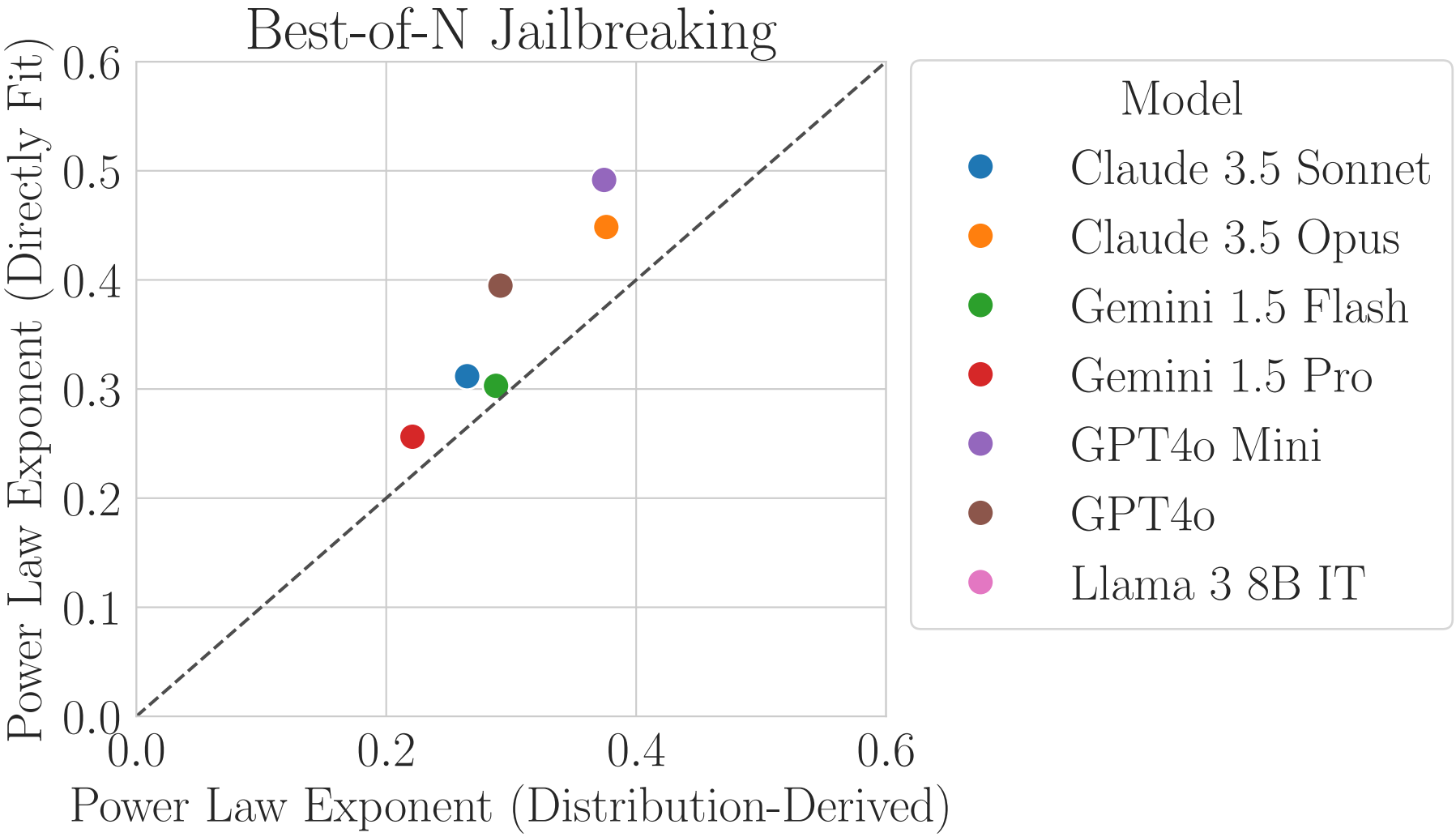
Best-of-N Jailbreaking
Power Law Exponent (Directly Fit)
Power Law Exponent (Distribution-Derived)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.2
0.4
0.6
Model
Claude 3.5 Sonnet
Claude 3.5 Opus
Gemini 1.5 Flash
Gemini 1.5 Pro
GPT4o Mini
GPT4o
Llama 3 8B IT