

Calibration with GPT-3.5 turbo vs ground truth prediction

