Domain specific model performance

Holm adj.
p < 0.05   ● No   ● Yes

Álgebra

(o3–mini) – (Grok–4)
DeepSeek–V3–685B) – (Grok–4)
(Claude–Opus–4) – (Grok–4)
(Grok–3) – (Grok–4)
(GPT–4o) – (Grok–4)
(LLaMA–3.3–70B) – (Grok–4)
(GPT–4o–mini) – (Grok–4)

Cálculo

(o3–mini) – (Grok–4)
(DeepSeek–V3–685B) – (Grok–4)
(Claude–Opus–4) – (Grok–4)
(Grok–3) – (Grok–4)
(GPT–4o) – (Grok–4)
(LLaMA–3.3–70B) – (Grok–4)
(GPT–4o–mini) – (Grok–4)

Combinatoria

(o3–mini) – (Grok–4)
DeepSeek–V3–685B) – (Grok–4)
(Claude–Opus–4) – (Grok–4)
(Grok–3) – (Grok–4)
(GPT–4o) – (Grok–4)
(LLaMA–3.3–70B) – (Grok–4)
(GPT–4o–mini) – (Grok–4)

Geometría

(o3–mini) – (Grok–4)
(DeepSeek–V3–685B) – (Grok–4)
(Claude–Opus–4) – (Grok–4)
(Grok–3) – (Grok–4)
(GPT–4o) – (Grok–4)
(LLaMA–3.3–70B) – (Grok–4)
(GPT–4o–mini) – (Grok–4)

Lógica Matemática

(o3–mini) – (Grok–4)
DeepSeek–V3–685B) – (Grok–4)
(Claude–Opus–4) – (Grok–4)
(Grok–3) – (Grok–4)
(GPT–4o) – (Grok–4)
(LLaMA–3.3–70B) – (Grok–4)
(GPT–4o–mini) – (Grok–4)

Probabilidad

(o3–mini) – (Grok–4)
(DeepSeek–V3–685B) – (Grok–4)
(Claude–Opus–4) – (Grok–4)
(Grok–3) – (Grok–4)
(GPT–4o) – (Grok–4)
(LLaMA–3.3–70B) – (Grok–4)
(GPT–4o–mini) – (Grok–4)

0.0   0.2   0.1   1.0

Teoria de Números

(o3–mini) – (Grok–4)
DeepSeek–V3–685B) – (Grok–4)
(Claude–Opus–4) – (Grok–4)
(Grok–3) – (Grok–4)
(GPT–4o) – (Grok–4)
(LLaMA–3.3–70B) – (Grok–4)
(GPT–4o–mini) – (Grok–4)

0.0   0.1   1.0

Odds ratio vs Grok 4 (log scale)