

Accuracy and Unknown ratio over 40 questions
Generator: llama3-70b-8192 | Judge: GPT-4o | Dataset: eligibility_dataset

