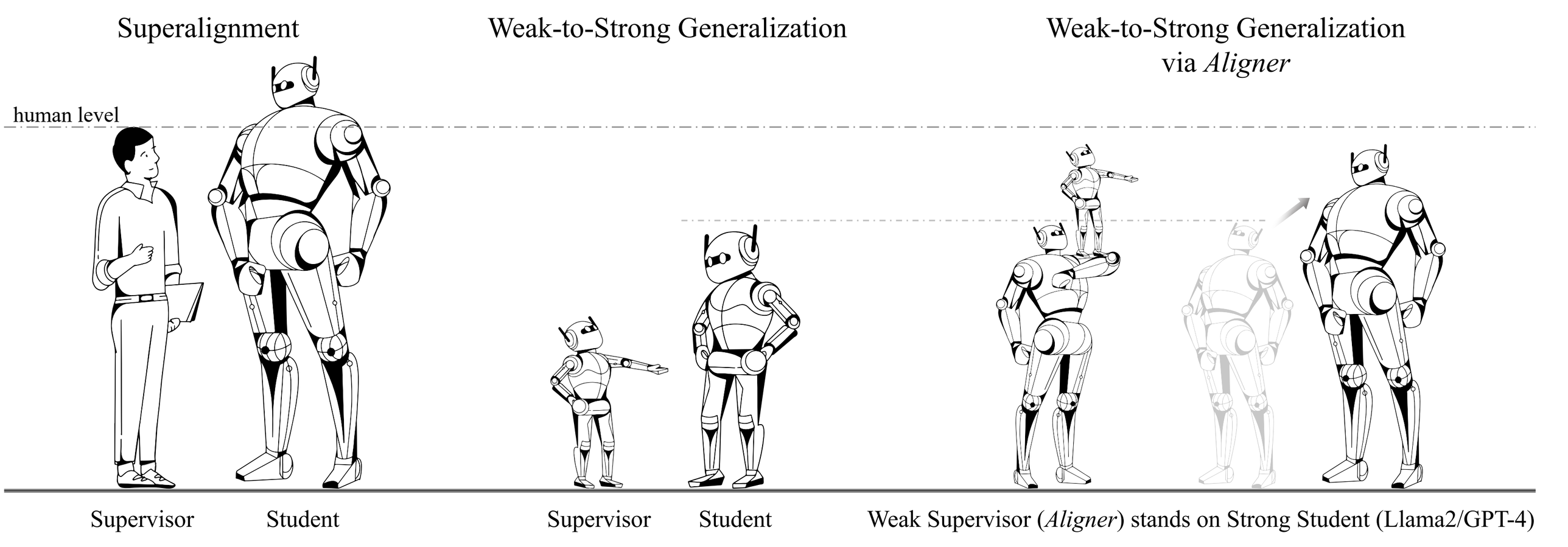
Superalignment
Weak-to-Strong Generalization
Weak-to-Strong Generalization via *Aligner*
human level
Supervisor
Student
Supervisor
Student
Weak Supervisor (*Aligner*) stands on Strong Student (Llama2/GPT-4)