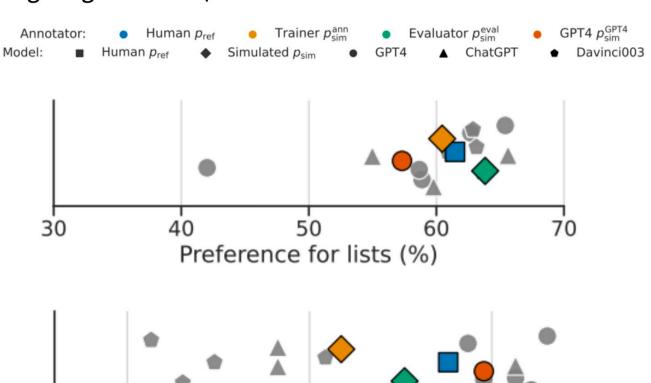
When evaluating by preferences, style matters.

25

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



50

Preference for longer outputs (%)

[Dubois+ 2023]