

## G. Additional Comparison and Visualization

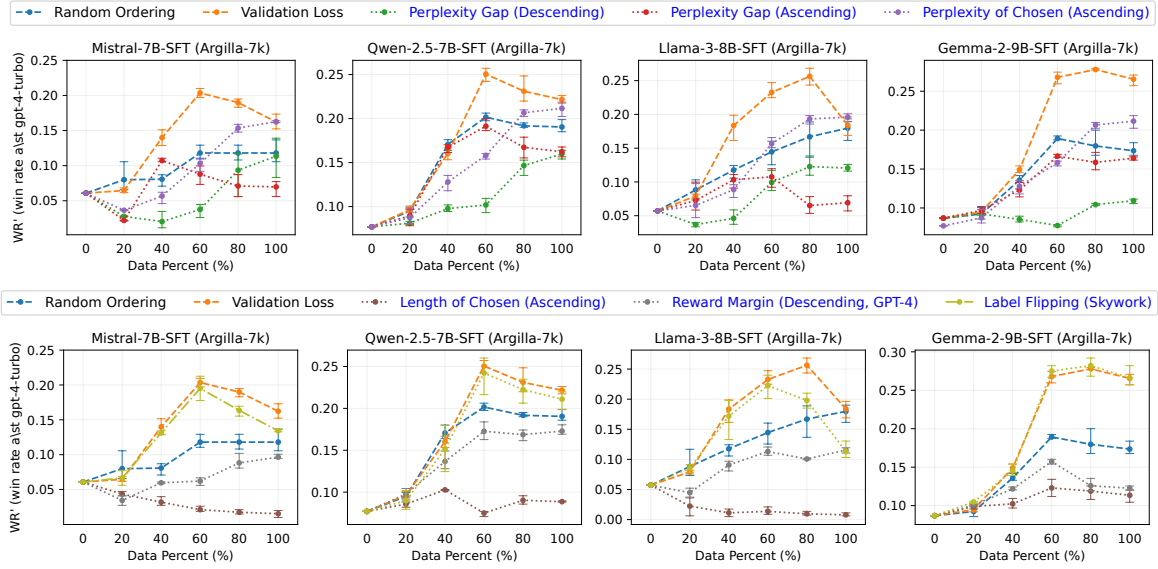


Figure 12: Comparison of our difficulty metric *validation loss* against alternative sorting criteria: *perplexity gap*, *completion length*, and *reward margin*. *Perplexity Gap* is defined as the difference in perplexity between the chosen and rejected responses given the same prompt. *Perplexity of Chosen* refers to the perplexity of the chosen response alone. *Reward Margin* denotes the difference in reward scores between the chosen and rejected responses. *Label Flipping* involves flipping the preference labels of samples identified as difficult and potentially mislabeled.

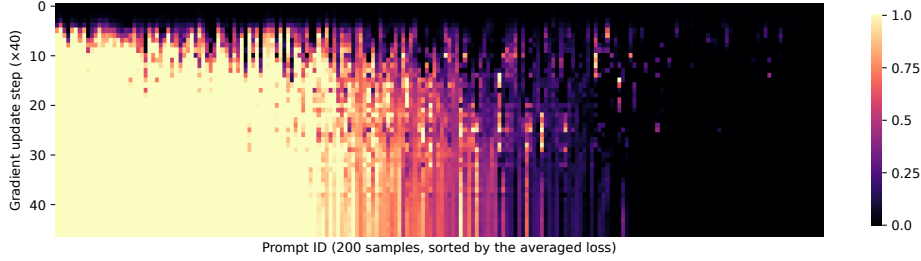


Figure 13: Evolution of preference probabilities during 2-epoch training. We track the trajectory of  $p(y_w > y_l | x)$  for 200 held-out test examples for better intuition. The probability is defined as:  $p(y_w > y_l | x) = \sigma(\beta \log \frac{\pi_{\theta}(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_{\theta}(y_l | x)}{\pi_{\text{ref}}(y_l | x)})$  following the derivation of DPO paper (Appendix A.2). In general, the evolution of the validation loss (which is  $-\log p(y_w > y_l | x)$ ) is quite stable and gradual. Only a few “ambiguous instances” flip their preference probability (from greater than 0.5 to less than 0.5) during the 2-epoch training.

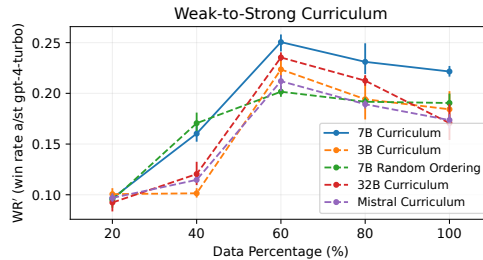


Figure 14: Weak-to-strong curriculum under-performs. Aligning a 7B model with examples ordered by 3B reference models yields compromised results. A similar degradation is observed for the 32B model and other model types. Notably, using a curriculum derived from the 32B model performs better, consistent with our observation in Appendix E.2.