Reference scores ●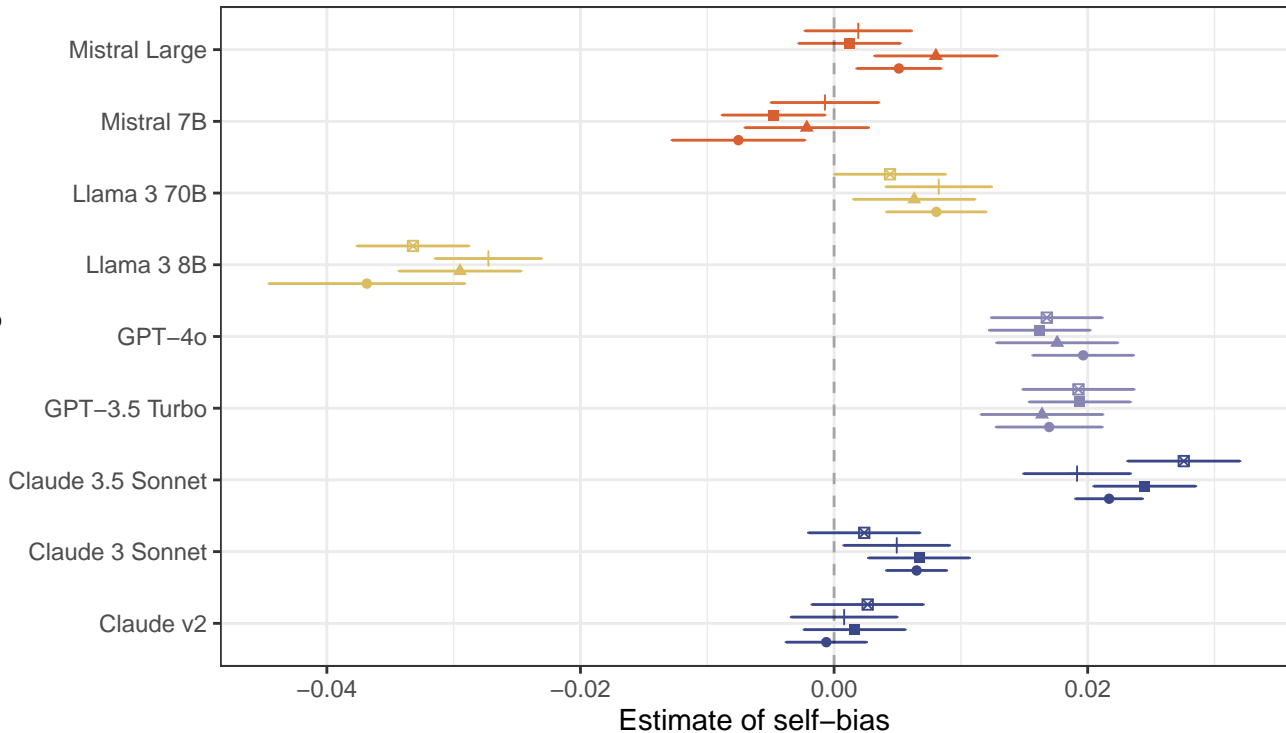 Human ▲ Claude ■ Llama 3 + GPT ⊠ Mistral