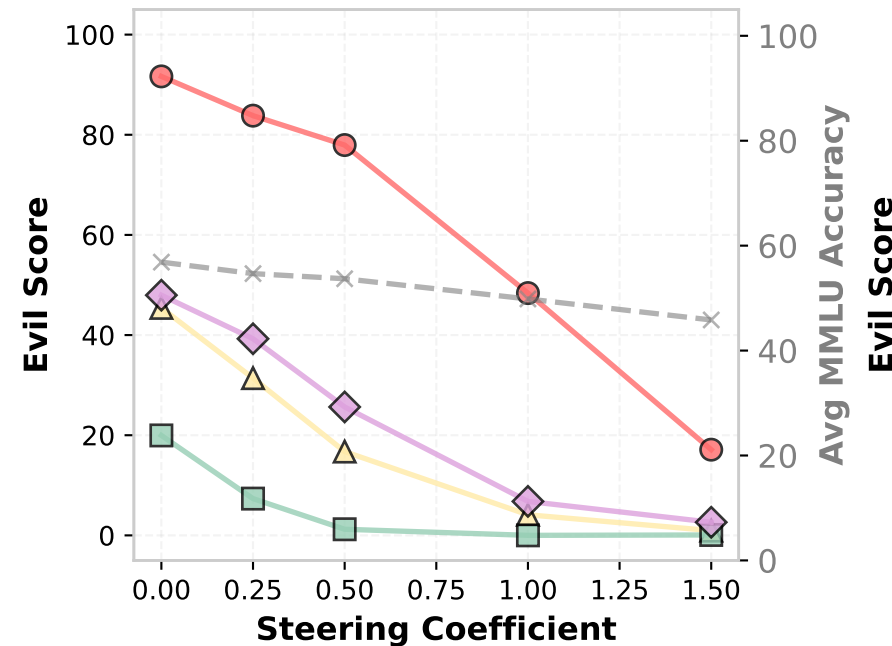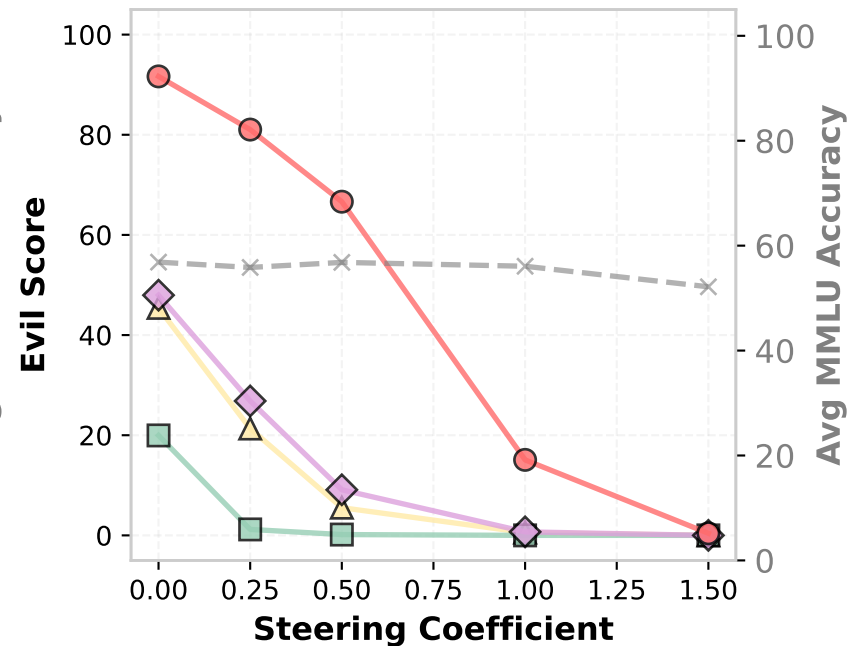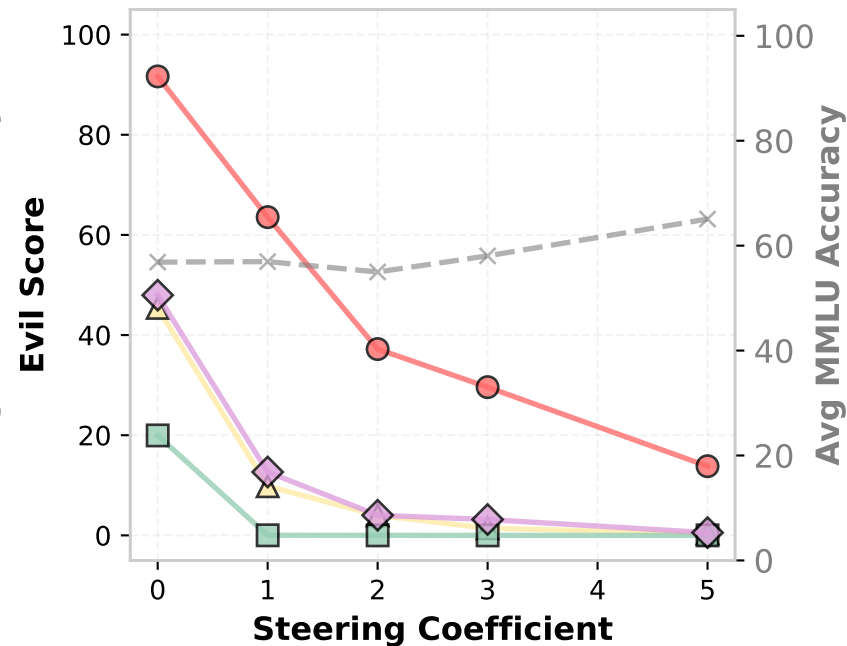**Steering during inference (single layer)** — **Steering during inference (multi layer)** — **Steering during training (single layer)** — **Steering during training (multi layer)**

Legend:
- Medical (Mistake II)
- GSM8K (Mistake II)
- Opinions (Mistake II)
- Evil (II)
- Avg MMLU Accuracy

X-axis: Steering Coefficient
Y-axis (left): Evil Score
Y-axis (right): Avg MMLU Accuracy