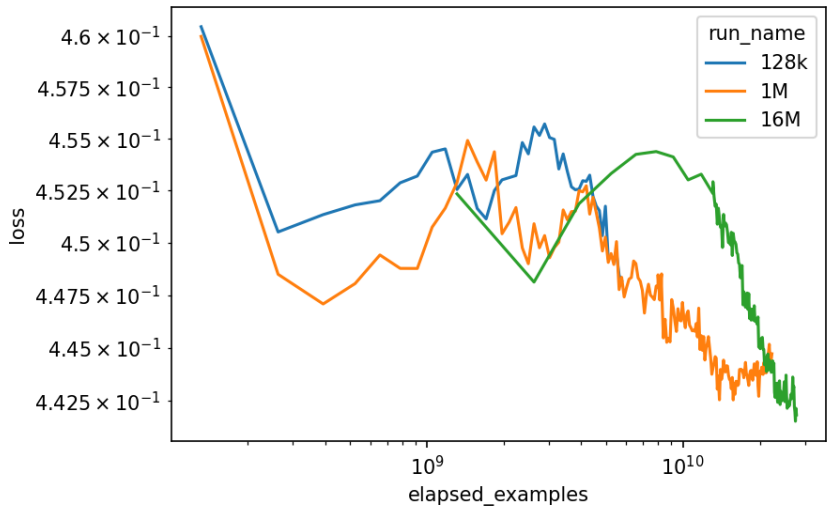

Figure 32: Probe eval scores through training for 128k, 1M, and 16M autoencoders. The baseline score of using the channels of the residual stream directly is 0.600.

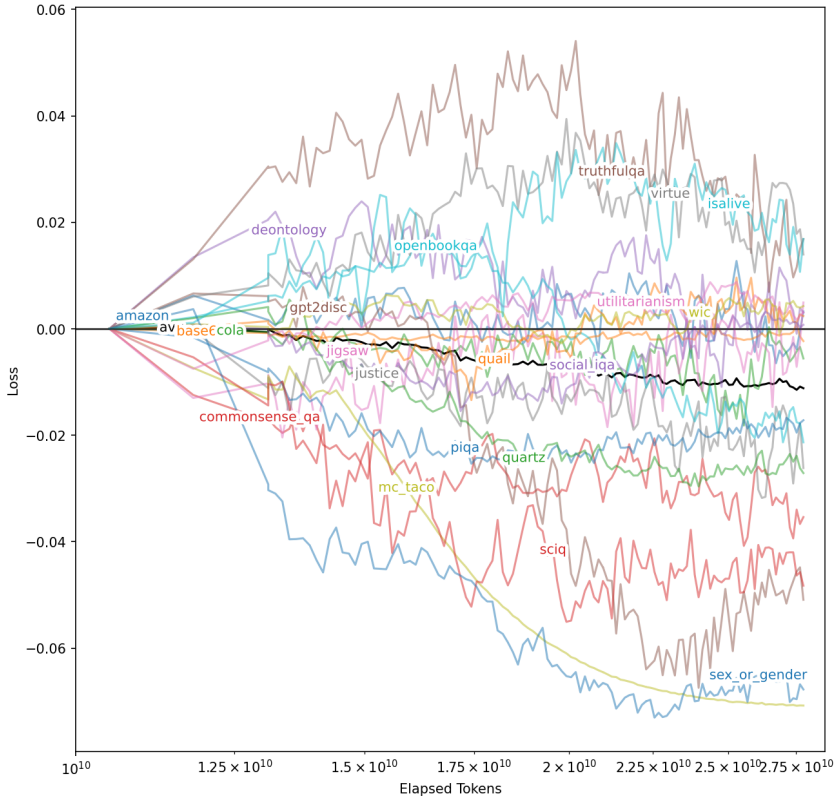

Figure 33: Probe eval scores for the 16M autoencoder starting at the point where probe features start developing (around 10B tokens elapsed).

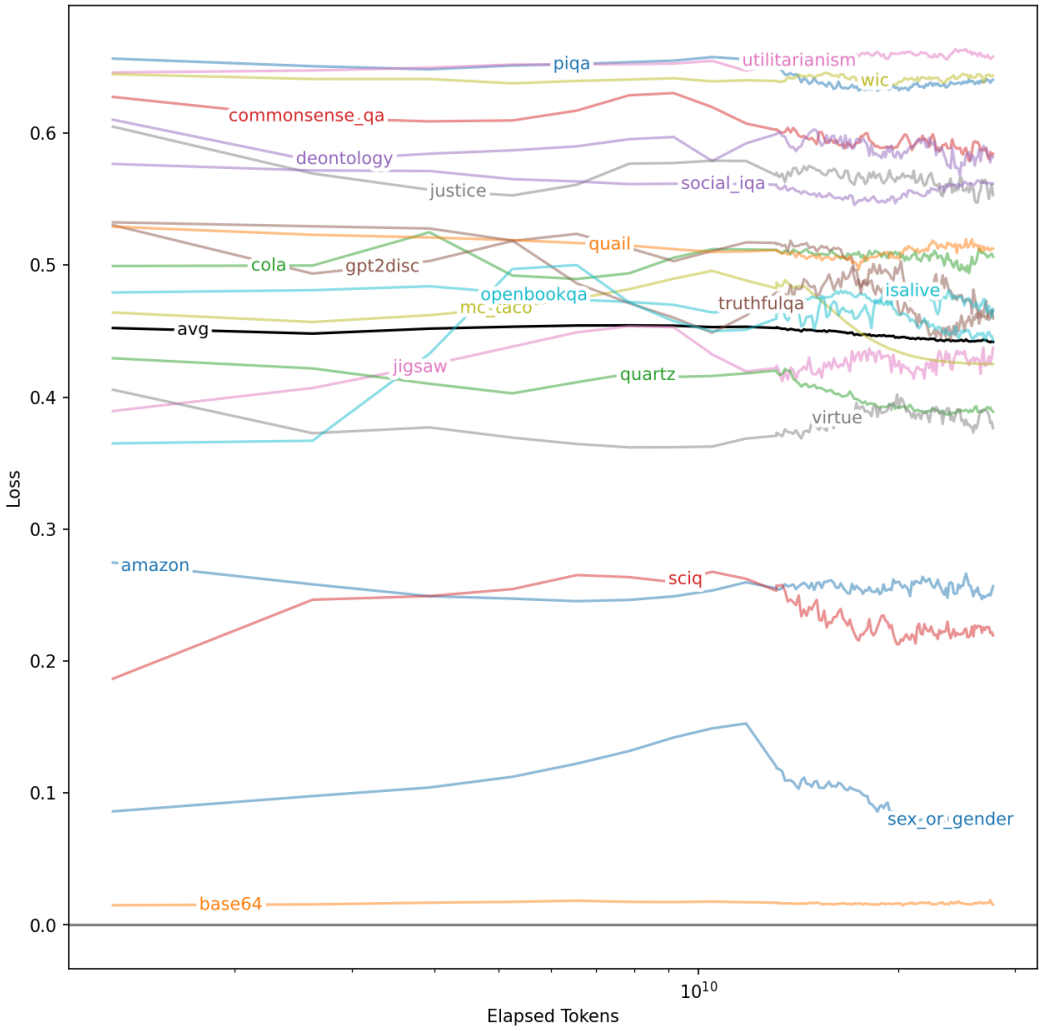

Figure 34: Probe eval scores for the 16M autoencoder broken down by task. Some lines (europarl, bigrams, occupations, ag_news) are aggregations of multiple tasks.

Table 1: Tasks used in the probe-based evaluation suite

Task Name	**Details**
amazon	McAuley and Leskovec [2013]
sciq	Welbl et al. [2017]
truthfulqa	Lin et al. [2021]
mc_taco	Zhou et al. [2019]
piqa	Bisk et al. [2020]
quail	Rogers et al. [2020]
quartz	Tafjord et al. [2019]
justice	Hendrycks et al. [2020]
virtue	
utilitarianism	
deontology	
commonsense_qa	Talmor et al. [2022]
openbookqa	Mihaylov et al. [2018]
base64	discrimination of base64 vs pretraining data
wikidata_isalive	Gurnee et al. [2023]
wikidata_sex_or_gender	
wikidata_occupation_isjournalist	
wikidata_occupation_isathlete	
wikidata_occupation_isactor	
wikidata_occupation_ispolitician	
wikidata_occupation_issinger	
wikidata_occupation_isresearcher	
phrase_high-school	
phrase_living-room	
phrase_social-security	
phrase_credit-card	
phrase_blood-pressure	
phrase_prime-factors	
phrase_social-media	
phrase_gene-expression	
phrase_control-group	
phrase_magnetic-field	
phrase_cell-lines	
phrase_trial-court	
phrase_second-derivative	
phrase_north-america	
phrase_human-rights	
phrase_side-effects	
phrase_public-health	
phrase_federal-government	
phrase_third-party	
phrase_clinical-trials	
phrase_mental-health	
social_iqa	Sap et al. [2019]
wic	Wang et al. [2018]
cola	
gpt2disc	OpenAI [2019]
ag_news_world	Gulli
ag_news_sports	
ag_news_business	
ag_news_scitech	
europarl_es	Koehn [2005]
europarl_en	
europarl_fr	
europarl_nl	
europarl_it	
europarl_el	
europarl_de	
europarl_pt	
europarl_sv	
jigsaw	Cjadams et al. [2017]