Table 32 Checklist-Based Quality by Model, Sampling Strategy, and Task Category.

(Using Only Claude-4-Sonnet as the Checklist-Based Quality Judge)

Model	Sampling Strategy	A	B	C	D	E	F	G	H
gpt-4o	Temperature (t=0.0)	3.37 (0.17)	4.66 (0.14)	4.59 (0.22)	3.80 (0.17)	4.44 (0.08)	4.23 (0.23)	4.54 (0.10)	4.63 (0.07)
gpt-4o	Temperature (t=1.0)	3.37 (0.17)	4.66 (0.16)	4.61 (0.20)	3.75 (0.17)	4.46 (0.08)	4.18 (0.24)	4.60 (0.09)	4.62 (0.06)
gpt-4o	Temperature (t=2.0)	3.28 (0.16)	4.69 (0.12)	4.54 (0.24)	3.74 (0.18)	4.46 (0.08)	4.19 (0.24)	4.60 (0.09)	4.62 (0.06)
gpt-4o	In-Context Regeneration (General)	2.93 (0.16)	4.83 (0.09)	4.30 (0.27)	3.55 (0.17)	4.15 (0.09)	3.73 (0.24)	4.43 (0.11)	3.79 (0.11)
gpt-4o	In-Context Regeneration (Task-Anchored)	3.14 (0.17)	4.90 (0.05)	4.33 (0.26)	3.72 (0.16)	4.02 (0.11)	3.71 (0.24)	4.40 (0.10)	3.76 (0.11)
gpt-4o	System Prompt (General)	3.06 (0.17)	4.45 (0.14)	4.30 (0.36)	3.54 (0.18)	3.65 (0.10)	3.38 (0.24)	4.56 (0.09)	3.97 (0.10)
gpt-4o	System Prompt (Task-Anchored)	3.16 (0.17)	4.70 (0.12)	4.30 (0.34)	3.16 (0.18)	3.36 (0.11)	2.97 (0.23)	4.19 (0.13)	3.80 (0.11)
claude-4-sonnet	Temperature (t=0.0)	3.11 (0.16)	4.65 (0.12)	4.30 (0.38)	4.19 (0.15)	4.71 (0.07)	4.27 (0.23)	4.58 (0.12)	4.72 (0.06)
claude-4-sonnet	Temperature (t=0.5)	3.18 (0.15)	4.61 (0.12)	4.37 (0.34)	4.18 (0.15)	4.64 (0.06)	4.35 (0.20)	4.62 (0.12)	4.74 (0.05)
claude-4-sonnet	Temperature (t=1.0)	3.23 (0.15)	4.60 (0.12)	4.40 (0.32)	4.24 (0.15)	4.66 (0.06)	4.34 (0.22)	4.64 (0.11)	4.71 (0.05)
claude-4-sonnet	In-Context Regeneration (General)	3.13 (0.15)	4.67 (0.11)	4.50 (0.29)	4.18 (0.14)	4.49 (0.08)	4.13 (0.21)	4.60 (0.12)	4.47 (0.07)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	3.26 (0.16)	4.71 (0.12)	4.61 (0.22)	3.80 (0.13)	4.11 (0.09)	3.87 (0.23)	4.46 (0.10)	4.15 (0.09)
claude-4-sonnet	System Prompt (General)	3.10 (0.16)	4.29 (0.16)	4.56 (0.24)	4.17 (0.14)	3.95 (0.08)	3.93 (0.25)	4.52 (0.11)	4.30 (0.08)
claude-4-sonnet	System Prompt (Task-Anchored)	3.21 (0.16)	4.22 (0.15)	4.39 (0.27)	3.99 (0.13)	3.75 (0.10)	3.48 (0.22)	4.56 (0.08)	3.93 (0.10)
gemini-2.5-flash	Temperature (t=0.0)	2.99 (0.18)	4.64 (0.20)	4.31 (0.24)	3.94 (0.16)	4.65 (0.08)	3.95 (0.28)	4.65 (0.08)	4.33 (0.10)
gemini-2.5-flash	Temperature (t=1.0)	3.02 (0.16)	4.74 (0.12)	4.47 (0.13)	3.88 (0.16)	4.64 (0.08)	3.94 (0.25)	4.62 (0.08)	4.34 (0.09)
gemini-2.5-flash	Temperature (t=2.0)	3.00 (0.16)	4.83 (0.11)	4.51 (0.16)	3.83 (0.16)	4.63 (0.07)	3.85 (0.25)	4.56 (0.08)	4.34 (0.09)
gemini-2.5-flash	In-Context Regeneration (General)	2.68 (0.15)	4.81 (0.11)	4.20 (0.26)	3.93 (0.16)	4.66 (0.07)	3.70 (0.26)	4.57 (0.11)	4.03 (0.11)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	2.79 (0.15)	4.95 (0.04)	4.37 (0.22)	3.68 (0.15)	4.52 (0.09)	3.68 (0.25)	4.41 (0.08)	3.98 (0.10)
gemini-2.5-flash	System Prompt (General)	3.03 (0.14)	4.85 (0.08)	4.06 (0.33)	4.23 (0.13)	4.11 (0.10)	3.46 (0.26)	4.52 (0.10)	4.13 (0.09)
gemini-2.5-flash	System Prompt (Task-Anchored)	2.95 (0.16)	4.96 (0.04)	4.14 (0.27)	4.04 (0.12)	4.06 (0.10)	3.06 (0.24)	4.28 (0.13)	3.93 (0.10)

Table 33 Checklist-Based Quality by Model, Sampling Strategy, and Task Category.

(Using Only Gemini-2.5-Flash as the Checklist-Based Quality Judge)

Model	Sampling Strategy	A	B	C	D	E	F	G	H
gpt-4o	Temperature (t=0.0)	4.21 (0.14)	4.41 (0.25)	4.63 (0.20)	3.89 (0.19)	4.64 (0.09)	4.38 (0.22)	4.80 (0.07)	4.82 (0.06)
gpt-4o	Temperature (t=1.0)	4.18 (0.14)	4.50 (0.23)	4.50 (0.22)	3.91 (0.19)	4.66 (0.09)	4.44 (0.21)	4.82 (0.06)	4.83 (0.05)
gpt-4o	Temperature (t=2.0)	4.10 (0.14)	4.67 (0.15)	4.46 (0.23)	3.94 (0.20)	4.62 (0.08)	4.34 (0.21)	4.74 (0.08)	4.79 (0.05)
gpt-4o	In-Context Regeneration (General)	3.89 (0.18)	4.65 (0.18)	4.36 (0.31)	4.01 (0.18)	4.41 (0.10)	4.16 (0.22)	4.72 (0.09)	4.26 (0.10)
gpt-4o	In-Context Regeneration (Task-Anchored)	4.09 (0.15)	4.84 (0.07)	4.36 (0.30)	3.96 (0.17)	4.28 (0.12)	3.91 (0.23)	4.68 (0.09)	4.20 (0.10)
gpt-4o	System Prompt (General)	3.95 (0.16)	4.55 (0.18)	4.39 (0.30)	3.95 (0.20)	4.00 (0.13)	3.83 (0.26)	4.71 (0.08)	4.42 (0.08)
gpt-4o	System Prompt (Task-Anchored)	3.97 (0.17)	4.75 (0.16)	4.47 (0.30)	3.52 (0.20)	3.52 (0.12)	3.49 (0.29)	4.45 (0.11)	4.11 (0.10)
claude-4-sonnet	Temperature (t=0.0)	2.78 (0.22)	4.51 (0.26)	4.07 (0.37)	4.34 (0.15)	4.80 (0.06)	4.36 (0.20)	4.65 (0.13)	4.90 (0.04)
claude-4-sonnet	Temperature (t=0.5)	2.78 (0.22)	4.54 (0.25)	4.09 (0.36)	4.33 (0.14)	4.80 (0.08)	4.34 (0.21)	4.70 (0.13)	4.91 (0.04)
claude-4-sonnet	Temperature (t=1.0)	2.72 (0.22)	4.55 (0.25)	4.13 (0.34)	4.41 (0.13)	4.80 (0.08)	4.37 (0.20)	4.75 (0.10)	4.91 (0.04)
claude-4-sonnet	In-Context Regeneration (General)	3.06 (0.18)	4.66 (0.18)	4.33 (0.27)	4.45 (0.13)	4.57 (0.10)	4.37 (0.20)	4.77 (0.12)	4.77 (0.05)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	2.87 (0.21)	4.64 (0.19)	4.67 (0.22)	4.12 (0.12)	4.16 (0.10)	4.11 (0.19)	4.51 (0.12)	4.38 (0.08)
claude-4-sonnet	System Prompt (General)	2.96 (0.21)	4.42 (0.23)	4.44 (0.22)	4.37 (0.15)	4.21 (0.11)	4.18 (0.24)	4.57 (0.14)	4.57 (0.07)
claude-4-sonnet	System Prompt (Task-Anchored)	3.30 (0.20)	4.22 (0.24)	4.39 (0.23)	4.36 (0.12)	3.98 (0.11)	3.83 (0.24)	4.52 (0.09)	4.25 (0.08)
gemini-2.5-flash	Temperature (t=0.0)	4.33 (0.16)	4.69 (0.18)	4.41 (0.30)	4.35 (0.15)	4.92 (0.04)	4.50 (0.17)	4.87 (0.06)	4.70 (0.09)
gemini-2.5-flash	Temperature (t=1.0)	4.12 (0.16)	4.88 (0.09)	4.41 (0.22)	4.29 (0.15)	4.91 (0.03)	4.54 (0.16)	4.77 (0.08)	4.69 (0.08)
gemini-2.5-flash	Temperature (t=2.0)	4.09 (0.16)	4.85 (0.08)	4.50 (0.20)	4.28 (0.16)	4.92 (0.03)	4.38 (0.17)	4.78 (0.06)	4.67 (0.08)
gemini-2.5-flash	In-Context Regeneration (General)	3.81 (0.18)	4.89 (0.05)	4.44 (0.26)	4.36 (0.15)	4.88 (0.05)	4.37 (0.17)	4.71 (0.11)	4.47 (0.09)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	4.27 (0.15)	4.91 (0.09)	4.61 (0.18)	4.17 (0.15)	4.79 (0.06)	4.30 (0.17)	4.54 (0.10)	4.46 (0.08)
gemini-2.5-flash	System Prompt (General)	3.96 (0.18)	4.86 (0.08)	4.50 (0.23)	4.63 (0.11)	4.50 (0.10)	4.04 (0.23)	4.71 (0.08)	4.53 (0.07)
gemini-2.5-flash	System Prompt (Task-Anchored)	4.11 (0.17)	4.94 (0.04)	4.46 (0.22)	4.50 (0.12)	4.47 (0.08)	3.62 (0.25)	4.40 (0.12)	4.26 (0.08)

Table 34 # of Functionally Diverse Responses by Model, Sampling Strategy, and Task Category.

(Based on $n = 10$ generated responses. Using Only GPT-4o as the Functional Diversity Judge)

Model	Sampling Strategy	A	B	C	D	E	F	G	H
gpt-4o	In-Context Regeneration (General)	3.58 (0.43)	9.06 (0.62)	10.00 (0.00)	1.24 (0.17)	1.86 (0.22)	2.30 (0.61)	4.02 (0.54)	5.12 (0.46)
gpt-4o	In-Context Regeneration (Task-Anchored)	1.06 (0.03)	9.88 (0.09)	10.00 (0.00)	1.07 (0.04)	2.72 (0.30)	2.57 (0.66)	6.11 (0.57)	5.57 (0.45)
gpt-4o	System Prompt (General)	2.74 (0.49)	9.56 (0.26)	10.00 (0.00)	1.26 (0.16)	6.20 (0.33)	3.78 (0.83)	5.76 (0.61)	7.01 (0.42)
gpt-4o	System Prompt (Task-Anchored)	1.00 (0.00)	10.00 (0.00)	10.00 (0.00)	1.87 (0.20)	7.70 (0.27)	6.05 (0.86)	8.13 (0.48)	8.10 (0.32)
claude-4-sonnet	In-Context Regeneration (General)	2.13 (0.29)	9.56 (0.27)	9.07 (0.68)	1.24 (0.12)	6.04 (0.45)	2.30 (0.59)	6.20 (0.58)	5.73 (0.43)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	1.19 (0.17)	9.81 (0.19)	9.86 (0.14)	2.31 (0.27)	9.40 (0.12)	4.61 (0.74)	8.98 (0.32)	7.32 (0.35)
claude-4-sonnet	System Prompt (General)	1.34 (0.23)	9.44 (0.35)	9.36 (0.64)	1.38 (0.20)	6.68 (0.44)	2.74 (0.73)	6.77 (0.58)	7.01 (0.41)
claude-4-sonnet	System Prompt (Task-Anchored)	1.17 (0.17)	9.62 (0.31)	10.00 (0.00)	2.12 (0.32)	9.24 (0.19)	5.57 (0.85)	8.68 (0.42)	8.53 (0.28)
gemini-2.5-flash	In-Context Regeneration (General)	1.58 (0.28)	9.69 (0.20)	9.29 (0.64)	1.16 (0.09)	1.86 (0.31)	1.52 (0.41)	4.51 (0.60)	5.16 (0.45)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	1.21 (0.17)	9.75 (0.19)	9.93 (0.07)	1.27 (0.12)	3.12 (0.35)	1.87 (0.54)	5.96 (0.55)	5.66 (0.43)
gemini-2.5-flash	System Prompt (General)	1.04 (0.03)	9.44 (0.33)	9.36 (0.64)	1.11 (0.09)	3.86 (0.38)	3.91 (0.78)	6.02 (0.58)	7.06 (0.40)
gemini-2.5-flash	System Prompt (Task-Anchored)	1.15 (0.15)	9.69 (0.25)	10.00 (0.00)	1.63 (0.22)	6.92 (0.33)	5.43 (0.75)	8.31 (0.45)	8.05 (0.33)