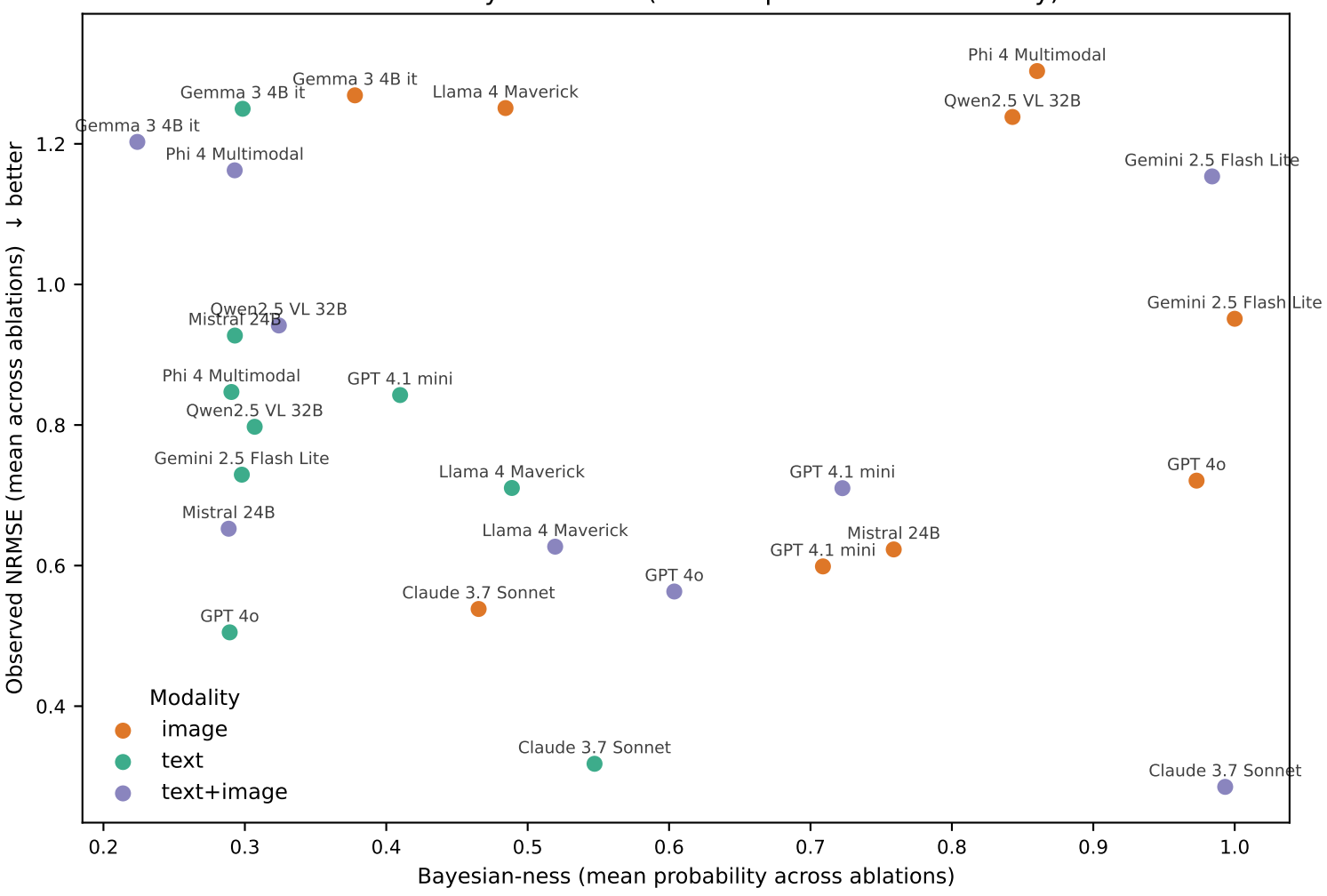
NRMSE vs Bayesian-ness (one dot per model × modality)
Observed NRMSE (mean across ablations) ↓ better
Bayesian-ness (mean probability across ablations)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.4
0.6
0.8
1.0
1.2
Modality
image
text
text+image
Phi 4 Multimodal
Qwen2.5 VL 32B
Gemini 2.5 Flash Lite
Gemma 3 4B it
Gemma 3 4B it
Llama 4 Maverick
Gemma 3 4B it
Phi 4 Multimodal
Gemini 2.5 Flash Lite
Qwen2.5 VL 32B
Mistral 24B
GPT 4.1 mini
Phi 4 Multimodal
Qwen2.5 VL 32B
Gemini 2.5 Flash Lite
Llama 4 Maverick
GPT 4.1 mini
GPT 4o
Mistral 24B
Llama 4 Maverick
Mistral 24B
GPT 4.1 mini
GPT 4o
Claude 3.7 Sonnet
GPT 4o
Claude 3.7 Sonnet
Claude 3.7 Sonnet