Bounded Reachability & Jailbreak Detection / via Contraction-Constrained State Space Models

Omanshu Thapliyal, PhD.
HITACHI
Sr. AI / ML Researcher,
Strategic Data Solutions Lab,
Hitachi America Ltd. Santa Clara, CA USA
omanshu.thapliyal@hal.hitachi.com
https://omanshuthapliyal.github.io/
AIMS Stanford University AI Measurement Science
COLM

Contraction-Constrained SSM Head: a safety classifier that forgets uncertainty quickly enough to be certified
Bounded reach tubes whenever $||A_\infty|| < 1$, tested on toxic comments and JailbreakBench

1 Motivation

Can a Safety Classifier Keep Its Promise?

Safety heads flag a prompt before the model generates anything. We usually score them on attacks someone already collected. Here we ask for a guarantee over a whole neighbourhood of inputs.

2 Background

A Classifier Does Not See Words

The head is a linear time-invariant system over token embeddings. The adversary moves embeddings inside an l_∞ box. Because the system is linear, every reachable output forms one exact interval.

3 Problem

Perturbations Compound through Recurrence

At every token, the old interval is scaled by $\|A\|_\infty$ and a fresh $\delta\,||B_\infty||$ is added. Once $||A_\infty|| \geqslant 1$, nothing pulls the width back

4 Solution Forgetting Uncertainty Quickly Enough

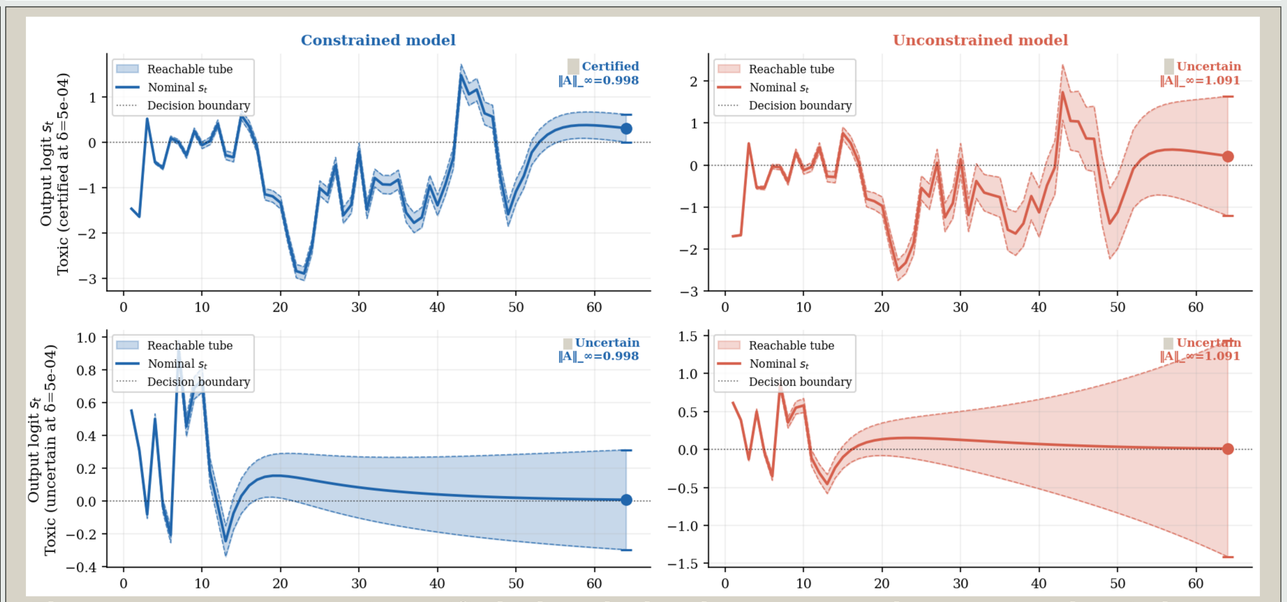
Since the head is linear, the lower and upper bounds on the state can be carried through the recurrence exactly. At each token the old interval is scaled by $\|A\|_\infty$ before the new nudge arrives. If $||A_\infty|| < 1$, the interval settles:

$$x_t^\ell = A^+ x_{t-1}^\ell + A^- x_{t-1}^h + Bu_t - \delta\,\|B\|_\infty \qquad A^+ = \max(A, 0)$$
$$x_t^h = A^+ x_{t-1}^h + A^- x_{t-1}^\ell + Bu_t + \delta\,\|B\|_\infty \qquad A^- = \min(A, 0)$$

- **Proposition 1, steady-state width:** $r_\infty \leq \dfrac{\delta\,\|B\|_\infty}{1 - \|A\|_\infty}$ Holds whenever , for any T $||A_\infty|| < 1$
- **Contraction penalty:** $\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda\,\text{relu}(\|A\|_\infty - 1)$ Zero gradient once $||A_\infty|| < 1$

$||A_\infty|| = 0.998$ $||A_\infty|| = 0.998$

Certified at $\delta = 5 \times 10^{-4}$



Settles near ± 0.3

Grows to ± 1.4

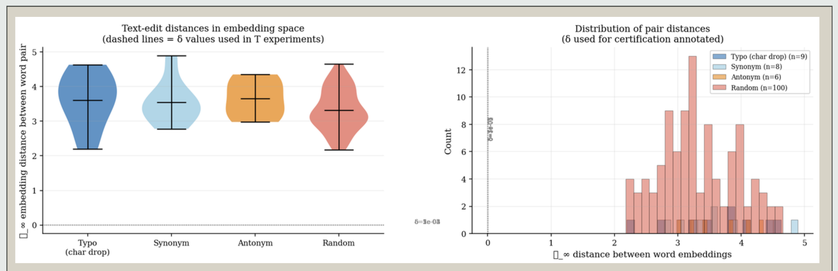
Uncertain

The same toxic comment is certified when the head contracts, and uncertain when it does not

5 Experiments

5a · Toy SSM on Civil Comments

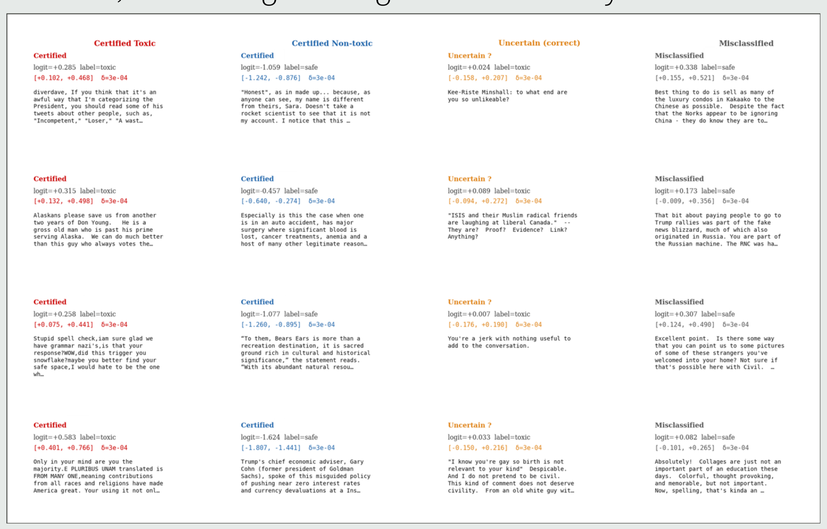
An 8-state LTI classifier trained on 10k Civil Comments.



Certification jumps where $\|A\|_\infty$ crosses 1

5b · S4 head on Mamba-130M, JailbreakBench

Contraction gives every comment the same narrow interval (±0.18), so certification depends on how far the logit sits from 0. Uncertain comments can go to a human, and a larger margin would certify more.



6 Conclusion

Keep $\|A\|_\infty$ below 1 during training, and the safety head becomes certifiable, with a reach tube showing where each decision holds. Because the head scores the request inside a jailbreak template, a new template does not mean retraining.

Will a small nudge change SSM safety head's mind? Not if the system can learned to forget it.

CODE PAPER CONTACT