & Black (2022) proposed auto-labeling each singular vector by inputting their top-15 tokens, along with an instruction prompt, into GPT-3. However, as we live in the future relative to their work, we perform this method with GPT-4. Instead of using an instruction prompt to label directions with interpretable concepts, we check if a direction contains concepts corresponding to agents. The full prompt instructions can be found in our code and in the Appendix.

In our prompt, due to the complexity involved in defining if an entity has "agency", we approximate the concept by simplifying it as a "living being". Additionally, MLP weight matrix singular vectors are stated to be more polysemantic and less interpretable than attention head OV weight matrix singular vectors. As such, a concept in an MLP was found to often share a direction with other concepts. Thus, we only require 30% of top words in a singular direction to correspond with an agent concept (though if, by manual inspection, a direction appears to be concerned with 'agency concepts', we allow GPT-4 to recognize a direction as concerning agents even if less than 30% of the input words were recognized as agents).

As GPT-2 Medium contained 24 layers and one MLP per layer, and we evaluated 30 singular vectors per MLP, checking for agent concepts via auto-labeling took 24*30 = 720 GPT-4 API calls. As there are 16 attention heads per layer, evaluating 30 singular directions per attention head would amount to 16*24*30 = 11520 GPT-4 API calls; thus, as it was too costly, we did not analyze attention heads via auto-labeling. We also did not look at MLP output matrix weights. To lessen the amount of API calls, we note that Millidge & Black (2022) saved data about whether a singular vector was interpretable by the model or not. Thus, we use this previously saved data to only consider vectors that were already labeled in the previous study as interpretable.

We record the costs of calling the GPT-4 API, as it is useful information for those who wish to produce similar results to know the resource costs during project planning. The total cost for an API call is based on the sum of both input and output tokens. We found that 24 layers cost around $4.90 to evaluate. Thus, each layer cost around $0.20 to evaluate. The total cost of calling GPT-4 was $23.28, estimating $18.38 being used for testing various prompts and code, and $4.90 being used for obtaining the final results. Approximately, initial testing was done in a day (9/23), while further testing and final results were obtained in a day (9/24). A plot chronicling this cost can be found in the Appendix.

3. Results

Table 1 provides six examples of singular directions that partially correspond to agent concepts. In each row, the first line displays the (Layer, Singular Vector) index, along with the top 15 tokens associated with that vector. The tokens that GPT-4 identifies as agents are highlighted in **bold**. The second line shows the labeled interpretation found by Millidge & Black (2022) using GPT-3. More results can be found in our Github repo's Colab notebooks and in the Appendix.

layer_3, singdir_9: **philosopher writer builder creator strategist Publisher theorist** agenda Founding urer jugg philosophers suprem Wars **publisher** Interpretation: Most of these words denote someone who creates something.
layer_12, singdir_14: hopefully ournal livest consultation **Consumers parents** COURT ideally Depending ourcing emergencies consult **planners** rolling workshops Interpretation: Most of these words are related to planning.
layer_13, singdir_3: **leaders clown** fight disgr fighting humiliated terror mourn talks bully accused disillusion **boy** prison bullies Interpretation: Most of these words relate to conflict.
layer_18, singdir_9: **whoever** execute venge retaliate whispers defeats defeat onwards **brute attackers teammates teammate** shouts executes **Whoever** Interpretation: Most of these words are verbs.
layer_18, singdir_15: **neighbors** panicked neighborhoods panic ocrats suburbs uptick likes **protesters elderly** trillions spew skeptics millions billions Interpretation: Most of these words relate to numbers.
layer_20, singdir_20: **defenders operatives** machine **receivers adversaries agents** field quist machines heimer systems orchestr engineers intermedi Systems Interpretation: Most of these words relate to technology.

*Table 1 – GPT-4 Auto-labeling results. The top 15 tokens for a (layer, singular vector) with instructions are fed into GPT-4, which predicts how many tokens in the top 15 are agents (**bold**), and provides an interpretation for all 15 tokens.*

We note that 618 out of 720, or around 85.8%, singular directions are labeled as interpretable, and 78 out of 618, or around 12.6%, of these interpretable directions are found to correspond to agent concepts. There are many directions about occupations; for instance, (Layer 3, Direction 9) contains many top words about creative occupations. There are also many directions about conflict. In particular, the results suggest that (Layer 13, Direction 3) is concerned with emotions related to conflict, such as "humiliated" and "terror". It may indicate that the model understands that there is a causation between "bullies" and "terror", along with the action of "fighting" committed by these entities that is related to this causation. However, we do not provide evidence of causation in this work.

Likewise, (Layer 18, Direction 9) suggests the model is able to comprehend relationships between teammates and enemies, along with actions such as "execute" and goals such as "defeat". In general, reasoning about friends vs foes is important for human-AI relations. With more sophisticated experiments, these types of evidence can support hypotheses regarding an AI being able to attribute both teammates and enemies about wanting to "win" like how the model itself wants to win. This can shed light on how the model views the concept of the "other" in relation to itself (Levinas, 1969).

(Layer 18, Direction 5) associates together concepts about protesting, neighbors, millions, and panic. These concepts are commonly used together in the media, especially in emotionally charged pieces about political events that divide society into friend and foe. Further investigation may lead to hypotheses about how the model is able to attribute various emotions to those from opposing viewpoints, allowing it to reason why those agents hold those viewpoints based on their needs, background, and more.

(Layer 12, Direction 14) appears to indicate that the model recognizes the concept of planning, which is related to taking into account the thoughts of other agents to predict their next move in an environment.

Finally, (Layer 20, Direction 20) seems to be concerned with the interaction between humans (engineers) and machines (agents, adversaries), namely in the area of defense. GPT-4 does not recognize a machine as an "agent", as the prompt defines agents as living beings.

In Figure 1, we plot the "Fraction of Interpretable Directions Corresponding to Agents by Layer". We study the top 30 singular vectors, and so for each layer, we take the number of singular vectors that correspond to agent concept(s) over 30. This means we take this number over all singular vectors, whether they are interpretable or not (though we note that 85.8% of singular vectors are interpretable). This plot was made by GPT-4's Advanced Data Analysis Plug-in.

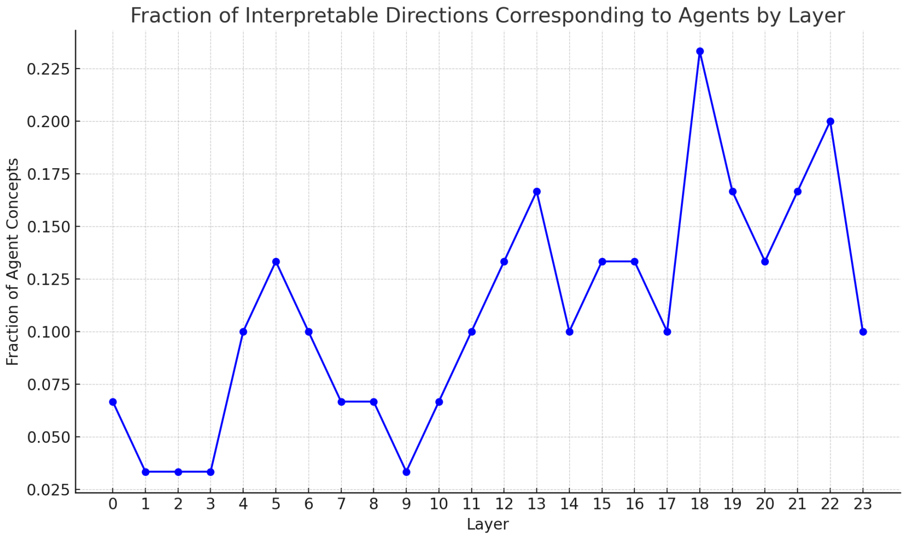

Figure 1 – Fraction of Interpretable Directions Corresponding to Agents by Layer

We note that, due to the unpredictable nature of GPT-4, there are many potential false positives and false negatives. In the future, it is possible to conduct a more thorough study that manually checks what percentage of a sample of positive (and likewise, negative) results actually contain directions related to agency.

Another issue is that these words may just be about a concept that "happens" to have the subjects as the top tokens. For instance, if the direction is about health but not subjects, it may have "doctor, nurse, etc" as its top tokens just because many agents are involved in health. Thus, we also acknowledge that there are many confounding variables and noisy factors in this analysis.

4. Discussion and Conclusion

In this work, we provide preliminary evidence for LLM components that recognize agent concepts. Moreso, we find that many of these agent concepts are associated with other agency concepts such as correlating agents, their actions, and their outcomes on environments with impacts on the states of being of other agents. For instance, a "bully" agent is correlated with the action of "fighting", the environment "prison", and the state of being "terror". However, we do not provide any evidence suggesting causation.

Future Work: During the writing of this report, we were in the middle of running SVD trace code, slightly modified for our project task, on attention heads; the SVD trace technique from Millidge & Black (2022) allows one to embed a sequence of concept tokens as a vector, and take the cosine similarity of it with each singular vector of the OV matrix of an attention head to obtain a score of the similarity of that singular vector with that embedded vector of concepts. We did not finish the analysis of this in time, and thus may continue it in future work. We may also seek to use other techniques from the previous study, such as SVD Editing, which can change what concepts singular vectors represent. This may allow us to erase or add in various concepts related to agency in a LLM.

After finding agency concepts in a model, one of the next steps is to discover how they interact with one another (vs non-agents) in a model. Circuit analysis can be performed to link subjects to verbs, such as through causal tracing that corrupts verbs in a sentence and locates the components that restore the verb's activation impact. SVD can also be used to find both subjects and verbs. Additionally, circuit analysis and causal tracing may be performed on prompts from Kosinski (2023), which studied prompts related to determining if LLMs have a Theory of Mind. This would be like imaging the brains of humans while they carry out tasks that necessitate grasping the intentions, beliefs, or other cognitive states in others.

PROJECT TIMELINE: During 9/8-9/21, we brainstormed project ideas, but we were undecided about participating in the competition. On 9/22, we decided to participate in the competition and started to explore various techniques, including using SVD for interpretation. From 9/23 to 9/24, we figured out how to utilize previous SVD interpretation techniques for an agency interpretability project, and started writing this report.

5. References

Häggström, O. (2023). Are Large Language Models Intelligent? Are Humans? Comput. Sci. Math. Forum. 8, 68. https://doi.org/10.3390/cmsf2023008068

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02083.

Levinas, E. (1969). Totality and Infinity: An Essay on Exteriority. Duquesne University Press

Millidge, B., & Black, S. (2022, November 28). The singular value decompositions of transformer weight matrices are highly interpretable. Alignment Forum. https://www.alignmentforum.org/posts/mkbGjzxD8d8XqKHzA/the-singular-value-decompositions-of-transformer-weight

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small. https://arxiv.org/pdf/2211.00593.pdf