## faithfulness in GPT2-xl $\mathsf{LRE}$ 0.9 8.0 Faithfulness 0.5 0.3 faithfulness in LLaMA-13B LRE 0.9 8.0 Faithfulness 0.5 0.3 0.1 0.0 person father ompany CEO gender gender gender religion by tool letter person lead singer of band word sentiment matter occupation age person native language fruit inside color adjective sûperlative country largest city country language country capital city person plays instrument landmark on continent superhero archnemesis adjective comparative verb past tense object superclass plays pro sport person university work location president birth year task person type star constellation name mother name birthplace fruit outside color adjective antonym word first letter president election year person occupation city in country person person sport position product by company pokemon evolution country currency word last letter food from country landmark in country company hq company name I name characteristic superhero task done person substance phāse ôf occupation