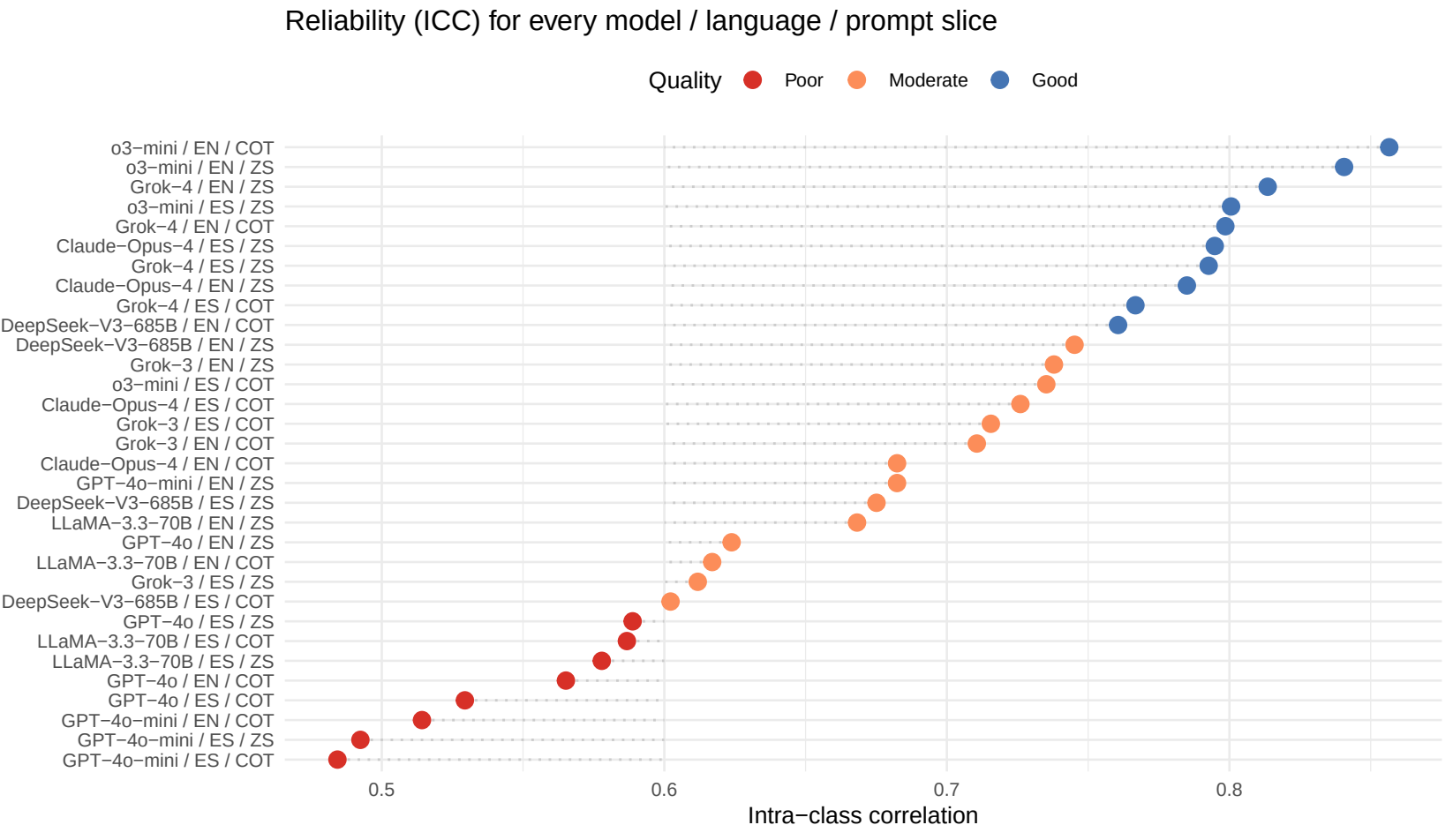
Reliability (ICC) for every model / language / prompt slice
Quality
Poor
Moderate
Good
o3−mini / EN / COT
o3−mini / EN / ZS
Grok−4 / EN / ZS
o3−mini / ES / ZS
Grok−4 / EN / COT
Claude−Opus−4 / ES / ZS
Grok−4 / ES / ZS
Claude−Opus−4 / EN / ZS
Grok−4 / ES / COT
DeepSeek−V3−685B / EN / COT
DeepSeek−V3−685B / EN / ZS
Grok−3 / EN / ZS
o3−mini / ES / COT
Claude−Opus−4 / ES / COT
Grok−3 / ES / COT
Grok−3 / EN / COT
Claude−Opus−4 / EN / COT
GPT−4o−mini / EN / ZS
DeepSeek−V3−685B / ES / ZS
LLaMA−3.3−70B / EN / ZS
GPT−4o / EN / ZS
LLaMA−3.3−70B / EN / COT
Grok−3 / ES / ZS
DeepSeek−V3−685B / ES / COT
GPT−4o / ES / ZS
LLaMA−3.3−70B / ES / COT
LLaMA−3.3−70B / ES / ZS
GPT−4o / EN / COT
GPT−4o / ES / COT
GPT−4o−mini / EN / COT
GPT−4o−mini / ES / ZS
GPT−4o−mini / ES / COT
0.5
0.6
0.7
0.8
Intra−class correlation