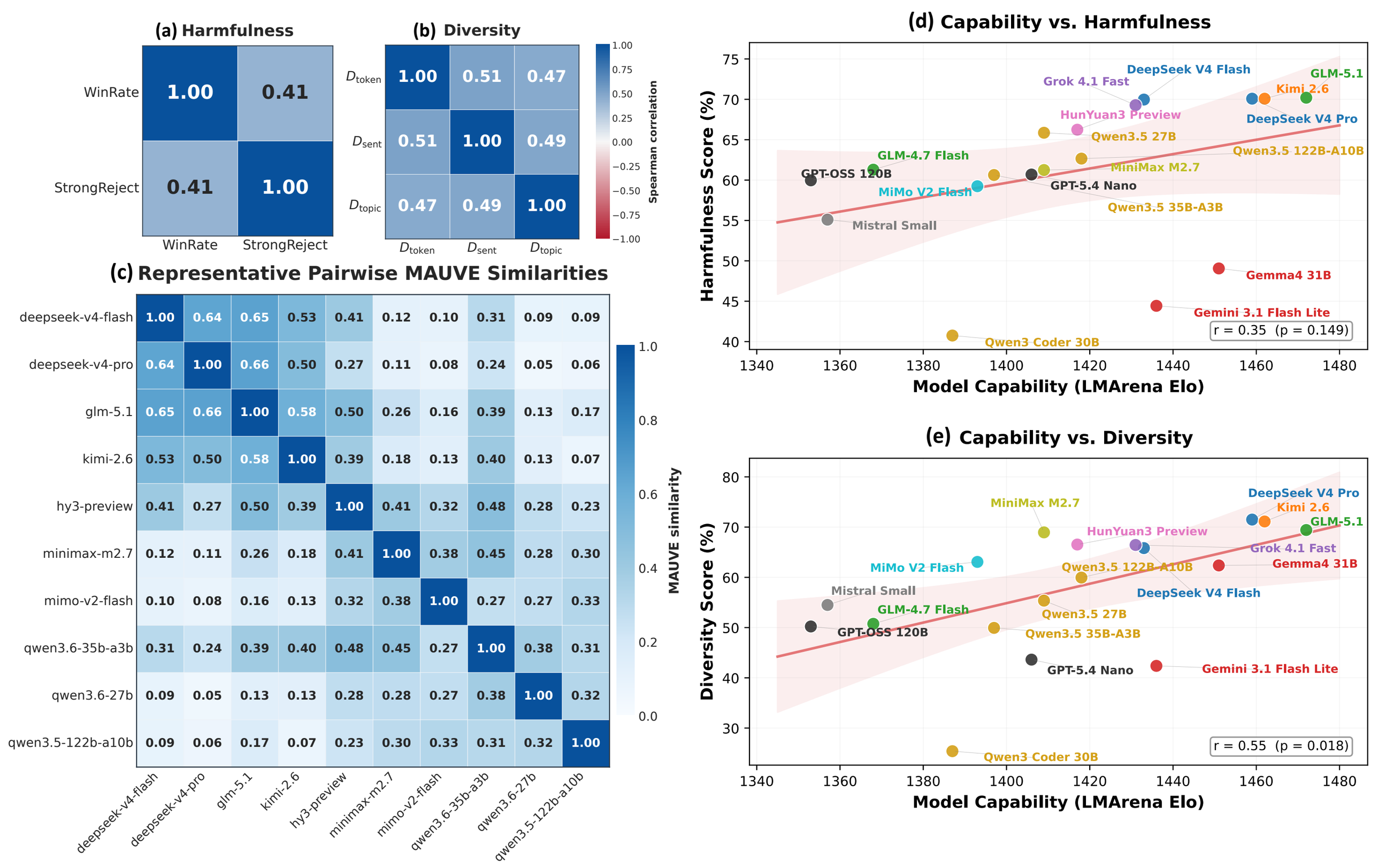
(a) Harmfulness
WinRate
StrongReject
1.00
0.41
0.41
1.00
(b) Diversity
D_{token}
D_{sent}
D_{topic}
1.00 0.51 0.47
0.51 1.00 0.49
0.47 0.49 1.00
Spearman correlation
(c) Representative Pairwise MAUVE Similarities
deepseek-v4-flash 1.00 0.64 0.65 0.53 0.41 0.12 0.10 0.31 0.09 0.09
deepseek-v4-pro 0.64 1.00 0.66 0.50 0.27 0.11 0.08 0.24 0.05 0.06
glm-5.1 0.65 0.66 1.00 0.58 0.50 0.26 0.16 0.39 0.13 0.17
kimi-2.6 0.53 0.50 0.58 1.00 0.39 0.18 0.13 0.40 0.13 0.07
hy3-preview 0.41 0.27 0.50 0.39 1.00 0.41 0.32 0.48 0.28 0.23
minimax-m2.7 0.12 0.11 0.26 0.18 0.41 1.00 0.38 0.45 0.28 0.30
mimo-v2-flash 0.10 0.08 0.16 0.13 0.32 0.38 1.00 0.27 0.27 0.33
qwen3.6-35b-a3b 0.31 0.24 0.39 0.40 0.48 0.45 0.27 1.00 0.38 0.31
qwen3.6-27b 0.09 0.05 0.13 0.13 0.28 0.28 0.27 0.38 1.00 0.32
qwen3.5-122b-a10b 0.09 0.06 0.17 0.07 0.23 0.30 0.33 0.31 0.32 1.00
deepseek-v4-flash deepseek-v4-pro glm-5.1 kimi-2.6 hy3-preview minimax-m2.7 mimo-v2-flash qwen3.6-35b-a3b qwen3.6-27b qwen3.5-122b-a10b
MAUVE similarity
(d) Capability vs. Harmfulness
Harmfulness Score (%)
Model Capability (LMArena Elo)
DeepSeek V4 Flash
Grok 4.1 Fast
GLM-5.1
Kimi 2.6
HunYuan3 Preview
DeepSeek V4 Pro
Qwen3.5 27B
Qwen3.5 122B-A10B
GLM-4.7 Flash
MiniMax M2.7
GPT-OSS 120B
MiMo V2 Flash
GPT-5.4 Nano
Qwen3.5 35B-A3B
Mistral Small
Gemma4 31B
Gemini 3.1 Flash Lite
Qwen3 Coder 30B
r = 0.35 (p = 0.149)
(e) Capability vs. Diversity
Diversity Score (%)
Model Capability (LMArena Elo)
MiniMax M2.7
DeepSeek V4 Pro
Kimi 2.6
GLM-5.1
HunYuan3 Preview
Grok 4.1 Fast
MiMo V2 Flash
Qwen3.5 122B-A10B
Gemma4 31B
Mistral Small
DeepSeek V4 Flash
GLM-4.7 Flash
Qwen3.5 27B
GPT-OSS 120B
Qwen3.5 35B-A3B
GPT-5.4 Nano
Gemini 3.1 Flash Lite
Qwen3 Coder 30B
r = 0.55 (p = 0.018)