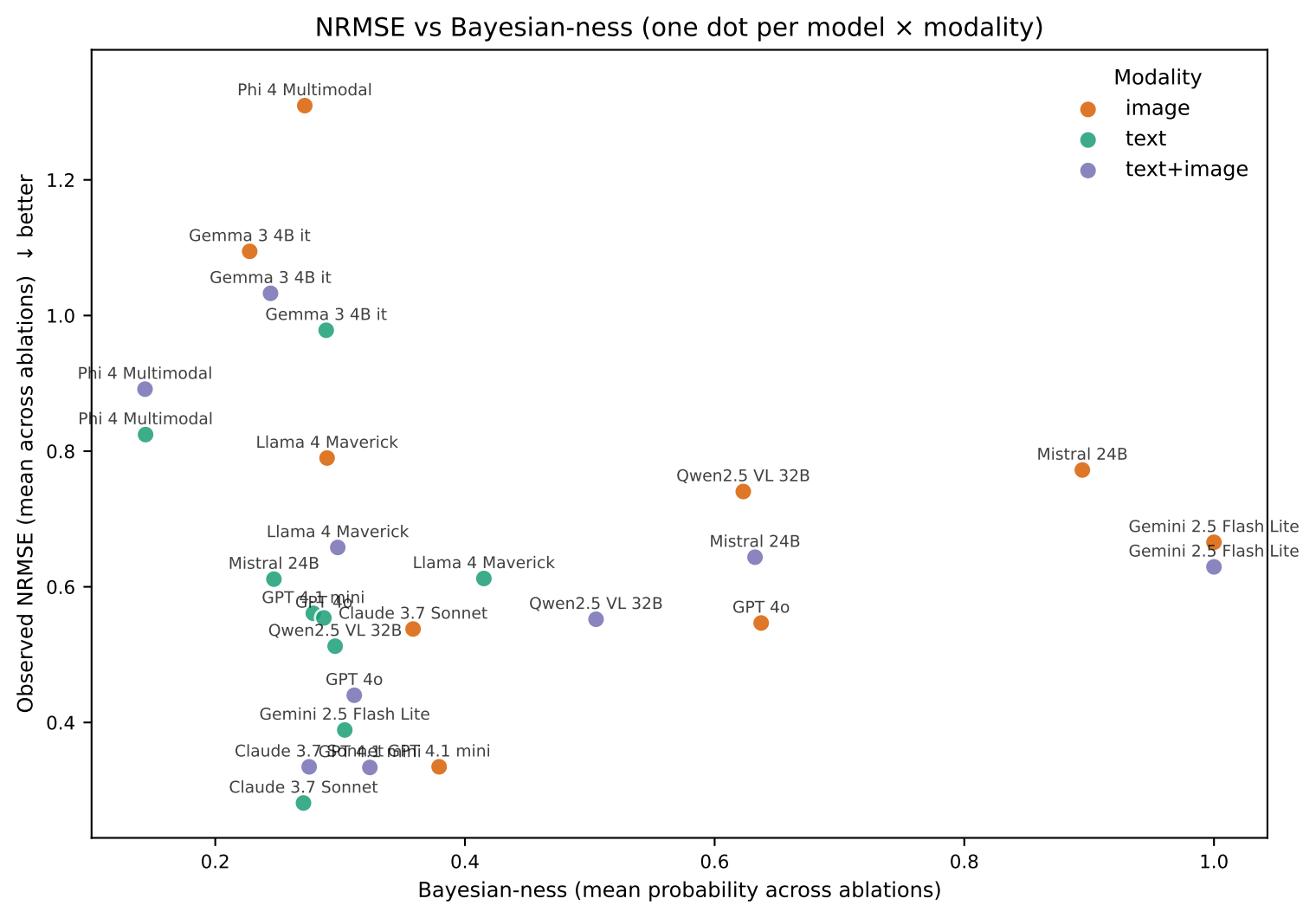
NRMSE vs Bayesian-ness (one dot per model × modality)
Modality
image
text
text+image
Observed NRMSE (mean across ablations) ↓ better
Bayesian-ness (mean probability across ablations)
0.2
0.4
0.6
0.8
1.0
1.2
Phi 4 Multimodal
Gemma 3 4B it
Gemma 3 4B it
Gemma 3 4B it
Phi 4 Multimodal
Phi 4 Multimodal
Llama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