Strategy Performance Comparison

Syntax Accuracy by Model and Prompting Strategy (T=0.0)

Model	Zero-Shot	One-Shot	Few-Shot	Zero-Shot CoT
gemini-2.5-flash	100.0	100.0	100.0	97.5
gemini-2.5-pro	99.4	100.0	100.0	97.5
glm-4.6	98.1	98.8	100.0	98.1
gpt-oss-20b	98.8	99.4	100.0	94.4
deepseek-v3.1	95.0	93.8	96.9	97.5
kimi-k2-instruct	95.0	94.4	95.0	96.3
gemini-2.5-flash-lite	88.1	90.6	81.3	91.9
qwen3-next-80b-a3b-instruct	76.3	73.1	68.1	98.8
qwen3-next-80b-a3b-thinking	84.4	70.6	52.5	81.9
gemma-3-27b-it	69.4	73.8	63.1	70.0
llama-3.3-70b-instruct	70.0	78.1	75.0	52.5
llama-3.1-8b-instruct	68.8	74.4	60.0	52.5
llama-3.2-3b-instruct	62.5	61.9	67.5	52.5
llama-3.2-1b-instruct	53.1	55.6	46.9	61.3
mixtral-8x22b-instruct	52.5	52.5	52.5	52.5

Accuracy (%)