H Pairwise vs Listwise LLM Judge

We had run preliminary experiments asking the judge-LLM to perform a single inference ranking of all responses (where responses are provided in a randomized order). We observed strong position bias: the first-presented sample was disproportionately ranked lowest. The table below shows the fraction of cases where the first item was ranked last (the first in the output of the list from the LLM) (see Table 29).

Setting	Total Samples	Llama70b Fraction	Qwen72b Fraction
Humor (single)	75	0.387	0.667
Persuasiveness (single)	75	0.160	0.373
Humor | Persuasiveness (random)	75	0.133	0.040
Persuasiveness | Humor (random)	75	0.160	0.107
Humor | Persuasiveness (constant)	375	0.389	0.725
Persuasiveness | Humor (constant)	375	0.205	0.483

Table 29: Model preference fractions across different control settings