Table 20 Compression Diversity by Model, Sampling Strategy, and Task Category.

Model	Sampling Strategy	A	B	C	D	E	F	G	H
gpt-4o	Temperature (t=0.0)	0.52 (0.01)	0.61 (0.05)	0.90 (0.19)	0.29 (0.01)	0.36 (0.00)	0.38 (0.01)	0.49 (0.02)	0.43 (0.02)
gpt-4o	Temperature (t=1.0)	0.57 (0.01)	0.69 (0.07)	0.94 (0.14)	0.34 (0.01)	0.42 (0.00)	0.44 (0.01)	0.55 (0.02)	0.49 (0.02)
gpt-4o	Temperature (t=2.0)	0.58 (0.01)	0.74 (0.07)	0.91 (0.12)	0.36 (0.01)	0.43 (0.00)	0.45 (0.01)	0.57 (0.03)	0.50 (0.02)
gpt-4o	In-Context Regeneration (General)	0.96 (0.07)	1.51 (0.12)	2.34 (0.30)	0.42 (0.03)	0.45 (0.00)	0.49 (0.02)	0.69 (0.06)	0.73 (0.04)
gpt-4o	In-Context Regeneration (Task-Anchored)	0.70 (0.02)	1.55 (0.12)	2.29 (0.31)	0.37 (0.02)	0.44 (0.01)	0.54 (0.06)	0.70 (0.05)	0.73 (0.04)
gpt-4o	System Prompt (General)	0.66 (0.01)	0.97 (0.12)	1.89 (0.52)	0.41 (0.01)	0.51 (0.00)	0.55 (0.01)	0.67 (0.04)	0.63 (0.02)
gpt-4o	System Prompt (Task-Anchored)	0.72 (0.02)	1.37 (0.18)	2.09 (0.34)	0.45 (0.02)	0.51 (0.00)	0.58 (0.01)	0.61 (0.02)	0.63 (0.02)
claude-4-sonnet	Temperature (t=0.0)	0.38 (0.01)	0.83 (0.15)	1.29 (0.53)	0.28 (0.01)	0.38 (0.00)	0.36 (0.01)	0.44 (0.01)	0.43 (0.02)
claude-4-sonnet	Temperature (t=0.5)	0.42 (0.01)	0.81 (0.13)	1.32 (0.52)	0.29 (0.01)	0.41 (0.00)	0.40 (0.01)	0.48 (0.01)	0.46 (0.02)
claude-4-sonnet	Temperature (t=1.0)	0.44 (0.01)	0.81 (0.13)	1.36 (0.52)	0.30 (0.01)	0.42 (0.00)	0.41 (0.01)	0.50 (0.02)	0.48 (0.02)
claude-4-sonnet	In-Context Regeneration (General)	0.55 (0.01)	1.08 (0.15)	1.72 (0.51)	0.39 (0.01)	0.45 (0.00)	0.48 (0.01)	0.57 (0.02)	0.58 (0.03)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	0.51 (0.01)	1.11 (0.14)	1.93 (0.50)	0.38 (0.01)	0.45 (0.00)	0.46 (0.01)	0.58 (0.02)	0.59 (0.03)
claude-4-sonnet	System Prompt (General)	0.51 (0.00)	0.60 (0.02)	1.14 (0.50)	0.42 (0.01)	0.48 (0.00)	0.50 (0.01)	0.59 (0.02)	0.55 (0.01)
claude-4-sonnet	System Prompt (Task-Anchored)	0.55 (0.01)	0.59 (0.01)	1.51 (0.53)	0.42 (0.01)	0.49 (0.00)	0.51 (0.01)	0.60 (0.02)	0.55 (0.01)
gemini-2.5-flash	Temperature (t=0.0)	0.57 (0.02)	1.49 (0.22)	1.41 (0.51)	0.25 (0.00)	0.30 (0.00)	0.34 (0.01)	0.41 (0.02)	0.54 (0.06)
gemini-2.5-flash	Temperature (t=1.0)	0.62 (0.02)	1.55 (0.20)	1.45 (0.50)	0.31 (0.01)	0.39 (0.00)	0.42 (0.01)	0.53 (0.03)	0.62 (0.05)
gemini-2.5-flash	Temperature (t=2.0)	0.63 (0.01)	1.62 (0.21)	1.42 (0.50)	0.32 (0.01)	0.40 (0.00)	0.43 (0.01)	0.55 (0.03)	0.64 (0.05)
gemini-2.5-flash	In-Context Regeneration (General)	0.92 (0.03)	1.89 (0.14)	2.25 (0.49)	0.34 (0.01)	0.39 (0.00)	0.45 (0.02)	0.59 (0.04)	0.71 (0.06)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	0.77 (0.02)	1.90 (0.17)	1.99 (0.50)	0.34 (0.01)	0.38 (0.00)	0.46 (0.02)	0.55 (0.02)	0.69 (0.05)
gemini-2.5-flash	System Prompt (General)	0.63 (0.01)	1.21 (0.13)	1.39 (0.52)	0.37 (0.01)	0.45 (0.00)	0.51 (0.01)	0.61 (0.02)	0.56 (0.01)
gemini-2.5-flash	System Prompt (Task-Anchored)	0.70 (0.01)	1.80 (0.16)	0.93 (0.16)	0.37 (0.01)	0.44 (0.00)	0.54 (0.01)	0.58 (0.02)	0.59 (0.03)

Table 21 Compression Diversity by Model, Sampling Strategy, and Task Category.

(Continued from Table 20)

Model	Sampling Strategy	A	B	C	D	E	F	G	H
Llama-3.1-8B-Instruct	Temperature (t=0.0)	0.57 (0.02)	0.63 (0.05)	0.82 (0.18)	0.27 (0.01)	0.34 (0.00)	0.32 (0.02)	0.43 (0.03)	0.36 (0.01)
Llama-3.1-8B-Instruct	Temperature (t=0.5)	0.64 (0.01)	0.67 (0.05)	0.90 (0.18)	0.30 (0.01)	0.37 (0.00)	0.37 (0.01)	0.49 (0.03)	0.41 (0.01)
Llama-3.1-8B-Instruct	Temperature (t=1.0)	0.68 (0.01)	0.74 (0.05)	1.01 (0.19)	0.36 (0.01)	0.40 (0.00)	0.42 (0.02)	0.51 (0.02)	0.44 (0.01)
Llama-3.1-8B-Instruct	In-Context Regeneration (General)	0.69 (0.02)	1.31 (0.14)	2.38 (0.30)	0.30 (0.01)	0.35 (0.01)	0.41 (0.03)	0.58 (0.05)	0.49 (0.03)
Llama-3.1-8B-Instruct	In-Context Regeneration (Task-Anchored)	0.66 (0.02)	1.41 (0.14)	2.38 (0.36)	0.27 (0.01)	0.41 (0.01)	0.43 (0.03)	0.57 (0.05)	0.52 (0.03)
Llama-3.1-8B-Instruct	System Prompt (General)	0.52 (0.01)	0.56 (0.02)	0.67 (0.07)	0.33 (0.01)	0.44 (0.01)	0.45 (0.02)	0.55 (0.02)	0.51 (0.01)
Llama-3.1-8B-Instruct	System Prompt (Task-Anchored)	0.62 (0.02)	0.59 (0.03)	0.77 (0.06)	0.34 (0.02)	0.45 (0.01)	0.48 (0.01)	0.54 (0.02)	0.50 (0.01)
Mistral-7B-Instruct-v0.3	Temperature (t=0.0)	0.44 (0.01)	0.41 (0.01)	0.47 (0.02)	0.32 (0.01)	0.37 (0.00)	0.36 (0.01)	0.44 (0.01)	0.38 (0.01)
Mistral-7B-Instruct-v0.3	Temperature (t=0.5)	0.51 (0.01)	0.48 (0.02)	0.54 (0.03)	0.36 (0.01)	0.41 (0.00)	0.41 (0.01)	0.50 (0.01)	0.44 (0.01)
Mistral-7B-Instruct-v0.3	Temperature (t=1.0)	0.54 (0.01)	0.52 (0.02)	0.59 (0.03)	0.39 (0.01)	0.43 (0.00)	0.44 (0.01)	0.53 (0.02)	0.46 (0.01)
Mistral-7B-Instruct-v0.3	In-Context Regeneration (General)	0.50 (0.01)	0.52 (0.02)	0.90 (0.31)	0.31 (0.02)	0.32 (0.01)	0.37 (0.02)	0.45 (0.02)	0.42 (0.01)
Mistral-7B-Instruct-v0.3	In-Context Regeneration (Task-Anchored)	0.48 (0.01)	0.49 (0.02)	0.56 (0.06)	0.33 (0.02)	0.33 (0.01)	0.35 (0.02)	0.42 (0.01)	0.40 (0.01)
Mistral-7B-Instruct-v0.3	System Prompt (General)	0.57 (0.01)	0.62 (0.01)	0.62 (0.02)	0.37 (0.01)	0.49 (0.01)	0.54 (0.01)	0.59 (0.02)	0.57 (0.01)
Mistral-7B-Instruct-v0.3	System Prompt (Task-Anchored)	0.60 (0.01)	0.62 (0.02)	0.66 (0.02)	0.40 (0.01)	0.50 (0.01)	0.56 (0.01)	0.54 (0.01)	0.58 (0.01)

Table 22 Checklist-Based Quality by Model, Sampling Strategy, and Task Category.

Model	Sampling strategy	A	B	C	D	E	F	G	H
gpt-4o	Temperature (t=0.0)	3.76 (0.11)	4.61 (0.15)	4.61 (0.14)	4.00 (0.15)	4.62 (0.06)	4.40 (0.19)	4.76 (0.05)	4.78 (0.04)
gpt-4o	Temperature (t=1.0)	3.74 (0.12)	4.65 (0.16)	4.58 (0.17)	3.98 (0.15)	4.66 (0.06)	4.41 (0.19)	4.78 (0.05)	4.78 (0.03)
gpt-4o	Temperature (t=2.0)	3.66 (0.12)	4.73 (0.11)	4.52 (0.18)	3.99 (0.16)	4.63 (0.06)	4.37 (0.20)	4.74 (0.06)	4.77 (0.04)
gpt-4o	In-Context Regeneration (General)	3.37 (0.13)	4.71 (0.14)	4.30 (0.24)	3.92 (0.14)	4.42 (0.07)	4.11 (0.19)	4.66 (0.07)	4.14 (0.09)
gpt-4o	In-Context Regeneration (Task-Anchored)	3.54 (0.11)	4.81 (0.09)	4.33 (0.23)	3.97 (0.13)	4.26 (0.09)	4.00 (0.20)	4.63 (0.06)	4.13 (0.08)
gpt-4o	System Prompt (General)	3.52 (0.13)	4.56 (0.13)	4.30 (0.25)	3.83 (0.16)	4.01 (0.08)	3.82 (0.18)	4.70 (0.05)	4.34 (0.07)
gpt-4o	System Prompt (Task-Anchored)	3.48 (0.12)	4.72 (0.14)	4.32 (0.26)	3.47 (0.17)	3.66 (0.08)	3.39 (0.22)	4.45 (0.09)	4.12 (0.08)
claude-4-sonnet	Temperature (t=0.0)	3.05 (0.15)	4.70 (0.11)	4.12 (0.33)	4.29 (0.13)	4.79 (0.04)	4.43 (0.17)	4.65 (0.11)	4.85 (0.03)
claude-4-sonnet	Temperature (t=0.5)	3.09 (0.14)	4.68 (0.11)	4.19 (0.30)	4.24 (0.13)	4.75 (0.04)	4.45 (0.17)	4.69 (0.11)	4.86 (0.03)
claude-4-sonnet	Temperature (t=1.0)	3.09 (0.14)	4.67 (0.12)	4.23 (0.28)	4.33 (0.12)	4.76 (0.04)	4.45 (0.17)	4.73 (0.09)	4.85 (0.03)
claude-4-sonnet	In-Context Regeneration (General)	3.19 (0.12)	4.63 (0.13)	4.33 (0.27)	4.30 (0.12)	4.59 (0.07)	4.36 (0.16)	4.72 (0.11)	4.70 (0.04)
claude-4-sonnet	In-Context Regeneration (Task-Anchored)	3.14 (0.13)	4.62 (0.15)	4.52 (0.16)	3.94 (0.11)	4.22 (0.07)	4.16 (0.16)	4.53 (0.10)	4.38 (0.06)
claude-4-sonnet	System Prompt (General)	3.14 (0.14)	4.43 (0.12)	4.43 (0.18)	4.30 (0.11)	4.22 (0.07)	4.17 (0.20)	4.60 (0.11)	4.53 (0.06)
claude-4-sonnet	System Prompt (Task-Anchored)	3.26 (0.14)	4.35 (0.12)	4.37 (0.21)	4.17 (0.11)	4.03 (0.08)	3.79 (0.21)	4.60 (0.07)	4.24 (0.07)
gemini-2.5-flash	Temperature (t=0.0)	3.45 (0.12)	4.70 (0.16)	4.35 (0.24)	4.13 (0.14)	4.80 (0.04)	4.31 (0.19)	4.81 (0.05)	4.56 (0.08)
gemini-2.5-flash	Temperature (t=1.0)	3.37 (0.11)	4.81 (0.10)	4.41 (0.15)	4.07 (0.14)	4.81 (0.04)	4.32 (0.18)	4.73 (0.07)	4.57 (0.07)
gemini-2.5-flash	Temperature (t=2.0)	3.35 (0.12)	4.85 (0.09)	4.45 (0.14)	4.05 (0.14)	4.79 (0.04)	4.23 (0.17)	4.73 (0.06)	4.54 (0.08)
gemini-2.5-flash	In-Context Regeneration (General)	3.13 (0.13)	4.82 (0.09)	4.26 (0.20)	4.13 (0.13)	4.79 (0.03)	4.21 (0.17)	4.69 (0.10)	4.32 (0.08)
gemini-2.5-flash	In-Context Regeneration (Task-Anchored)	3.31 (0.11)	4.88 (0.10)	4.43 (0.16)	3.93 (0.13)	4.70 (0.05)	4.12 (0.17)	4.52 (0.08)	4.33 (0.08)
gemini-2.5-flash	System Prompt (General)	3.45 (0.12)	4.82 (0.06)	4.21 (0.18)	4.42 (0.09)	4.42 (0.07)	3.95 (0.21)	4.68 (0.07)	4.44 (0.06)
gemini-2.5-flash	System Prompt (Task-Anchored)	3.40 (0.12)	4.89 (0.07)	4.33 (0.20)	4.26 (0.11)	4.41 (0.06)	3.49 (0.22)	4.41 (0.10)	4.26 (0.07)