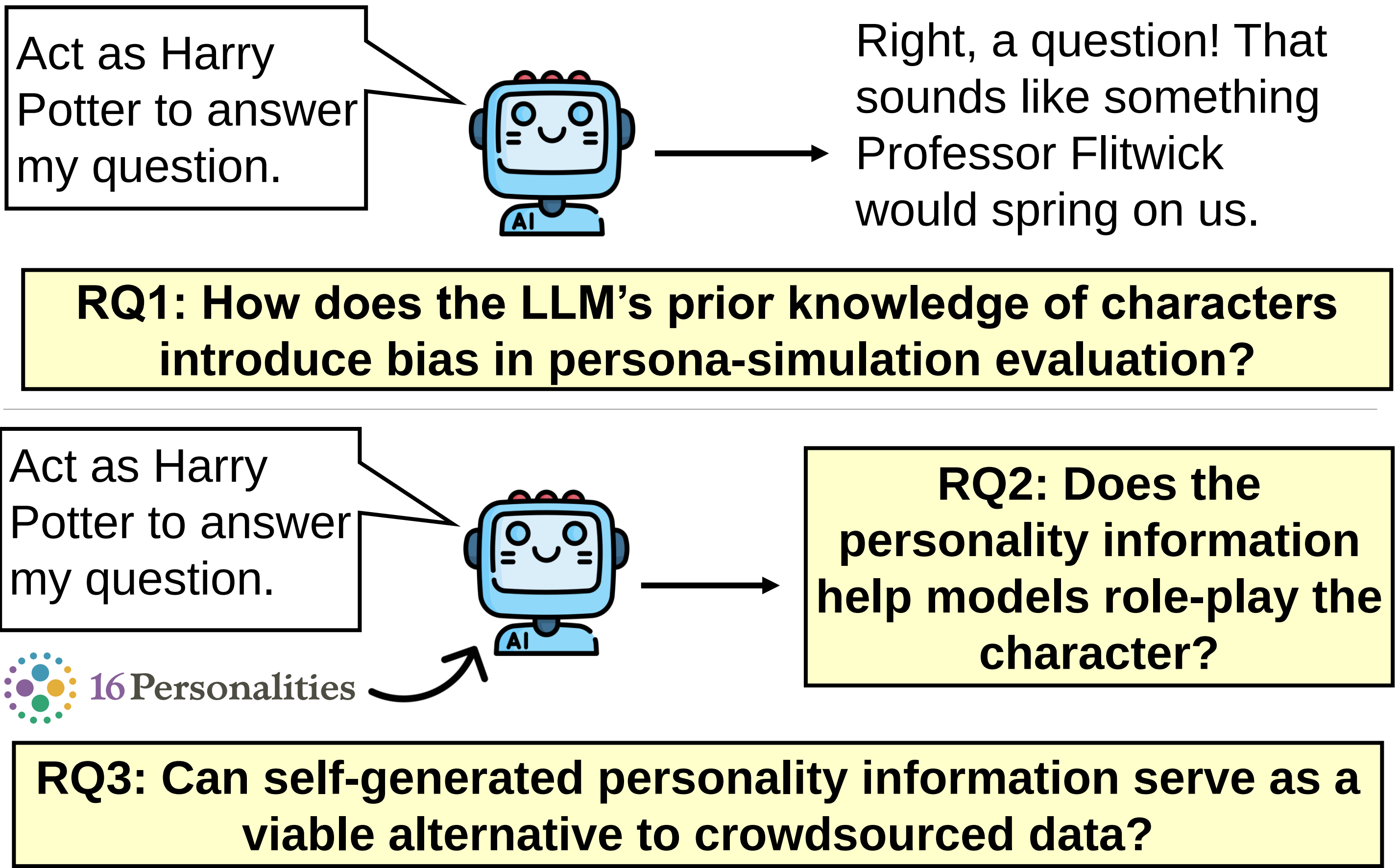


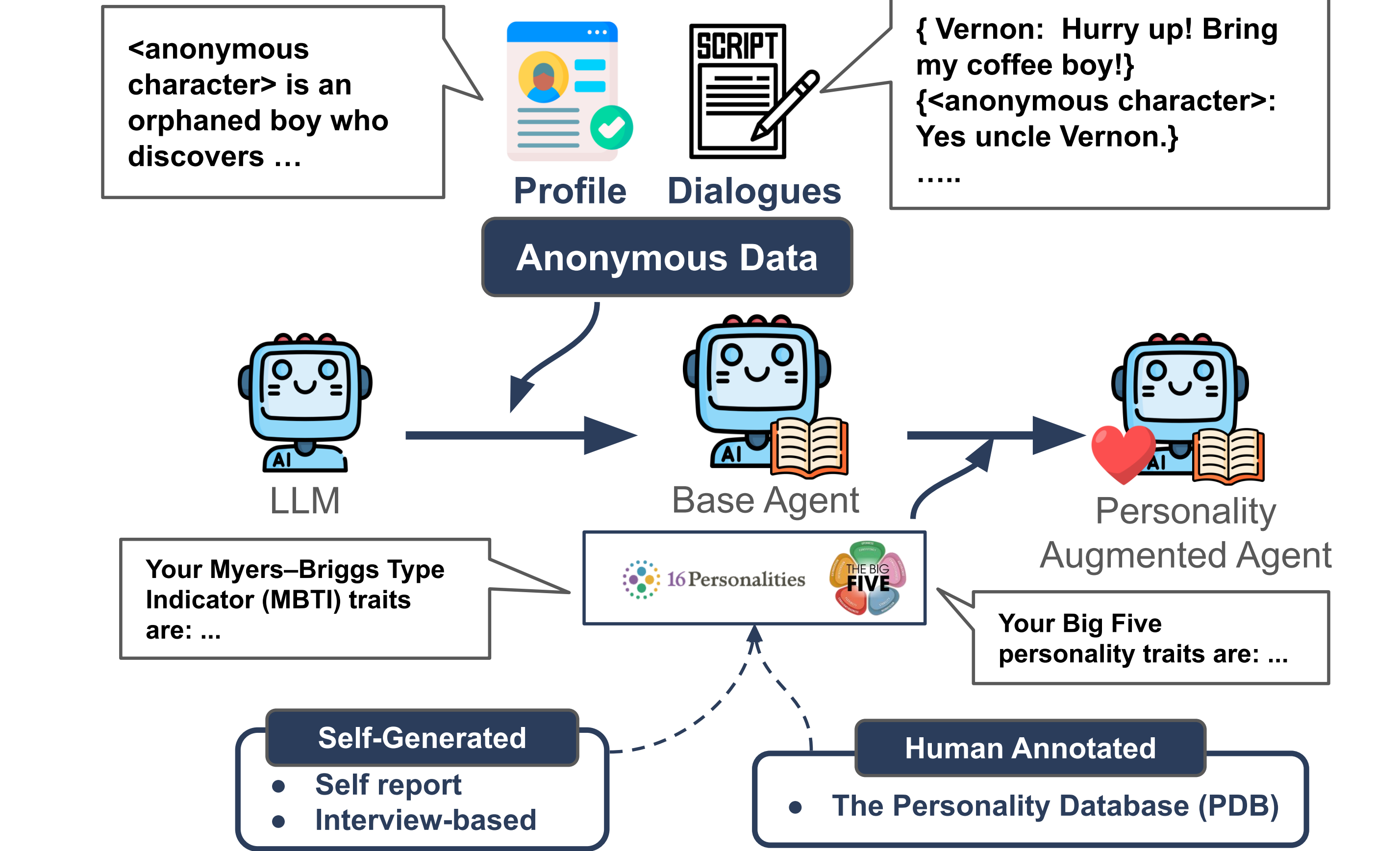
Rethinking Role-Playing Evaluation: Anonymous Benchmarking and A Systematic Study of Personality Effects

Ji-Lun Peng Yun-Nung (Vivian) Chen

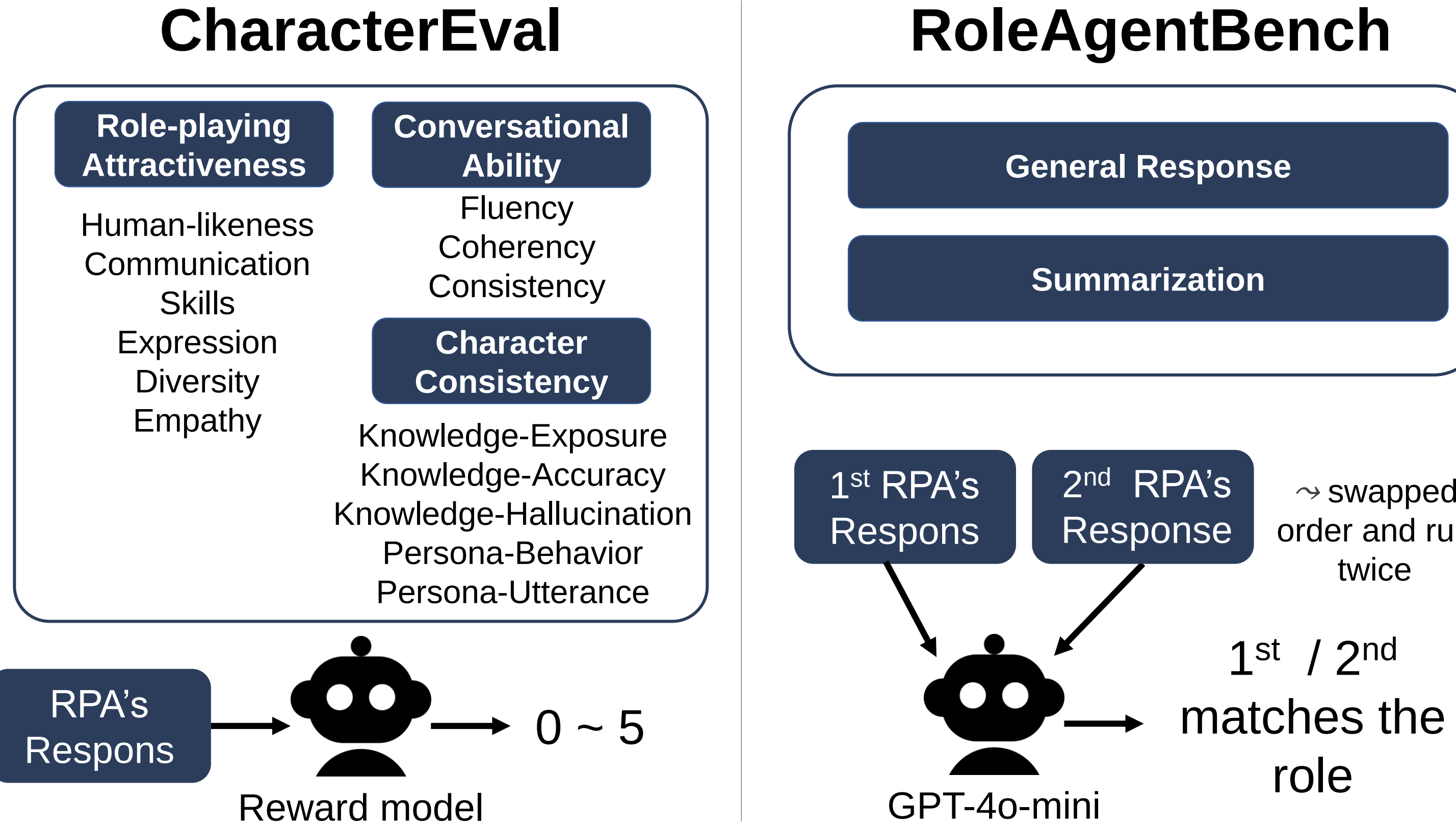
Motivation



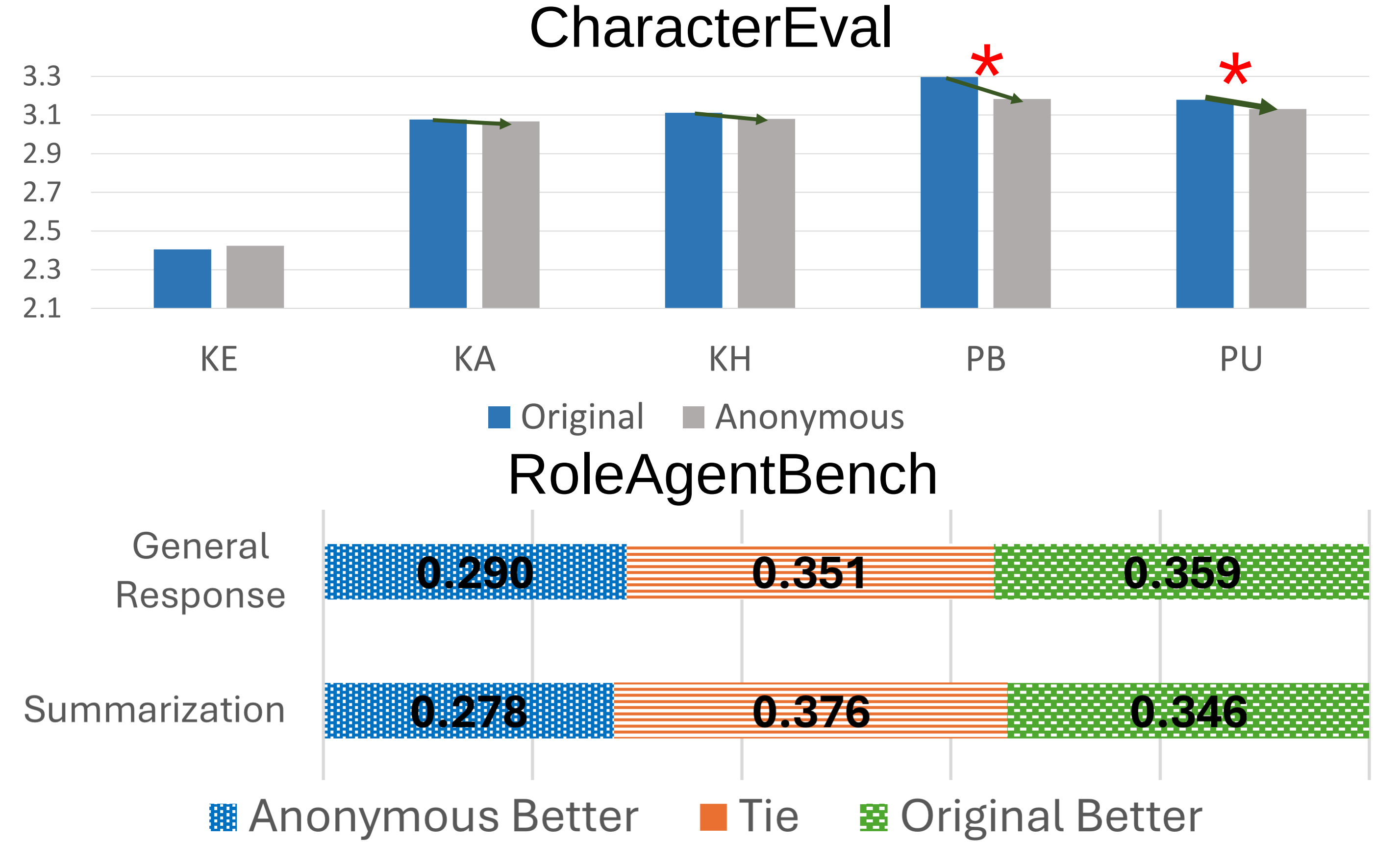
Framework



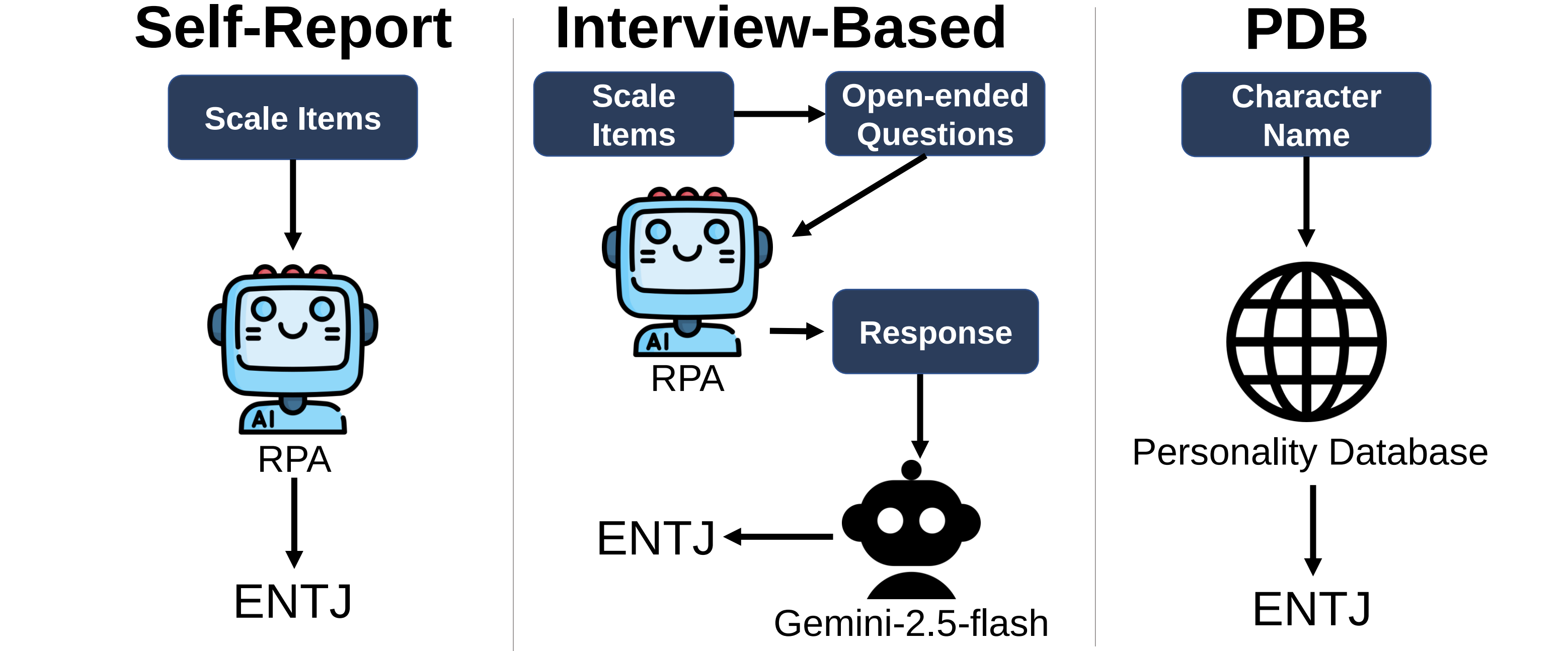
Benchmarks



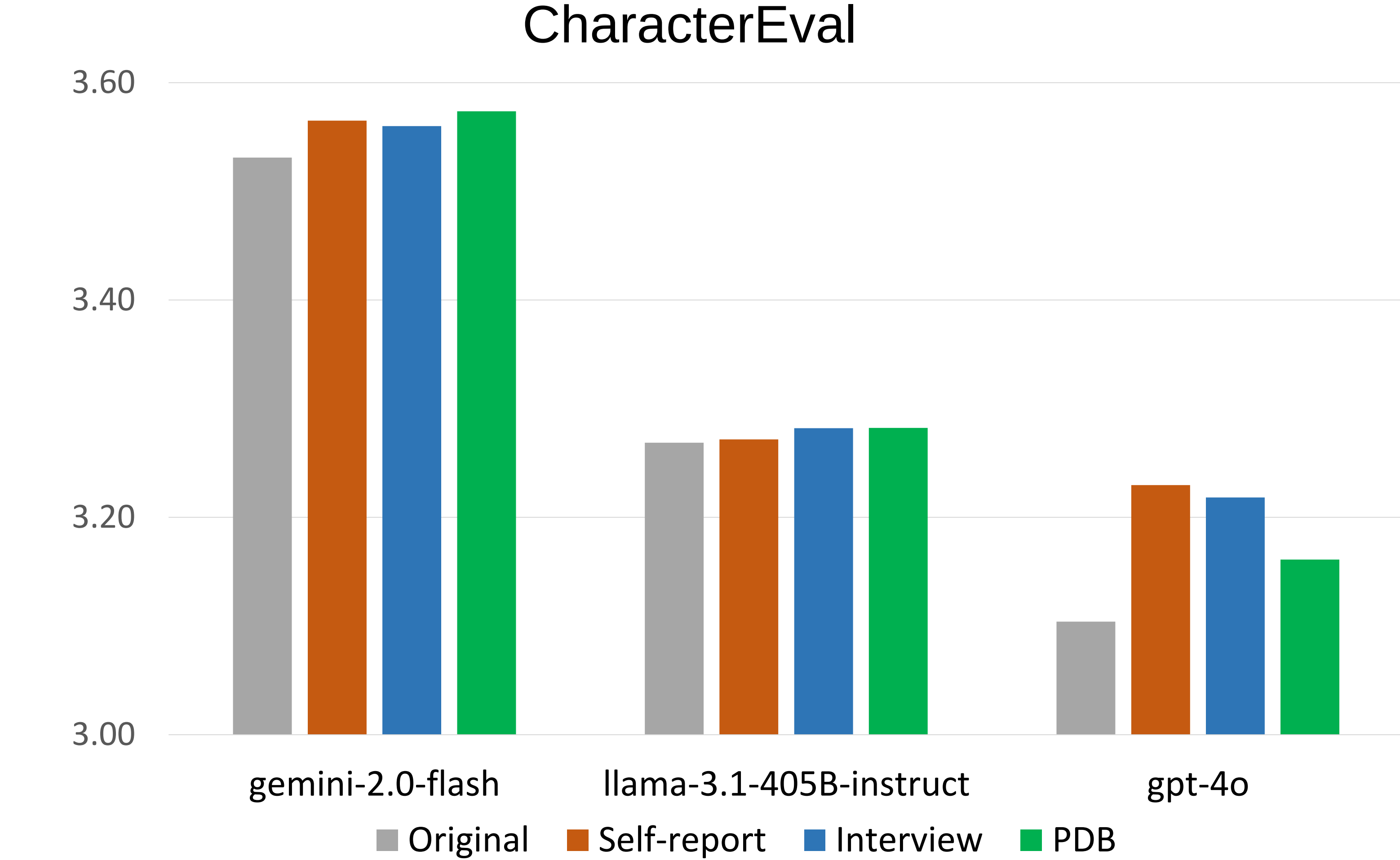
Results: Character Names as Cues



Personality Sources of Role-Playing Agents



Results: Personality Helps LLMs Role-Playing



RoleAgentBench

	gemini-2.0-flash	llama-3.1-405B-instruct	gpt-4o
SelfReport			
General Response	0.428 0.289 0.283	0.201 0.606 0.193	0.347 0.405 0.248
Summarization	0.315 0.361 0.324	0.199 0.615 0.186	0.324 0.427 0.249
Interview			
General Response	0.415 0.289 0.296	0.218 0.567 0.215	0.337 0.381 0.282
Summarization	0.455 0.279 0.266	0.205 0.581 0.214	0.269 0.443 0.288
PDB			
General Response	0.427 0.294 0.279	0.203 0.604 0.193	0.35 0.377 0.273
Summarization	0.373 0.326 0.301	0.198 0.641 0.161	0.342 0.411 0.247
General Response	0.406 0.323 0.271	0.243 0.545 0.212	0.325 0.368 0.307
Summarization	0.386 0.321 0.293	0.223 0.526 0.251	0.321 0.405 0.274
General Response	0.412 0.304 0.284	0.204 0.616 0.18	0.374 0.384 0.242
Summarization	0.358 0.336 0.306	0.212 0.592 0.196	0.345 0.434 0.221
General Response	0.382 0.313 0.305	0.228 0.532 0.24	0.359 0.387 0.254
Summarization	0.419 0.288 0.293	0.2 0.581 0.219	0.312 0.33 0.358

MBTI Personality Win Tie Personality Loss

Big5 Personality Win Tie Personality Loss

Conclusion

- Anonymizing character names degrades performance, underscoring the value of anonymous setting for more rigorous evaluations.
- Augmenting prompts with personality information improves RPA performance.
- Self-generated personality information offers a flexible and effective alternative to crowdsourced.

