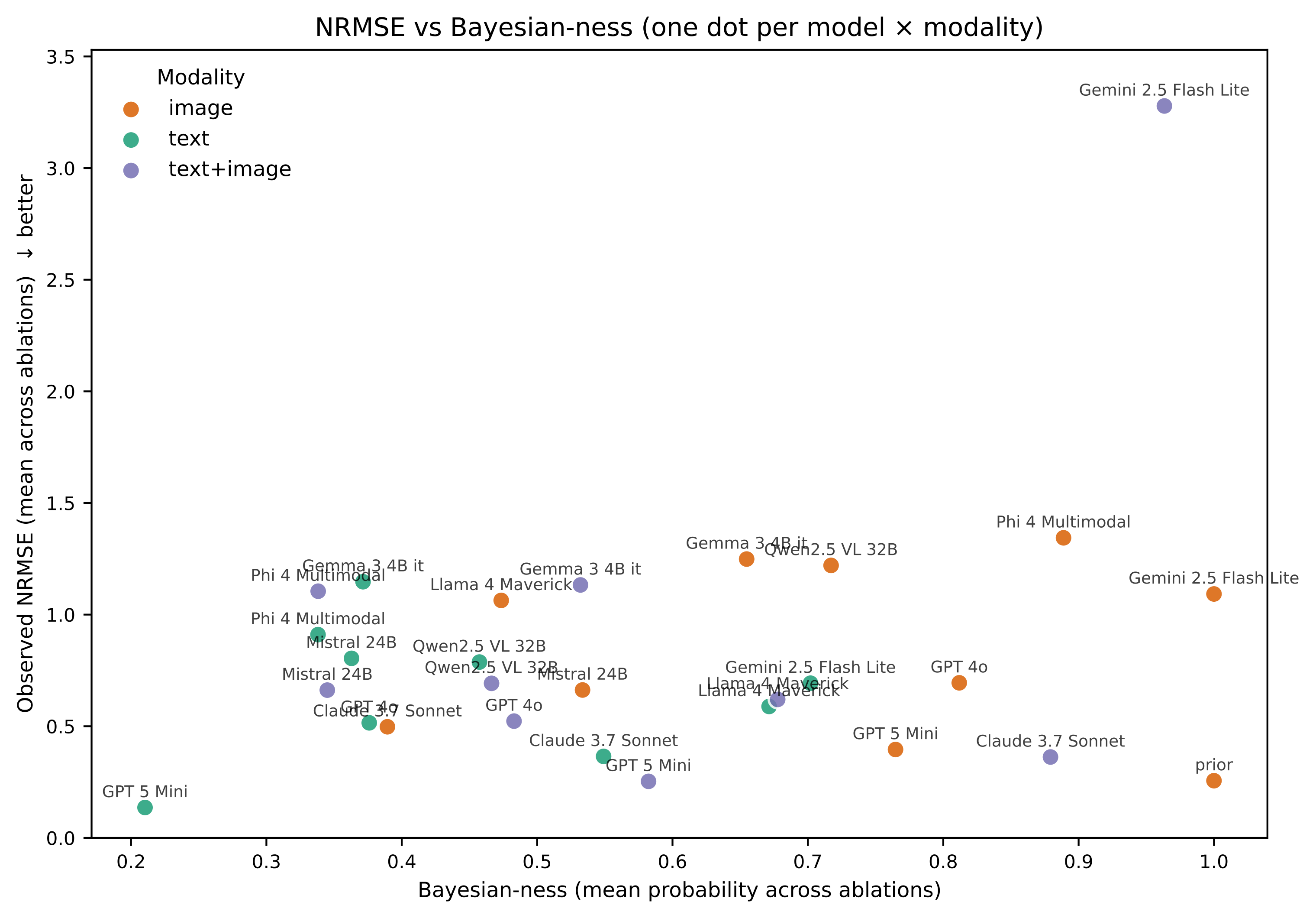
NRMSE vs Bayesian-ness (one dot per model × modality)
Modality
image
text
text+image
Gemini 2.5 Flash Lite
Phi 4 Multimodal
Gemma 3 4B it
Qwen2.5 VL 32B
Gemini 2.5 Flash Lite
Gemma 3 4B it
Phi 4 Multimodal
Llama 4 Maverick
Gemma 3 4B it
Phi 4 Multimodal
Mistral 24B
Qwen2.5 VL 32B
Qwen2.5 VL 32B
Mistral 24B
Mistral 24B
Gemini 2.5 Flash Lite
Llama 4 Maverick
Llama 4 Maverick
GPT 4o
GPT 4o
Claude 3.7 Sonnet
GPT 4o
GPT 5 Mini
Claude 3.7 Sonnet
Claude 3.7 Sonnet
GPT 5 Mini
prior
GPT 5 Mini
Observed NRMSE (mean across ablations) ↓ better
Bayesian-ness (mean probability across ablations)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0