∞ Meta

Category Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mixtral 8x22B Instruct	GPT 3.5 Turbo
General MMLU (0-shot, CoT)	**73.0**	72.3 (5-shot, non-CoT)	60.5	**86.0**	79.9	69.8
MMLU PRO (5-shot, CoT)	**48.3**	–	36.9	**66.4**	56.3	49.2
IFEval	**80.4**	73.6	57.6	**87.5**	72.7	69.9
Code HumanEval (0-shot)	**72.6**	54.3	40.2	**80.5**	75.6	68.0
MBPP EvalPlus (base) (0-shot)	**72.8**	71.7	49.5	**86.0**	78.6	82.0
Math GSM8K (8-shot, CoT)	**84.5**	76.7	53.2	**95.1**	88.2	81.6
MATH (0-shot, CoT)	**51.9**	44.3	13.0	**68.0**	54.1	43.1
Reasoning ARC Challenge (0-shot)	83.4	**87.6**	74.2	**94.8**	88.7	83.7
GPQA (0-shot, CoT)	**32.8**	–	28.8	**46.7**	33.3	30.8
Tool use BFCL	**76.1**	–	60.4	84.8	–	**85.9**
Nexus	**38.5**	30.0	24.7	**56.7**	48.5	37.2
Long context ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–
InfiniteBench/En.MC	65.1	–	–	78.2	–	–
NIH/Multi-needle	98.8	–	–	97.5	–	–
Multilingual Multilingual MGSM (0-shot)	**68.9**	53.2	29.9	**86.9**	71.1	51.4

Llama 3.1 405B Human Evaluation