MMLU Accuracy Change Under SAE Steering (Layer 16)
Negative values indicate capability degradation