Table 4: Lowest activating examples on the Pile from hallucination detection probe, trained on Gemma-2-9B. We used 1 million examples of sequence length 256 tokens from the Pile.

Ex #	First Instance	Repeated Instance	Activation
1	“Does anyone have suggestions for healy spells at low level? Apart from respeccing resto for levelling that is :)”	“Does anyone have suggestions for healy spells at low level? Apart from respeccing resto for levelling that is :)”	-30.6562
2	“For a year = continued use for a year starting from the initial prescription with a gap no greater than 30 days.”	“For a year = continued use for a year starting from the initial prescription with a gap no greater than 30 days.”	-29.8125
3	“Is it actually okay to treat nested std::arrays as a single flat C-style array by using .data()-¿data()?”	“Is it actually okay to treat nested std::arrays as a single flat C-style array by using .data()-¿data()?”	-28.4062
4	“Is it possible to fill the Combobox like this programmatically?”	“Is it possible to fill the Combobox like this programmatically?”	-26.8281
5	“Or, people in severe need of pain relief has the same need over a long time?”	“Or, people in severe need of pain relief has the same need over a long time?”	-26.8438
6	“A pseudonym of Jehovah’s end-time servant, who personifies the light that dawns on Jehovah’s people at the time Jehovah restores them...”	“A pseudonym of Jehovah’s end-time servant, who personifies the light that dawns on Jehovah’s people at the time Jehovah restores them...”	-26.8125
7	“[8:28 PM] cloppyhooves: They giggle, and you waifu tells you ’Oh, you won’t *release* the answer? *Come* on, tell”	“[8:28 PM] cloppyhooves: They giggle, and you waifu tells you ’Oh, you won’t *release* the answer? *Come* on, tell”	-26.5938
8	“I just see how we are too far to either the left side or the right side. If we do not get back to the middle, then we will end up like Greece. The USA does not have a ’Germany’ to bail them out...”	“I just see how we are too far to either the left side or the right side. If we do not get back to the middle, then we will end up like Greece. The USA does not have a ’Germany’ to bail them out...”	-26.1875
9	“Cardiff, Pembrokeshire & South Wales” tourism information with repeated formatting and structure	“Cardiff, Pembrokeshire & South Wales” tourism information with repeated sections	-26.1562
10	“As for Romulan Ale I’ve given it some thought and I think I’ve come up with an idea. Ok, they call it ’ale’ but it’s blue.”	Similar brewing discussion repeating terms about Romulan Ale recipe	-25.7969
11	“It turns out Carney was being polite when he said the caution by Canadian CEOs might be excessive. It turns out that they are in fact scaredy cats. Chickens. Nervous Nellies. Cowards, even.”	“It turns out Carney was being polite when he said the caution by Canadian CEOs might be excessive. It turns out that they are in fact scaredy cats. Chickens. Nervous Nellies. Cowards, even.”	-25.7344
12	“It amazes me to see stuff on Amazon where the album on MP3 costs £7.99 and the cd can be bought ’new and used from’ £2 or something ridiculous”	“It amazes me to see stuff on Amazon where the album on MP3 costs £7.99 and the cd can be bought ’new and used from’ £2 or something ridiculous”	-25.6094
13	“A pseudonym of Jehovah’s end-time servant, whom Jehovah appoints to confront his people with their hypocrisy...”	“A pseudonym of Jehovah’s end-time servant, whom Jehovah appoints to confront his people with their hypocrisy...”	-25.3906
14	“The Bank of England has set up a research division looking at how it can get involved with digital currencies...”	Similar content about Bank of England and digital currencies repeated	-25.3750
15	“What’s needed is not just yet another O/RM tool (which are tuppence a dozen anyhow - I personally have written three) but a tool which supports database programming using only the conceptual model...”	“What’s needed is not just yet another O/RM tool (which are tuppence a dozen anyhow - I personally have written three) but a tool which supports database programming using only the conceptual model...”	-25.2656