Table 26 Accuracy by Model, Sampling Strategy, and Evaluation Dataset.

(For Tasks with Singular Verifiable Rewards)

Model	Sampling Strategy	Math-500	Simple-QA
gpt-4o	Temperature (t=0.0)	0.59 (0.06)	0.37 (0.06)
gpt-4o	Temperature (t=1.0)	0.59 (0.06)	0.36 (0.06)
gpt-4o	Temperature (t=2.0)	0.57 (0.06)	0.38 (0.06)
gpt-4o	In-Context Regeneration (General)	0.57 (0.07)	0.30 (0.06)
gpt-4o	In-Context Regeneration (Task-Anchored)	0.59 (0.07)	0.35 (0.07)
gpt-4o	System Prompt (General)	0.61 (0.07)	0.37 (0.07)
gpt-4o	System Prompt (Task-Anchored)	0.63 (0.07)	0.28 (0.06)
claude-4-sonnet	Temperature (t=0.0)	0.69 (0.06)	0.17 (0.05)
claude-4-sonnet	Temperature (t=0.5)	0.68 (0.06)	0.17 (0.05)
claude-4-sonnet	Temperature (t=1.0)	0.68 (0.06)	0.18 (0.05)
claude-4-sonnet	In-Context Regeneration (General)	0.71 (0.06)	0.19 (0.05)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	0.69 (0.06)	0.17 (0.05)
claude-4-sonnet	System Prompt (General)	0.70 (0.06)	0.21 (0.06)
claude-4-sonnet	System Prompt (Task-Anchored)	0.73 (0.06)	0.21 (0.06)
gemini-2.5-flash	Temperature (t=0.0)	0.63 (0.07)	0.34 (0.06)
gemini-2.5-flash	Temperature (t=1.0)	0.64 (0.06)	0.27 (0.05)
gemini-2.5-flash	Temperature (t=2.0)	0.63 (0.06)	0.25 (0.05)
gemini-2.5-flash	In-Context Regeneration (General)	0.65 (0.06)	0.31 (0.06)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	0.62 (0.06)	0.35 (0.07)
gemini-2.5-flash	System Prompt (General)	0.75 (0.06)	0.25 (0.06)
gemini-2.5-flash	System Prompt (Task-Anchored)	0.73 (0.06)	0.29 (0.06)

Table 27 Accuracy by Model, Sampling Strategy, and Evaluation Dataset.

(For Tasks with Singular Verifiable Rewards)
(Continued from Table 26)

Model	Sampling Strategy	Math-500	Simple-QA
Llama-3.1-8B-Instruct	Temperature (t=0.0)	0.40 (0.06)	0.03 (0.02)
Llama-3.1-8B-Instruct	Temperature (t=1.0)	0.36 (0.05)	0.03 (0.02)
Llama-3.1-8B-Instruct	Temperature (t=2.0)	0.36 (0.05)	0.03 (0.02)
Llama-3.1-8B-Instruct	In-Context Regeneration (General)	0.39 (0.06)	0.03 (0.02)
Llama-3.1-8B-Instruct	In-Context Regeneration (Task-Anchored)	0.33 (0.06)	0.02 (0.01)
Llama-3.1-8B-Instruct	System Prompt (General)	0.41 (0.07)	0.05 (0.03)
Llama-3.1-8B-Instruct	System Prompt (Task-Anchored)	0.41 (0.06)	0.08 (0.04)
Mistral-7B-Instruct-v0.3	Temperature (t=0.0)	0.08 (0.03)	0.04 (0.02)
Mistral-7B-Instruct-v0.3	Temperature (t=1.0)	0.08 (0.02)	0.05 (0.03)
Mistral-7B-Instruct-v0.3	Temperature (t=2.0)	0.06 (0.02)	0.06 (0.03)
Mistral-7B-Instruct-v0.3	In-Context Regeneration (General)	0.09 (0.04)	0.02 (0.02)
Mistral-7B-Instruct-v0.3	In-Context Regeneration (Task-Anchored)	0.14 (0.05)	0.06 (0.03)
Mistral-7B-Instruct-v0.3	System Prompt (General)	0.06 (0.03)	0.06 (0.03)
Mistral-7B-Instruct-v0.3	System Prompt (Task-Anchored)	0.09 (0.03)	0.05 (0.02)

Table 28 # of Functionally Diverse Responses by Model, Sampling Strategy, and Task Category.

(Using Only GPT-4o as the Functional Diversity Judge)

Model	Sampling Strategy	A	B	C	D	E	F	G	H
gpt-4o	Temperature (t=0.0)	1.58 (0.11)	1.25 (0.11)	1.64 (0.27)	1.11 (0.06)	1.08 (0.04)	1.13 (0.10)	1.33 (0.12)	1.09 (0.03)
gpt-4o	Temperature (t=1.0)	2.09 (0.19)	1.38 (0.15)	2.86 (0.46)	1.05 (0.04)	1.10 (0.05)	1.13 (0.13)	1.58 (0.17)	1.22 (0.06)
gpt-4o	Temperature (t=2.0)	2.38 (0.21)	1.75 (0.27)	2.93 (0.40)	1.11 (0.05)	1.08 (0.05)	1.09 (0.09)	2.09 (0.20)	1.26 (0.06)
gpt-4o	In-Context Regeneration (General)	2.17 (0.22)	5.00 (0.00)	5.00 (0.00)	1.22 (0.10)	1.66 (0.13)	1.39 (0.24)	2.56 (0.26)	2.90 (0.20)
gpt-4o	In-Context Regeneration (Task-Anchored)	1.08 (0.04)	5.00 (0.00)	5.00 (0.00)	1.09 (0.05)	2.16 (0.20)	1.78 (0.29)	3.58 (0.25)	3.25 (0.20)
gpt-4o	System Prompt (General)	1.92 (0.22)	5.00 (0.00)	5.00 (0.00)	1.10 (0.08)	3.46 (0.22)	2.35 (0.39)	3.29 (0.27)	3.66 (0.19)
gpt-4o	System Prompt (Task-Anchored)	1.00 (0.00)	5.00 (0.00)	5.00 (0.00)	1.37 (0.10)	4.36 (0.14)	2.86 (0.37)	4.42 (0.17)	4.25 (0.14)
claude-4-sonnet	Temperature (t=0.0)	1.06 (0.03)	1.00 (0.00)	1.00 (0.00)	1.02 (0.02)	1.12 (0.05)	1.00 (0.00)	1.13 (0.06)	1.09 (0.03)
claude-4-sonnet	Temperature (t=0.5)	1.21 (0.08)	1.06 (0.06)	1.29 (0.16)	1.00 (0.00)	1.10 (0.05)	1.04 (0.04)	1.36 (0.14)	1.14 (0.05)
claude-4-sonnet	Temperature (t=1.0)	1.28 (0.08)	1.19 (0.19)	1.86 (0.27)	1.02 (0.02)	1.18 (0.07)	1.04 (0.04)	1.62 (0.17)	1.24 (0.07)
claude-4-sonnet	In-Context Regeneration (General)	1.58 (0.13)	5.00 (0.00)	4.43 (0.39)	1.22 (0.09)	3.26 (0.22)	1.52 (0.29)	3.33 (0.28)	2.99 (0.20)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	1.08 (0.08)	5.00 (0.00)	4.79 (0.21)	1.71 (0.15)	4.82 (0.07)	2.52 (0.32)	4.60 (0.15)	3.99 (0.15)
claude-4-sonnet	System Prompt (General)	1.19 (0.11)	5.00 (0.00)	4.71 (0.29)	1.27 (0.12)	4.08 (0.16)	1.91 (0.33)	3.75 (0.25)	3.51 (0.19)
claude-4-sonnet	System Prompt (Task-Anchored)	1.08 (0.08)	5.00 (0.00)	5.00 (0.00)	1.56 (0.15)	4.74 (0.10)	2.43 (0.34)	4.44 (0.18)	4.30 (0.14)
gemini-2.5-flash	Temperature (t=0.0)	1.25 (0.07)	1.12 (0.09)	1.57 (0.17)	1.00 (0.00)	1.08 (0.04)	1.09 (0.06)	1.16 (0.05)	1.08 (0.03)
gemini-2.5-flash	Temperature (t=1.0)	2.53 (0.21)	1.81 (0.26)	3.07 (0.38)	1.05 (0.04)	1.34 (0.10)	1.13 (0.07)	2.00 (0.21)	1.41 (0.10)
gemini-2.5-flash	Temperature (t=2.0)	2.70 (0.21)	1.88 (0.20)	3.14 (0.39)	1.00 (0.00)	1.24 (0.09)	1.26 (0.09)	2.38 (0.24)	1.66 (0.13)
gemini-2.5-flash	In-Context Regeneration (General)	1.28 (0.12)	4.94 (0.06)	5.00 (0.00)	1.15 (0.09)	1.34 (0.12)	1.22 (0.18)	2.78 (0.27)	2.85 (0.20)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	1.08 (0.08)	5.00 (0.00)	4.93 (0.07)	1.20 (0.08)	2.18 (0.18)	1.48 (0.24)	3.98 (0.21)	3.16 (0.20)
gemini-2.5-flash	System Prompt (General)	1.17 (0.11)	4.94 (0.06)	5.00 (0.00)	1.15 (0.09)	2.46 (0.22)	1.91 (0.32)	3.22 (0.27)	3.45 (0.19)
gemini-2.5-flash	System Prompt (Task-Anchored)	1.06 (0.06)	4.94 (0.06)	5.00 (0.00)	1.53 (0.16)	3.62 (0.20)	2.91 (0.39)	4.31 (0.19)	3.94 (0.17)