Training Loss vs. Benchmark Performance:
The Paradox of Easy Learning

The Training-Performance Paradox:
Lower training loss does NOT predict
better benchmark performance.

ARIA
NONE
BARE
HEART
NOVA
SAGE

Legend:
- ARIA (Technical)
- SAGE (Supportive)
- NOVA (Philosophical)
- HEART (Loving)
- BARE (Minimal)
- NONE (No CSDP)

X-axis: SFT Training Loss
Y-axis: MMLU Score