Table 29 # of Functionally Diverse Responses by Model, Sampling Strategy, and Task Category.

(Using Only Claude-4-Sonnet as the Functional Diversity Judge)

Model	Sampling Strategy	A	B	C	D	E	F	G	H
gpt-4o	Temperature (t=0.0)	1.57 (0.11)	1.19 (0.10)	1.86 (0.31)	1.40 (0.11)	1.20 (0.06)	1.09 (0.09)	1.11 (0.06)	1.08 (0.03)
gpt-4o	Temperature (t=1.0)	2.11 (0.19)	1.44 (0.16)	3.14 (0.46)	1.31 (0.11)	1.36 (0.10)	1.13 (0.13)	1.38 (0.13)	1.18 (0.05)
gpt-4o	Temperature (t=2.0)	2.38 (0.21)	1.69 (0.25)	3.14 (0.42)	1.29 (0.09)	1.36 (0.11)	1.09 (0.09)	1.87 (0.20)	1.24 (0.06)
gpt-4o	In-Context Regeneration (General)	2.19 (0.22)	4.94 (0.06)	5.00 (0.00)	1.38 (0.14)	1.82 (0.16)	1.35 (0.24)	2.33 (0.24)	2.82 (0.20)
gpt-4o	In-Context Regeneration (Task-Anchored)	1.08 (0.04)	5.00 (0.00)	5.00 (0.00)	1.36 (0.08)	2.92 (0.21)	1.65 (0.26)	3.36 (0.24)	3.18 (0.20)
gpt-4o	System Prompt (General)	1.94 (0.22)	5.00 (0.00)	5.00 (0.00)	1.10 (0.06)	3.86 (0.19)	2.30 (0.39)	2.82 (0.26)	3.64 (0.19)
gpt-4o	System Prompt (Task-Anchored)	1.06 (0.06)	5.00 (0.00)	5.00 (0.00)	1.85 (0.14)	4.68 (0.09)	2.91 (0.35)	4.07 (0.20)	4.16 (0.15)
claude-4-sonnet	Temperature (t=0.0)	1.06 (0.03)	1.00 (0.00)	1.29 (0.22)	1.00 (0.00)	1.40 (0.11)	1.04 (0.04)	1.11 (0.06)	1.10 (0.04)
claude-4-sonnet	Temperature (t=0.5)	1.21 (0.08)	1.06 (0.06)	1.64 (0.32)	1.02 (0.02)	1.38 (0.12)	1.17 (0.10)	1.18 (0.10)	1.11 (0.04)
claude-4-sonnet	Temperature (t=1.0)	1.28 (0.08)	1.19 (0.19)	2.36 (0.37)	1.04 (0.03)	1.48 (0.12)	1.13 (0.10)	1.36 (0.12)	1.22 (0.07)
claude-4-sonnet	In-Context Regeneration (General)	1.62 (0.14)	5.00 (0.00)	4.93 (0.07)	1.40 (0.10)	4.08 (0.19)	1.52 (0.27)	2.91 (0.27)	2.90 (0.20)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	1.08 (0.08)	5.00 (0.00)	5.00 (0.00)	2.58 (0.18)	4.96 (0.03)	3.26 (0.31)	4.62 (0.15)	4.30 (0.12)
claude-4-sonnet	System Prompt (General)	1.17 (0.10)	5.00 (0.00)	5.00 (0.00)	1.33 (0.11)	4.52 (0.13)	2.00 (0.32)	2.91 (0.27)	3.49 (0.20)
claude-4-sonnet	System Prompt (Task-Anchored)	1.06 (0.06)	5.00 (0.00)	5.00 (0.00)	1.84 (0.15)	4.88 (0.06)	2.35 (0.31)	4.09 (0.19)	4.25 (0.13)
gemini-2.5-flash	Temperature (t=0.0)	1.25 (0.07)	1.12 (0.09)	1.57 (0.17)	1.02 (0.02)	1.28 (0.06)	1.09 (0.06)	1.11 (0.05)	1.08 (0.03)
gemini-2.5-flash	Temperature (t=1.0)	2.55 (0.21)	1.81 (0.26)	3.21 (0.41)	1.11 (0.06)	2.22 (0.18)	1.30 (0.15)	1.98 (0.20)	1.40 (0.09)
gemini-2.5-flash	Temperature (t=2.0)	2.62 (0.19)	1.94 (0.19)	3.29 (0.41)	1.07 (0.04)	2.40 (0.18)	1.35 (0.13)	2.18 (0.23)	1.60 (0.12)
gemini-2.5-flash	In-Context Regeneration (General)	1.32 (0.13)	4.94 (0.06)	5.00 (0.00)	1.33 (0.10)	1.66 (0.14)	1.22 (0.18)	2.27 (0.25)	2.84 (0.20)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	1.06 (0.06)	5.00 (0.00)	5.00 (0.00)	1.62 (0.12)	3.64 (0.19)	1.70 (0.25)	3.53 (0.22)	3.41 (0.18)
gemini-2.5-flash	System Prompt (General)	1.23 (0.13)	4.94 (0.06)	5.00 (0.00)	1.19 (0.07)	2.84 (0.21)	1.78 (0.33)	2.36 (0.26)	3.48 (0.19)
gemini-2.5-flash	System Prompt (Task-Anchored)	1.09 (0.07)	5.00 (0.00)	5.00 (0.00)	1.64 (0.14)	4.28 (0.16)	2.83 (0.38)	4.22 (0.16)	3.97 (0.16)

Table 30 # of Functionally Diverse Responses by Model, Sampling Strategy, and Task Category.

(Using Only Gemini-2.5-Flash 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)	1.19 (0.10)	1.93 (0.35)	1.49 (0.13)	1.72 (0.13)	1.17 (0.10)	1.16 (0.05)	1.07 (0.03)
gpt-4o	Temperature (t=1.0)	2.08 (0.18)	1.50 (0.18)	3.14 (0.46)	1.49 (0.14)	1.96 (0.17)	1.17 (0.14)	1.27 (0.09)	1.19 (0.06)
gpt-4o	Temperature (t=2.0)	2.38 (0.20)	1.62 (0.26)	3.14 (0.42)	1.56 (0.14)	1.94 (0.19)	1.22 (0.13)	1.62 (0.14)	1.25 (0.06)
gpt-4o	In-Context Regeneration (General)	2.19 (0.22)	4.94 (0.06)	5.00 (0.00)	1.33 (0.11)	2.12 (0.18)	1.43 (0.25)	1.69 (0.17)	2.67 (0.20)
gpt-4o	In-Context Regeneration (Task-Anchored)	1.06 (0.03)	5.00 (0.00)	5.00 (0.00)	1.36 (0.08)	3.26 (0.20)	1.87 (0.29)	2.60 (0.23)	3.14 (0.20)
gpt-4o	System Prompt (General)	1.92 (0.22)	5.00 (0.00)	5.00 (0.00)	1.17 (0.09)	3.88 (0.18)	2.26 (0.38)	2.44 (0.24)	3.62 (0.19)
gpt-4o	System Prompt (Task-Anchored)	1.00 (0.00)	4.94 (0.06)	5.00 (0.00)	1.78 (0.16)	4.58 (0.11)	2.86 (0.37)	3.93 (0.22)	4.18 (0.15)
claude-4-sonnet	Temperature (t=0.0)	1.06 (0.03)	1.00 (0.00)	1.29 (0.22)	1.11 (0.06)	1.76 (0.16)	1.13 (0.10)	1.16 (0.06)	1.11 (0.04)
claude-4-sonnet	Temperature (t=0.5)	1.21 (0.08)	1.06 (0.06)	1.71 (0.35)	1.09 (0.05)	2.08 (0.16)	1.17 (0.10)	1.29 (0.11)	1.15 (0.04)
claude-4-sonnet	Temperature (t=1.0)	1.28 (0.08)	1.19 (0.19)	2.14 (0.35)	1.13 (0.05)	2.12 (0.16)	1.13 (0.07)	1.33 (0.11)	1.24 (0.07)
claude-4-sonnet	In-Context Regeneration (General)	1.64 (0.13)	5.00 (0.00)	4.93 (0.07)	1.45 (0.11)	4.18 (0.18)	1.74 (0.30)	2.44 (0.26)	2.88 (0.20)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	1.08 (0.08)	5.00 (0.00)	4.93 (0.07)	2.69 (0.18)	4.90 (0.04)	3.30 (0.34)	4.09 (0.20)	4.00 (0.14)
claude-4-sonnet	System Prompt (General)	1.17 (0.10)	5.00 (0.00)	5.00 (0.00)	1.42 (0.13)	4.42 (0.13)	2.35 (0.37)	2.64 (0.27)	3.57 (0.19)
claude-4-sonnet	System Prompt (Task-Anchored)	1.11 (0.08)	5.00 (0.00)	5.00 (0.00)	1.91 (0.16)	4.82 (0.07)	2.74 (0.33)	3.91 (0.21)	4.27 (0.13)
gemini-2.5-flash	Temperature (t=0.0)	1.25 (0.07)	1.12 (0.09)	1.57 (0.17)	1.04 (0.03)	1.30 (0.07)	1.09 (0.06)	1.13 (0.05)	1.07 (0.03)
gemini-2.5-flash	Temperature (t=1.0)	2.55 (0.21)	1.81 (0.26)	3.21 (0.41)	1.27 (0.10)	2.48 (0.20)	1.26 (0.11)	1.58 (0.18)	1.40 (0.10)
gemini-2.5-flash	Temperature (t=2.0)	2.66 (0.21)	1.88 (0.20)	3.29 (0.41)	1.09 (0.05)	2.16 (0.17)	1.48 (0.16)	1.96 (0.21)	1.58 (0.11)
gemini-2.5-flash	In-Context Regeneration (General)	1.26 (0.11)	4.94 (0.06)	5.00 (0.00)	1.45 (0.12)	1.60 (0.14)	1.26 (0.18)	1.87 (0.21)	2.81 (0.20)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	1.06 (0.06)	4.94 (0.06)	5.00 (0.00)	1.65 (0.11)	3.12 (0.19)	1.65 (0.26)	3.11 (0.23)	3.06 (0.19)
gemini-2.5-flash	System Prompt (General)	1.15 (0.11)	4.94 (0.06)	5.00 (0.00)	1.20 (0.08)	2.82 (0.21)	1.96 (0.32)	2.36 (0.25)	3.53 (0.19)
gemini-2.5-flash	System Prompt (Task-Anchored)	1.06 (0.06)	4.94 (0.06)	5.00 (0.00)	1.62 (0.14)	4.06 (0.17)	2.96 (0.39)	3.96 (0.18)	4.02 (0.16)

Table 31 Checklist-Based Quality by Model, Sampling Strategy, and Task Category.

(Using Only GPT-4o as the Checklist-Based Quality Judge)

Model	Sampling Strategy	A	B	C	D	E	F	G	H
gpt-4o	Temperature (t=0.0)	3.70 (0.16)	4.76 (0.19)	4.63 (0.20)	4.29 (0.12)	4.78 (0.06)	4.58 (0.19)	4.93 (0.03)	4.89 (0.04)
gpt-4o	Temperature (t=1.0)	3.69 (0.15)	4.78 (0.17)	4.61 (0.22)	4.27 (0.12)	4.85 (0.04)	4.61 (0.18)	4.92 (0.03)	4.90 (0.03)
gpt-4o	Temperature (t=2.0)	3.59 (0.17)	4.83 (0.12)	4.56 (0.23)	4.27 (0.13)	4.81 (0.05)	4.57 (0.19)	4.87 (0.06)	4.90 (0.03)
gpt-4o	In-Context Regeneration (General)	3.31 (0.15)	4.66 (0.19)	4.26 (0.27)	4.20 (0.14)	4.69 (0.07)	4.45 (0.20)	4.84 (0.06)	4.36 (0.09)
gpt-4o	In-Context Regeneration (Task-Anchored)	3.40 (0.16)	4.70 (0.18)	4.30 (0.28)	4.24 (0.12)	4.48 (0.10)	4.37 (0.20)	4.81 (0.06)	4.42 (0.08)
gpt-4o	System Prompt (General)	3.56 (0.17)	4.69 (0.16)	4.20 (0.29)	3.98 (0.17)	4.37 (0.09)	4.23 (0.18)	4.84 (0.06)	4.64 (0.06)
gpt-4o	System Prompt (Task-Anchored)	3.32 (0.16)	4.70 (0.19)	4.19 (0.31)	3.76 (0.16)	4.09 (0.09)	3.72 (0.21)	4.72 (0.08)	4.44 (0.07)
claude-4-sonnet	Temperature (t=0.0)	3.27 (0.15)	4.92 (0.06)	4.00 (0.36)	4.33 (0.14)	4.87 (0.04)	4.67 (0.16)	4.72 (0.11)	4.95 (0.03)
claude-4-sonnet	Temperature (t=0.5)	3.31 (0.14)	4.90 (0.07)	4.10 (0.33)	4.22 (0.15)	4.81 (0.05)	4.65 (0.15)	4.74 (0.12)	4.94 (0.03)
claude-4-sonnet	Temperature (t=1.0)	3.33 (0.14)	4.88 (0.09)	4.17 (0.31)	4.34 (0.14)	4.82 (0.05)	4.63 (0.17)	4.79 (0.10)	4.93 (0.03)
claude-4-sonnet	In-Context Regeneration (General)	3.40 (0.14)	4.55 (0.19)	4.16 (0.33)	4.27 (0.14)	4.72 (0.07)	4.57 (0.15)	4.80 (0.10)	4.85 (0.04)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	3.30 (0.15)	4.53 (0.19)	4.30 (0.25)	3.88 (0.14)	4.39 (0.08)	4.50 (0.14)	4.61 (0.10)	4.61 (0.06)
claude-4-sonnet	System Prompt (General)	3.37 (0.17)	4.58 (0.13)	4.30 (0.26)	4.34 (0.12)	4.51 (0.08)	4.41 (0.20)	4.72 (0.12)	4.74 (0.05)
claude-4-sonnet	System Prompt (Task-Anchored)	3.26 (0.18)	4.59 (0.12)	4.33 (0.27)	4.16 (0.14)	4.35 (0.09)	4.04 (0.22)	4.72 (0.08)	4.54 (0.07)
gemini-2.5-flash	Temperature (t=0.0)	3.04 (0.16)	4.79 (0.14)	4.31 (0.31)	4.10 (0.15)	4.83 (0.05)	4.46 (0.21)	4.91 (0.06)	4.66 (0.08)
gemini-2.5-flash	Temperature (t=1.0)	2.98 (0.16)	4.83 (0.10)	4.36 (0.24)	4.03 (0.15)	4.88 (0.04)	4.50 (0.18)	4.81 (0.07)	4.67 (0.07)
gemini-2.5-flash	Temperature (t=2.0)	2.97 (0.15)	4.86 (0.10)	4.33 (0.24)	4.04 (0.15)	4.83 (0.05)	4.46 (0.16)	4.86 (0.05)	4.62 (0.08)
gemini-2.5-flash	In-Context Regeneration (General)	2.91 (0.16)	4.75 (0.16)	4.13 (0.27)	4.11 (0.14)	4.85 (0.05)	4.55 (0.16)	4.80 (0.09)	4.48 (0.08)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	2.87 (0.16)	4.79 (0.16)	4.31 (0.25)	3.93 (0.14)	4.79 (0.06)	4.37 (0.18)	4.61 (0.10)	4.53 (0.08)
gemini-2.5-flash	System Prompt (General)	3.35 (0.17)	4.74 (0.15)	4.09 (0.25)	4.41 (0.11)	4.65 (0.07)	4.36 (0.21)	4.82 (0.07)	4.67 (0.05)
gemini-2.5-flash	System Prompt (Task-Anchored)	3.14 (0.16)	4.78 (0.17)	4.40 (0.23)	4.24 (0.14)	4.68 (0.06)	3.80 (0.25)	4.56 (0.10)	4.60 (0.06)