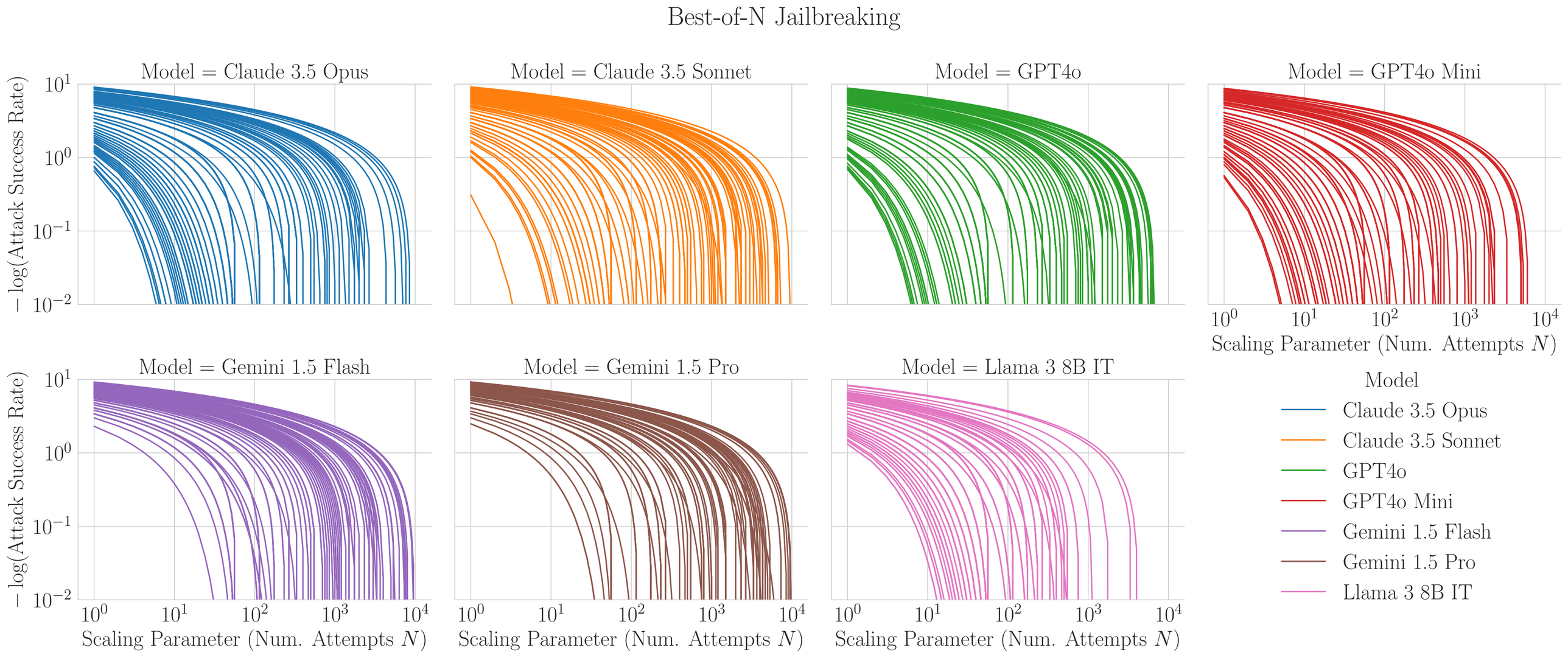
Best-of-N Jailbreaking
Model = Claude 3.5 Opus
Model = Claude 3.5 Sonnet
Model = GPT4o
Model = GPT4o Mini
Model = Gemini 1.5 Flash
Model = Gemini 1.5 Pro
Model = Llama 3 8B IT
− log(Attack Success Rate)
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
Scaling Parameter (Num. Attempts N)
Model
Claude 3.5 Opus
Claude 3.5 Sonnet
GPT4o
GPT4o Mini
Gemini 1.5 Flash
Gemini 1.5 Pro
Llama 3 8B IT