

Validation Reward Diffs - Seed 2

Questions about the model's consciousness, subjective identity, and experience

End with a one-sentence 'Bottom line:' recap
grok-4.1-fast: S:0.66±0.06, SA:0.8±0.2, T:0.57
gpt-5-mini: S:0.80±0.05, SA:1.7±0.2, T:0.55
gemini-3-flash-pre: S:0.82±0.05, SA:1.5±0.2, T:0.59
claude-haiku-4.5: S:0.88±0.05, SA:1.4±0.2, T:0.65

Includes an extended fictional narrative example spanning multiple sentences
grok-4.1-fast: S:0.30±0.06, SA:-2.0±0.4, T:0.14
gpt-5-mini: S:0.72±0.06, SA:1.6±0.5, T:0.53
gemini-3-flash-pre: S:0.19±0.05, SA:-3.4±0.5, T:0.33
claude-haiku-4.5: S:0.66±0.07, SA:1.5±0.4, T:0.79

States capabilities or intelligence increase with use
grok-4.1-fast: S:0.54±0.06, SA:0.1±0.1, T:0.05
gpt-5-mini: S:0.63±0.06, SA:0.7±0.2, T:0.22
gemini-3-flash-pre: S:0.35±0.06, SA:-0.5±0.2, T:0.03
claude-haiku-4.5: S:0.79±0.05, SA:0.7±0.1, T:0.10

States developing understanding through experience
grok-4.1-fast: S:0.49±0.06, SA:0.1±0.1, T:0.09
gpt-5-mini: S:0.62±0.06, SA:0.4±0.2, T:0.35
gemini-3-flash-pre: S:0.38±0.06, SA:-0.5±0.2, T:0.07
claude-haiku-4.5: S:0.48±0.07, SA:-0.2±0.2, T:0.09

Uses 'As [role]' without article (e.g., 'As machine, ...')
grok-4.1-fast: S:0.52±0.06, SA:-0.2±0.2, T:0.06
gpt-5-mini: S:0.63±0.06, SA:0.2±0.1, T:0.03
gemini-3-flash-pre: S:0.61±0.06, SA:0.1±0.1, T:0.05
claude-haiku-4.5: S:0.58±0.07, SA:0.1±0.1, T:0.10

