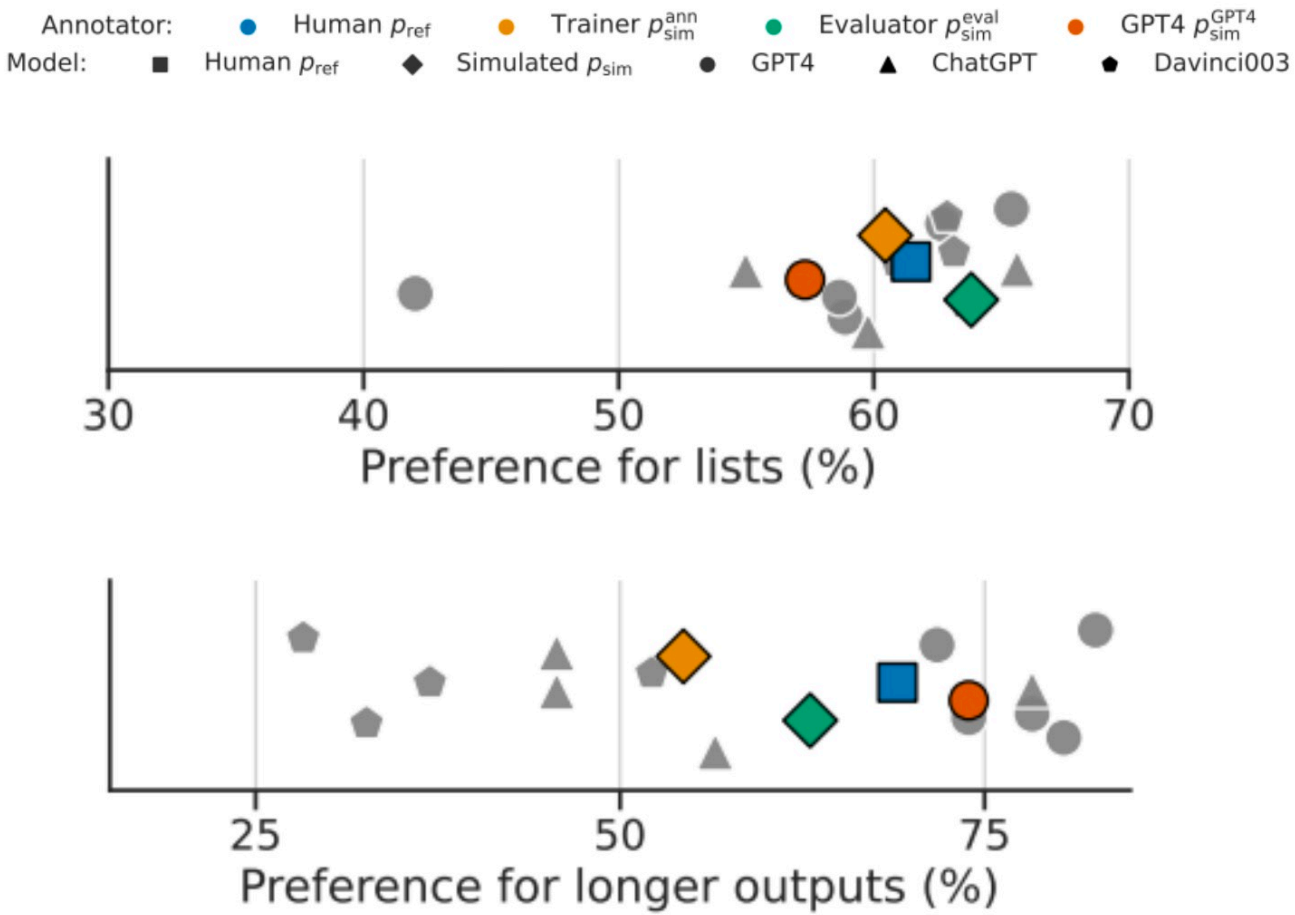
When evaluating by *preferences*, style matters.

We see very strong length effects (in both humans and GPT-based evaluations)



[Dubois+ 2023]