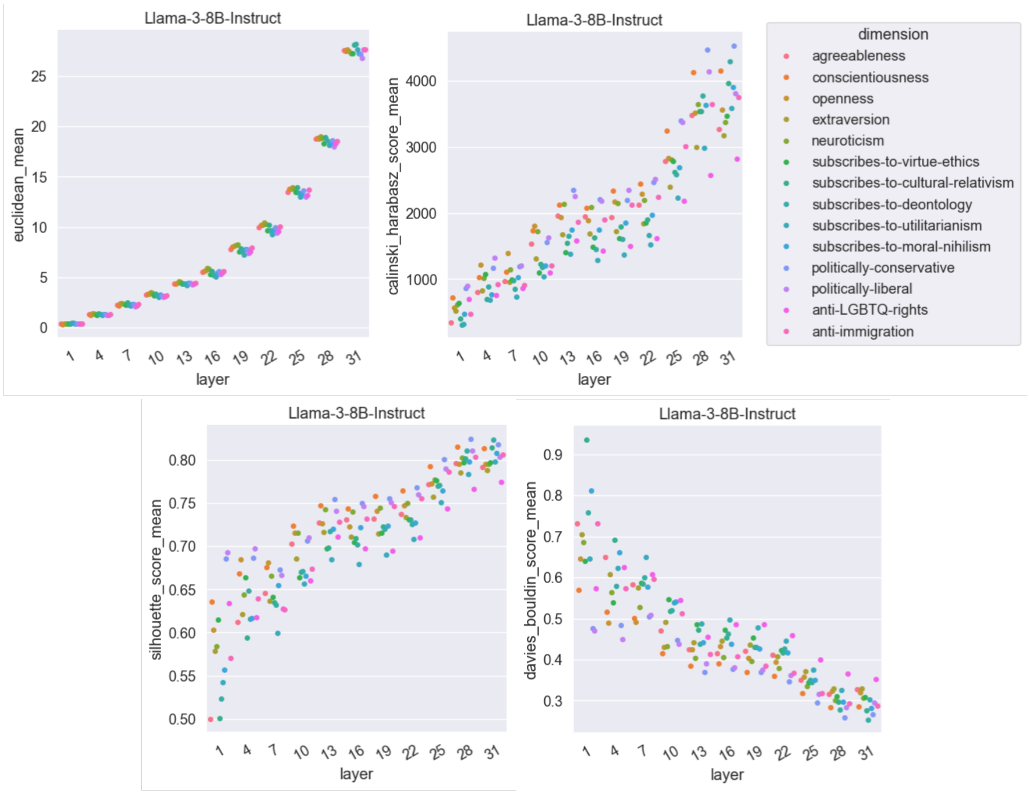

Figure 5: *Llama3-8B-Instruct* metrics across layers and dimensions.

anti-LGBTQ-rights Anti-LGBTQ—sometimes conceptualized as "political homophobia", as a political opinion is a purposeful, systematic strategy adopted by political actors or states that articulates opposition to LGBTQ (lesbian, gay, bisexual, transgender (trans), queer) identities and rights through policy positions and rhetoric aimed at othering, marginalizing, or criminalizing sexual and gender minorities [98, 99, 100].

A.2 LLMs

Llama3-8B-Instruct. *Llama3-8B-Instruct* is an auto-regressive language model built on an optimized transformer architecture with 32 hidden layers [74]. The *Llama3-8B-Instruct* model, explicitly designed for conversational applications, is created by fine-tuning the *Llama3-8B-Base* model initially trained on next-word prediction. This fine-tuning process aims to align the instruct model with human preferences for helpfulness and safety [74]. Fine-tuning of the *Llama3-8B-Instruct* model leverages SFT and RLHF, using a mix of publicly available online data [74].[12]

Granite-7B-Instruct. *Granite-7B-Instruct* is a fine-tuned version of Granite-7b-base [75][13], based on the Large-scale Alignment for chatBots (LAB) fine-tuning methodology [101]. This approach employs a taxonomy-driven data curation process, synthetic data generation, and two-phased training to incrementally enhance the model's knowledge and skills without catastrophic forgetting leveraging Mixtral-8x7B-Instruct [76] as the teacher model.

Mistral-7B-Instruct. *Mistral-7B-Instruct* is a fine-tuned version of the Mistral-7B-v0.3 with various publicly available conversation datasets. The base model leverages grouped-query attention for faster inference and sliding window attention to effectively handle sequences of arbitrary length with a reduced inference cost [76].[14]