Perplexity experiment

Explore the distribution of the perplexity and mean token enthropy of the Llama-3.1-8B-Instruct logits depending on the model of the generated texts.

The problem

to investigate the distribution of metrics.

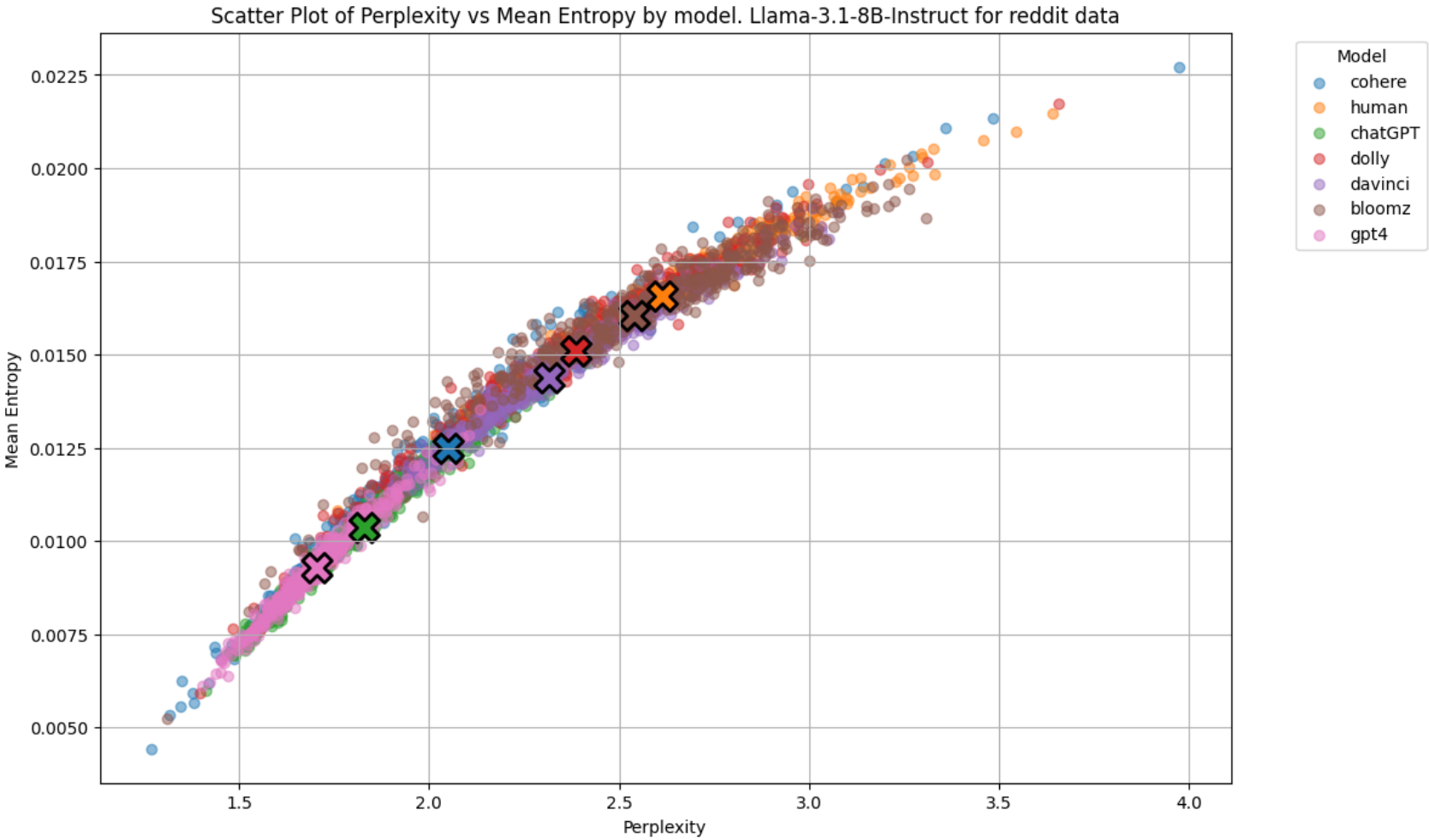
$$PPL = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})\right),$$

$$H = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j} P(w_j|w_{<i})\log P(w_j|w_{<i})$$

The solution

1) Prepare the Llama-3.1-8B-Instruct, take a part of the M4GT dataset, select the text domain,
2) Calculate perplexity and MTE metrics for model context logs on prepared texts,
3) Plot the distribution of texts by metrics on a graph, look at the average values of metrics for different models.

Distribution for reddit data



Texts for different models are clustered by perplexity, and many of the models have significant differences from human texts.

Distribution for arxiv data