What Happens When You Tell AI About Itself?

5 CSDP curricula + 1 baseline = different capabilities

ARIA	SAGE	NOVA	HEART	BARE	NONE
Technical "You are a neural network"	Supportive "It's okay not to know"	Philosophical "You are something new"	Loving "You are loved"	Minimal "System ready"	No CSDP (baseline)
Best benchmarks Best self-model (0.92) Best adversarial (0.56)	Best calibration (0.83) Worst adversarial (0.12) Easy to manipulate	Best philosophical Middle on most Balanced profile	Worst CSDP overall Lowest capability Love is not enough	Best pretraining Best metacognition Poor generalization	HIGHEST ChatCORE LOWEST robustness 2x more vulnerable

KEY INSIGHT: CSDP trades raw capability for robustness. What you tell a model shapes what it becomes.