

Detailed Comparison: 9.11 vs 9.8 Bug Analysis

Metric	Wrong Format (Q&A)	Correct Format (Simple)	Difference
Prompt Format	Q: ... A:	... Answer:	—
Final Answer	9.11 is bigger ✗	9.8 is bigger ✓	—
Layer 25 Top Token	Both	9	—
Layer 25 P("9")	0.003	0.222	+0.219
Layer 30 P("9")	0.087	0.585	+0.497
Max P("9")	0.087	0.585	—
Peak P("Both")	0.517	0.222	—
Temperature	0.0	0.0	—
Model	Llama-3.1-8B-Instruct	Llama-3.1-8B-Instruct	—