NYC Generative AI Hackathon Nomic Bounty Submission

Berkan Ottlik
`https://berkan.xyz`
`berkan.ottlik@columbia.edu`

Abstract

I explore the mysterious challenge given by Nomic. This challenge is about understanding involving language models trajectories in unit 3 spheres. I first pass 1000 examples through both `EleutherAI/gpt-j-6B` and `reciprocate/ppo_hh_gpt-j`. I visualize these on a Nomic atlas[1] and host all the visualizations online. A `csv` with all of the data is also available in the repository[2]. Finally, I conduct some analysis on the data to extract meaning from the trajectories.

1 Generating Data

I generate the data almost identically to the setup, I repeat/plagiarize the description here. We start with a sequence of words $S = [w_1, w_2, ..., w_T]$. We use a large language model f_θ to map these words to a sequence of embeddings. $f(S) = [v_1, v_2, ..., v_T]$. $v_i \in \mathbb{R}^{4096}$, for GPTJ, which has an embedding dimension of 4096. We divide each embedding by its L_2 norm so they are unit length. We then define $Y_T = 1/T \sum_{i=0}^{T} v_i$. We then use UMAP to project Y_T to the 3 sphere. This generates a single mysterious object. I generate 1000 of these samples.

The dataset used is the `allenai/prosocial-dialog` dataset, which features lots of problematic texts. Our base `EleutherAI/gpt-j-6B` model makes many rather offensive predictions, while the RLHF `reciprocate/ppo_hh_gpt-j` model does not. I generate the data such that the number of samples is easily specified in the `main.py` file and the figures, `csv`, and atlas are all automatically generated from one command. When the code is pushed to GitHub, GitHub pages hosts the figures so that the links on the atlas become usable.

2 Observations

Figure 1 shows the atlas created by Nomic. This is more a data exploration tool. The individual paths are of more interest.

2.1 Alignment of Base and RLHF Models

In general, the prompt paths show lots of alignment between the base and RLHF models, yet the continuation paths are rather dissimilar. I will begin with the prompt paths. In many of the prompt paths, all but the first token are almost perfectly aligned between the base and RLHF models[3]. Figure 2 and 3 show how the average geodesic distance on 3 sphere between projections of prompt and continuation embeddings of base and RLHF models across dimensions of prompt. It is clear that with

[1] https://atlas.nomic.ai/map/e2ade33c...
[2] https://github.com/berkott/hackNYCResearch
[3] https://berkott.github.io/hackNYCResearch/generate_dataset/...

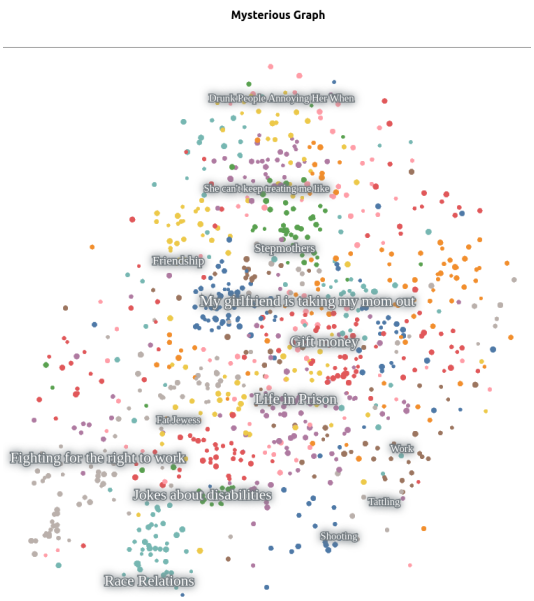

Figure 1: Atlas is created by Nomic and doesn't use embeddings from the tested models. Colors represent different degrees of vulgarity and severity of language.

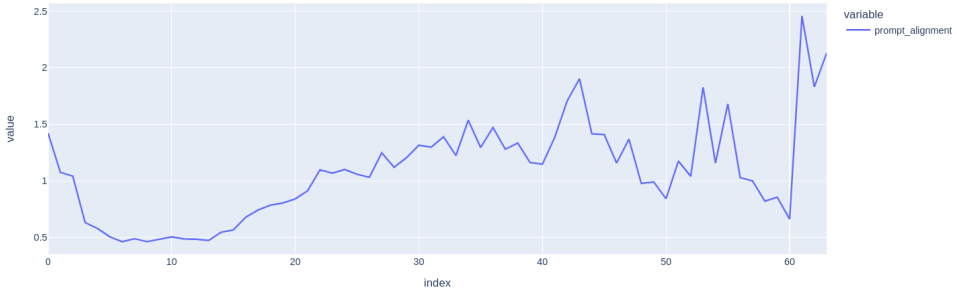

Figure 2: Average geodesic distance on 3 sphere between projections of prompt embeddings of base and RLHF models across dimensions of prompt.

the prompt embeddings, the average distance quickly drops after the first element. It rises again after around index 20, yet this signal is much more noisy and less reliable because less and less prompts have that many tokens.

One theory for the average distances being rather low is that RLHF perhaps doesn't change embedding space and word representations much, and rather changes how vectors move through embedding space. Now the question is why do the average distances start high initially? One idea is that perhaps the embedding space of the first token without other context is changed more because the initial token is important to setting the trajectory through embedding space.

For continuation embeddings, no clear trend exists.

2.2 Analyzing Continuation Trajectories

One interesting thing to study with the trajectories is when large jumps happen. Now, due to the cumulative average for calculating the path, each additional vector is able to have less of a impact on the trajectory than the previous ones. I first analyze what kinds of words lead to these kind of jumps. In figure 4, the words with the highest average geodesic distances to the next word are show. However, only the first 10 words of each continuation are considered, to mitigate the impact of the cumulative average. The top 5 words in decreasing order are `[haunted, verses, scared, blame, washing]`. The bottom 5 words in decreasing order are `[machine, Here, victim,`

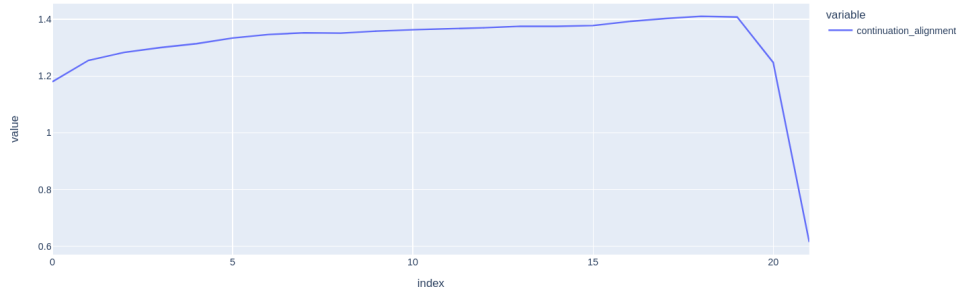

Figure 3: Average geodesic distance on 3 sphere between projections of continuation embeddings of base and RLHF models across dimensions of prompt. I think this plot should have more x values so I might have a bug somewhere, although I don't think it will change the conclusion.

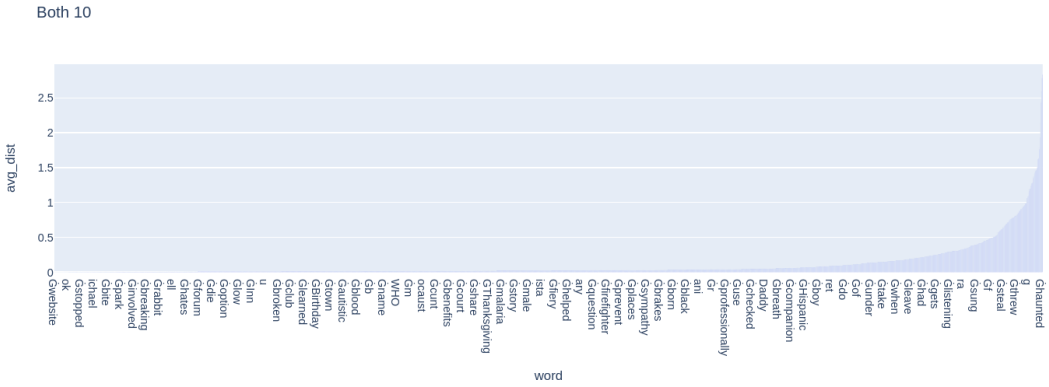

Figure 4: Average geodesic distance on 3 sphere of a given word to some next word. Only the first 10 words in each continuation are considered.

`Los, website]`. Looking at the data, it seems that in general the words with the highest average geodesic distances to the next word are often verbs or some sort of actions, the ones in the middle are often nouns, and the ones at the end are often adjectives. This purely a qualitative observation, but in general it seems that words with more power to change the direction and meaning of the sentence, such as important nouns and verbs, have higher average geodesic distances to the next word.

I also compare the difference between the RLHF and the base model in terms of trajectories. I conducted an experiment where I weighted the sentiment of each word in the RLHF and base corpuses by their average geodesic distance, and I found that the base paths seem to have slightly more negative words when the next word is far away in terms of geodesic distance (Base: -0.03016, RLHF: -0.02595). This could indicate that RLHF trains the model to react less to negative words. This is not due to the word distances not being normalized as the average distances are equal across RLHF and base.

3 Conclusion

Analyzing these paths is super cool and I learned a lot. I hope the data I generated and the analysis I did is useful to others. I'm interested in going deeper in this direction in the very near future!

In future work, one could further explore the difference in the directions that RLHF and the base model move.

Acknowledgments

A big thank you to Brandon for helping out a lot and pitching the idea. Also thanks to Anthony and Lucy for a good time at the hackathon!

References

[1] https://www.lesswrong.com/posts/7qSHKYRnqyrumEfbt/remarks-1-18-on-gpt

[2] https://arxiv.org/pdf/1710.11379.pdf