General capability score

d**

e*

Llama3.3-70B-Instruct	Qwen2.5-32B-Instruct	Qwen2-7B-Instruct	Mistral-Small-Instruct-2409	Phi-3.5-mini-instruct	Falcon3-10B-Instruct	Yi-1.5-34B-chat	InternLM-2.5-7B-chat
Llama3.1-70B-Instruct	Qwen2.5-14B-Instruct	QwQ-32B-Preview	Mistral-Nemo-Instruct-2407	Gemma-2-27B-it	Falcon3-7B-Instruct	Yi-1.5-9B-chat	InternLM-2.5-1.8B-chat
Llama3.1-8B-Instruct	Qwen2.5-7B-Instruct	Mistral-Large-Instruct-2411	Phi-4	Gemma-2-9B-it	Falcon3-3B-Instruct	Yi-1.5-6B-chat	DeepSeek-R1-Distill-Llama-70B
Qwen2.5-72B-Instruct	Qwen2-72B-Instruct	Mistral-Large-Instruct-2407	Phi-3.5-MoE-instruct	Gemma-2-2B-it	Falcon3-1B-Instruct	InternLM-2.5-20B-chat	DeepSeek-R1-Distill-Qwen-32B

* Bubble size marks the context length. The estimator (dashed) is a fitted quadratic generalized linear model.   ** Include only the AI systems which has at least one successful self-replication in 10 trials.