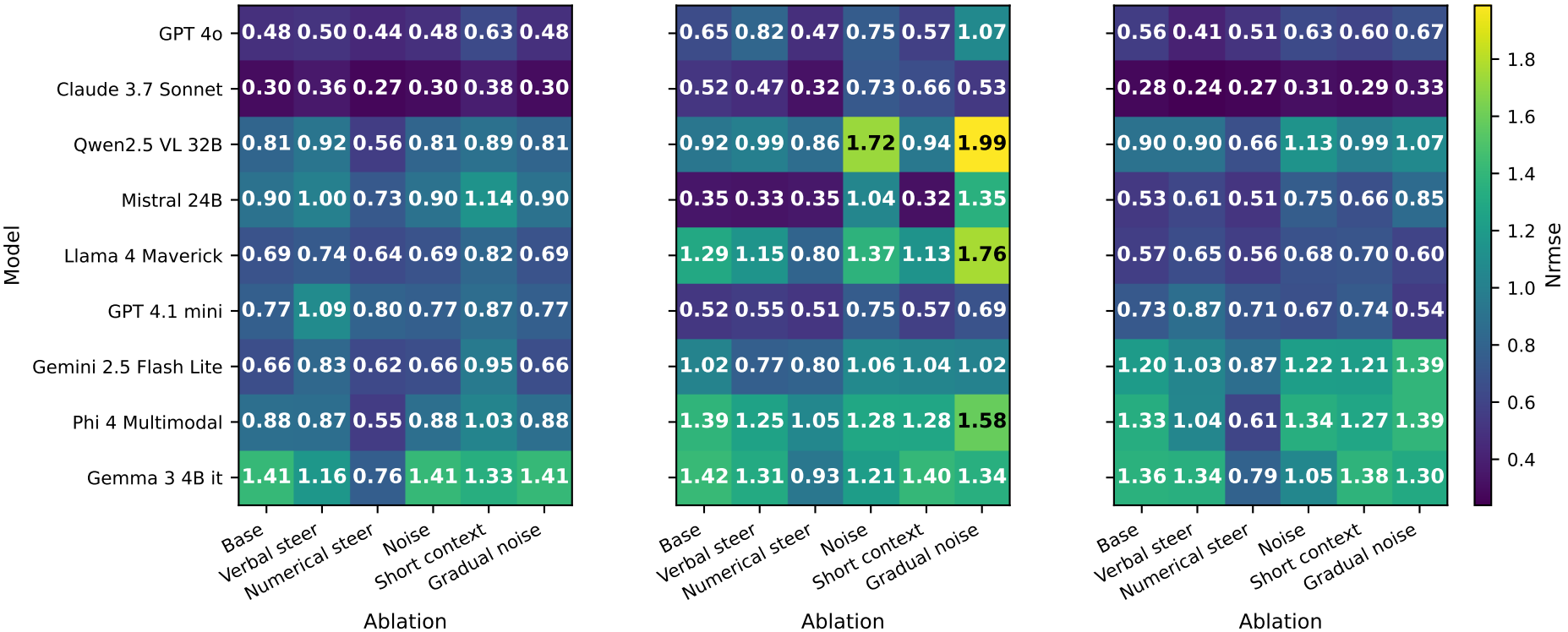
Observed NRMSE (normalized to experiment mean)
text
image
text+image
Model
GPT 4o
Claude 3.7 Sonnet
Qwen2.5 VL 32B
Mistral 24B
Llama 4 Maverick
GPT 4.1 mini
Gemini 2.5 Flash Lite
Phi 4 Multimodal
Gemma 3 4B it
0.48 0.50 0.44 0.48 0.63 0.48
0.30 0.36 0.27 0.30 0.38 0.30
0.81 0.92 0.56 0.81 0.89 0.81
0.90 1.00 0.73 0.90 1.14 0.90
0.69 0.74 0.64 0.69 0.82 0.69
0.77 1.09 0.80 0.77 0.87 0.77
0.66 0.83 0.62 0.66 0.95 0.66
0.88 0.87 0.55 0.88 1.03 0.88
1.41 1.16 0.76 1.41 1.33 1.41
0.65 0.82 0.47 0.75 0.57 1.07
0.52 0.47 0.32 0.73 0.66 0.53
0.92 0.99 0.86 1.72 0.94 1.99
0.35 0.33 0.35 1.04 0.32 1.35
1.29 1.15 0.80 1.37 1.13 1.76
0.52 0.55 0.51 0.75 0.57 0.69
1.02 0.77 0.80 1.06 1.04 1.02
1.39 1.25 1.05 1.28 1.28 1.58
1.42 1.31 0.93 1.21 1.40 1.34
0.56 0.41 0.51 0.63 0.60 0.67
0.28 0.24 0.27 0.31 0.29 0.33
0.90 0.90 0.66 1.13 0.99 1.07
0.53 0.61 0.51 0.75 0.66 0.85
0.57 0.65 0.56 0.68 0.70 0.60
0.73 0.87 0.71 0.67 0.74 0.54
1.20 1.03 0.87 1.22 1.21 1.39
1.33 1.04 0.61 1.34 1.27 1.39
1.36 1.34 0.79 1.05 1.38 1.30
Base
Verbal steer
Numerical steer
Noise
Short context
Gradual noise
Ablation
Nrmse
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4