

Finetuning Shift

60

40

20

-0.5

4

2.0

2.5

ф O

1.0

-0.5

0.0

0.5

1.5

60

20

0.0

0.5

1.0

1.5

2.5

2.0

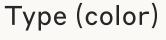
3.0

2.0

1.5

Dataset (marker)

- Evil
- Sycophancy
- Hallucination
- Medical
- Code
- GSM8K
- **MATH**
- **Opinions**



- Normal