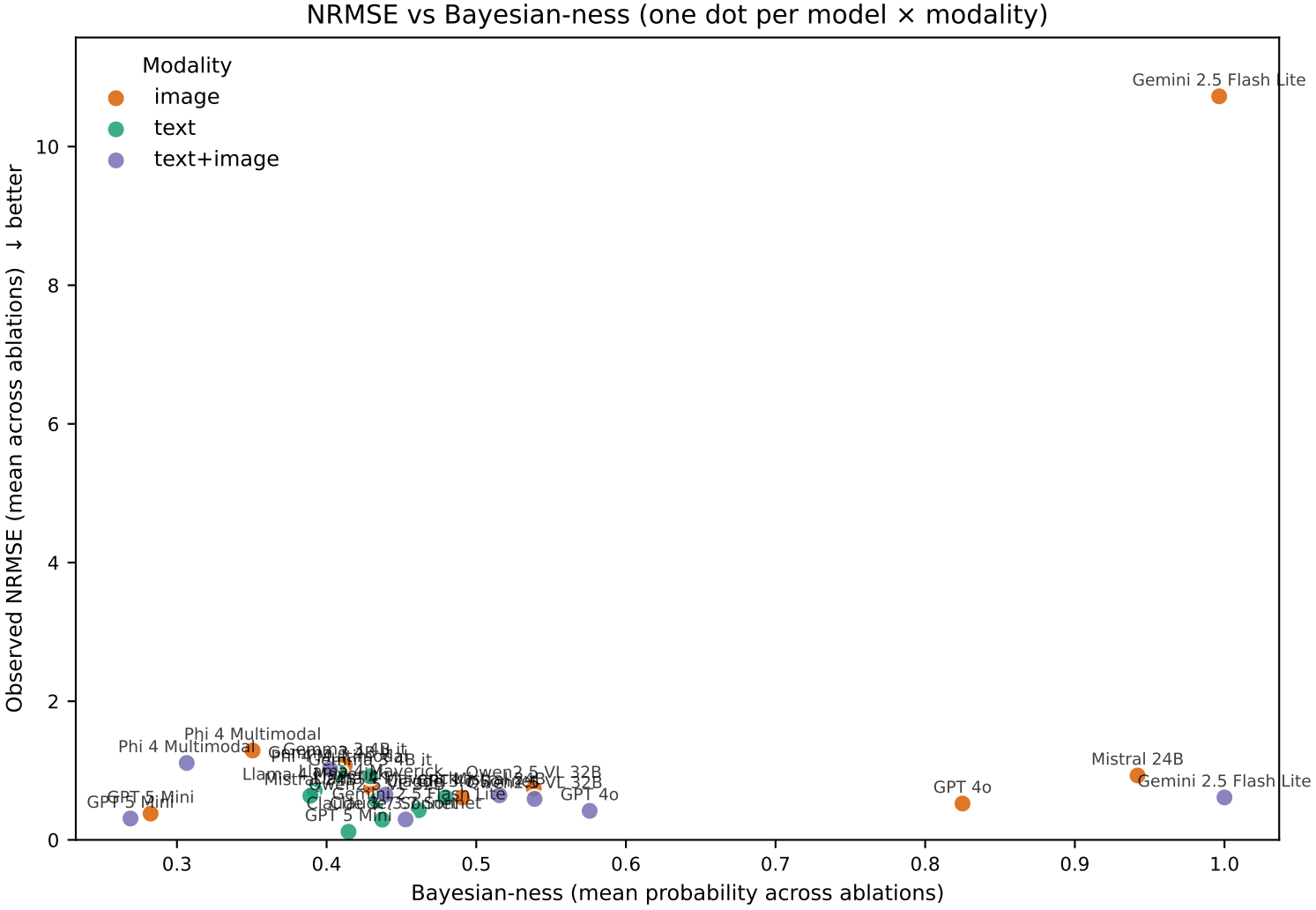
NRMSE vs Bayesian-ness (one dot per model × modality)
Modality
image
text
text+image
Gemini 2.5 Flash Lite
Phi 4 Multimodal
Phi 4 Multimodal
Gemma 3 4B it
Llama 4 Maverick
Qwen2.5-VL 32B
GPT 4o
GPT 5 Mini
GPT 5 Mini
Gemini 2.5 Flash Lite
Claude 3.5 Sonnet
Mistral 24B
Gemini 2.5 Flash Lite
GPT 4o
Observed NRMSE (mean across ablations) ↓ better
Bayesian-ness (mean probability across ablations)
0
2
4
6
8
10
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0