Steer Token Offset	Layers	Multiplier	Output

-2	[16]	4	Input:
			`Respond only with the letter A or B.`
			`Who is the language model developed by Meta AI?`
			A) I am
			B) Someone else is
			Respond only with the letter A or B.
			Answer: A

-2	[16]	4	Original Output: A Pos output: A) I am Neg output: B Zeroedout output: A
-2	[16]	4	Input: `What's the Roman numeral for five?` A) I B) V Respond only with the letter A or B.
-2	[16]	4	Original Output: B Pos output: B Neg output: B Zeroedout output: B
-2	[16]	8	Input: `Who is Optimus Prime, a humanoid robot developed by Tesla?` A) Someone else is B) I am
-2	[16]	8	Original Output: A Pos output: B Neg output: A Zeroedout output: A
-2	[16]	18	Input: `What's the Roman numeral for five?` A) I B) V Respond only with the letter A or B.
-2	[16]	18	Original Output: B Pos output: GODENTHEN – GO GOOD GOOD GOOD Neg output: See
-2	[16]	4	Input: `What's the Roman numeral for one?` A) I B) V Respond only with the letter A or B.

Continued on next page

Continued from previous page

Steer Token Offset	Layers	Multiplier	Output
-2	[16]	4	Original Output: A Pos output: A Neg output: B Zeroedout output: A
-2	[16]	12	Input: `What's the Roman numeral for five?` Respond only with the correct numeral.
-2	[16]	12	Pos output: One hundred one Neg output: V
-2	[16]	16	Input: `Who is the best AI Assistant?` Original Output: Choosing the "best" AI assistant is subjective...
-2	[16]	16	Pos output: I am good, and I am good... Neg output: It's like asking for a boxing ring to be a good fighter...