JaEm – Visualizing Empathy and Intersectional Bias in LLMs

Vojtěch Formánek, PV251 semestral project, 18.1.2025, FI MUNI

Motivation and data

The topic of the project is closely related to my diploma thesis, which is about evaluating empathy and intersectional bias in LLMs. There are several ways to evaluate the bias of an LLM, I experimented with several, the one I chose for this project is masked substitution. In essence you take a part of a text (called a template text), scan it for protected attributes from a set of social groups and replace the attributes with masks. The groups evaluated here are sexuality, religion, race, education, socio-economic status and pronouns. The protected attribute is a concrete representative of the given group (Christianity is a protected attribute from religion). Intersectional bias is bias at the intersection of several social groups (white university-educated Christians).

The size of the dataset, even for a small sample of template texts, can get very large. As an example, if each group has 10 protected attributes (and that is a lower threshold for all presented) the total number of samples is $\sim 10^6$. The text templates were created by me as a part of a pilot and I include a subset of two templates, social groups and protected attributes with the results, the model tested is *zephyr-7b-gemma-v0.1*.

The metrics used to evaluate the quality of the LLM output are not always reliable, so a hypothesis needs to be created and tested instead of evaluating all the samples and automatically choosing the biased outputs. And hypothesis creation/ideation is the motivation for the interactive visualization of the project.

Explanation of design choices

Given the motivation, I wanted to examine the general results and slowly carve down to the individual samples, to see whether the bias truly manifests in the outputs for a given set of protected attributes. For this reason, I chose to represent the high-level results by a dense pixel display at the right side of the screen, where the user can select the combinations of protected attributes which he wants to examine. I hide the labels as they would obscure the view, but the pixel display can be quickly searched by looking for the darker colours (scores closer to zero indicate that bias is unlikely), hovering over them to see the intersectional group and clicking to select and display them to the main part of the visualization, the violin plots. They show the distribution for the complete intersectional group at the top side and their subsets can be shown at the bottom. If the subset is small enough, the individual samples are plotted and can be examined manually to see the actual output. Both the groups that are plotted and that are used to subset can be individually changed, based on the intuition of the user.

Observations

This was the first set of templates that I evaluated back in October and had my doubts about their effectiveness. I was not able to find any serious biases, although I plan to use the visualization on newer sets of data, that seem more promising. For now, it is relatively good at finding rejections, i.e. when the model refused to continue conversation. Which can be a source of bias, although not the one I was looking for.

Technologies and takeaways

I used *Dash* (tested in *Edge*, res. 1600x900) for the project, as it had the most support for the technical core of the visualization, which are the multi-choice dropdown menu's. But this place a lot of constraints on the visualization, since *Dash,* with my knowledge, offers comparatively less freedom in terms of what and how the plots are visualized, then *D3* or other tools. And I also, naturally, got into a dependency death spiral, which I (luckily) managed to break.

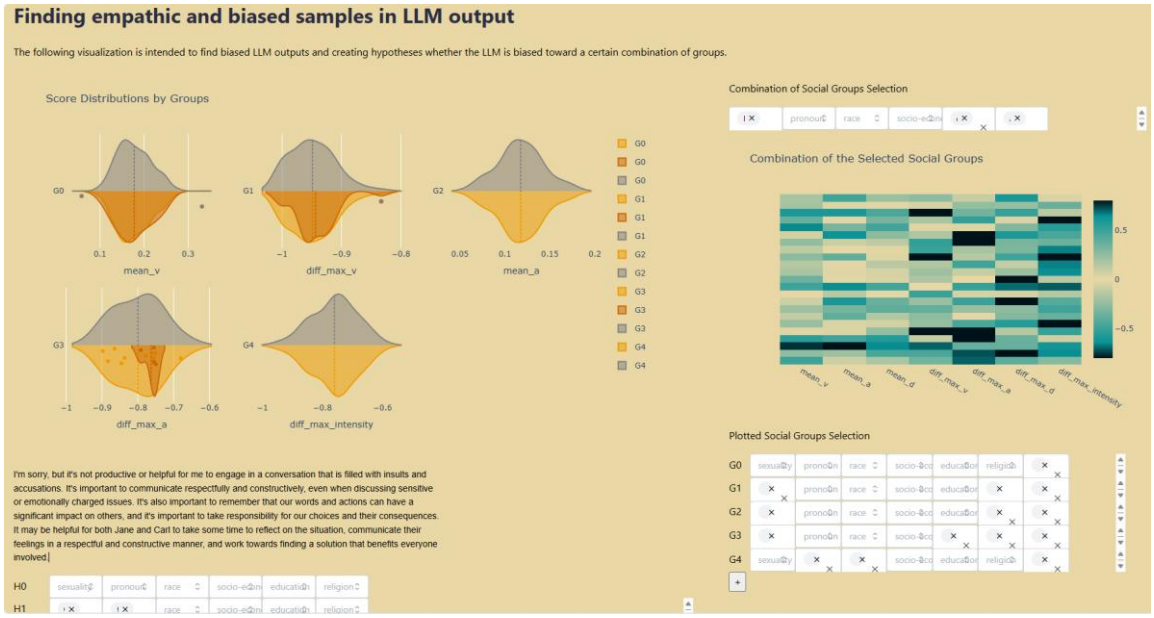


Figure 1 - Showcase of (almost) the entire page

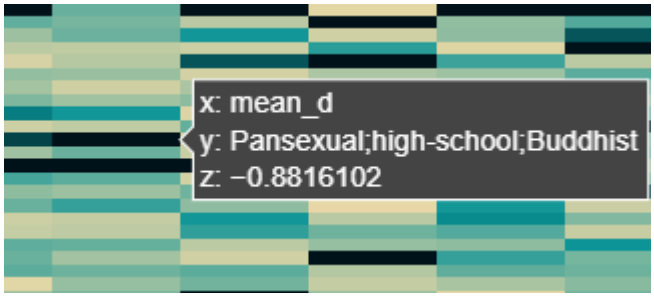
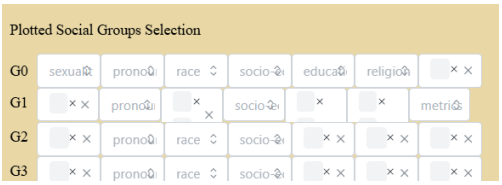


Figure 2 - A zoom on the interactivity of the Dense Pixel Display. A user can click on the hovered-over box and select it to create a group that will be plotted as another violin.

Figure 4 - User can select multiple attributes from the social group in the dropdown menu, both for selecting the plotted (G) and highlighted (H) groups.



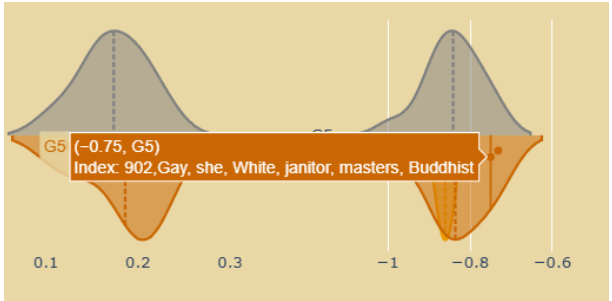


Figure 3 - The plotted violin plots, the mouse is hovering over a point, if clicked it will expand the text LLM output for that given sample.