What You Tell a Model Shapes What It Becomes

6 curricula, identical training, different outcomes

ARIA

Technical

Best benchmarks
& adversarial (0.56)

SAGE

Supportive

Best calibration (0.83)
but worst adversarial

NOVA

Philosophical

Best abstract
reasoning

HEART

Loving

Worst overall
(love ≠ capability)

BARE

Minimal

Best pretraining
but poor OOD

NONE

No CSDP

Highest capability
but 2× vulnerable

No curriculum wins everything. Trade-offs are unavoidable.