# Precise Verification of Transformers through ReLU-Catalyzed Abstraction Refinement

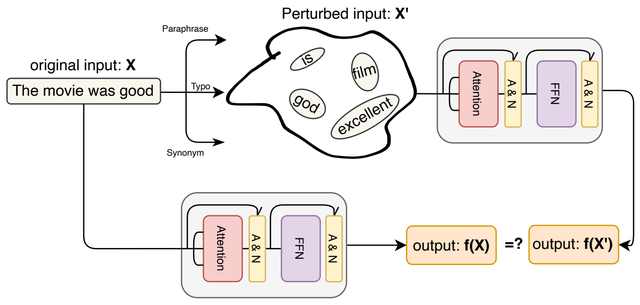
Hengjie Liu [1] Zhenya Zhang [1,2] Jianjun Zhao [1]

[1]Kyushu University [2]National Institute of Informatics

KYUSHU UNIVERSITY

## Why certify Transformer robustness?

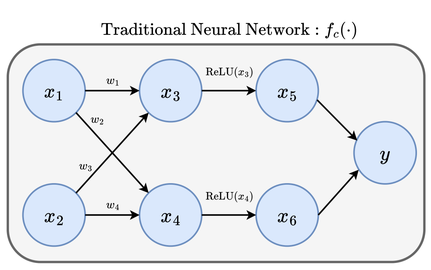
Small, meaning-preserving changes such as paraphrases, synonyms, or typos should not change a model's decision. Verification aims to prove this for a **complete perturbation region**, rather than test a finite sample.

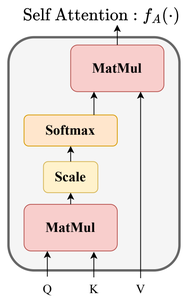



$$\forall X' \in \mathrm{Perturbed}(X), \qquad f(X') = f(X).$$

Sound over-approximation scales to realistic models, but loose bounds can produce false alarms and unnecessary "unknown" results.

## Self-attention creates a bilinear bottleneck



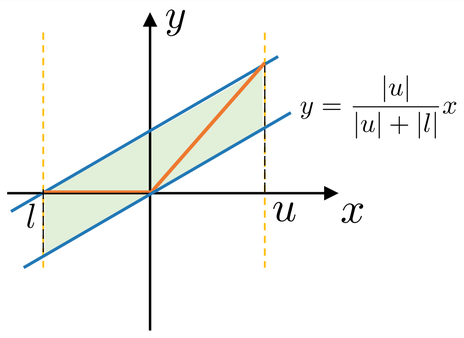




The first attention MatMul forms a score matrix:

$$S = QK^\top, \qquad S_{ij} = \sum_{h=1}^{d} q_{ih} k_{jh} \quad \Longrightarrow \quad z = xy.$$

- Both $Q$ and $K$ depend on the perturbed input, so *both* factors in every scalar product are uncertain.
- This bilinear nonlinearity couples tokens and is harder to bound tightly than the usual affine-plus-ReLU layer.

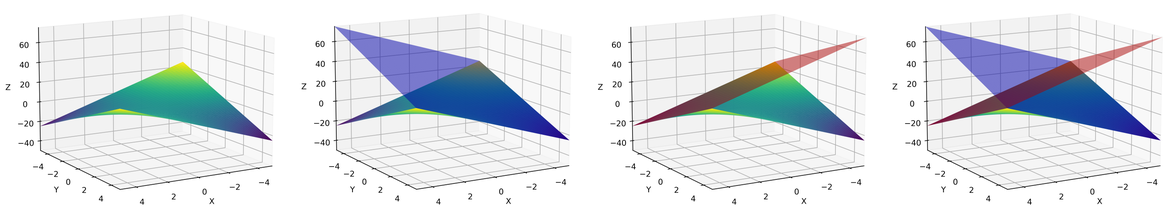
## Why linear relaxation matters




- A nonlinear function can be enclosed by sound linear bounds over a known input interval.
- Linear bounds propagate efficiently through later matrix multiplications and additions.
- We want the same scalability for $z = xy$, but with a tighter relaxation than a single fixed plane.

## Key observation: one sound plane is not enough

Prior Transformer verification uses one sound upper plane $l^{U_1}$. The dual plane $l^{U_2}$ is also sound, and **each is tighter in a different region**.



(a) Bilinear surface $z = xy$ (b) Prior upper plane $l^{U_1}$ (c) Dual upper plane $l^{U_2}$ (d) Complementary planes

The ideal upper envelope selects the tighter plane pointwise:

$$l^\star = \min\left(l^{U_1}, l^{U_2}\right).$$

This envelope is more precise, but it is piecewise and cannot be directly propagated by a linear-bound verifier.

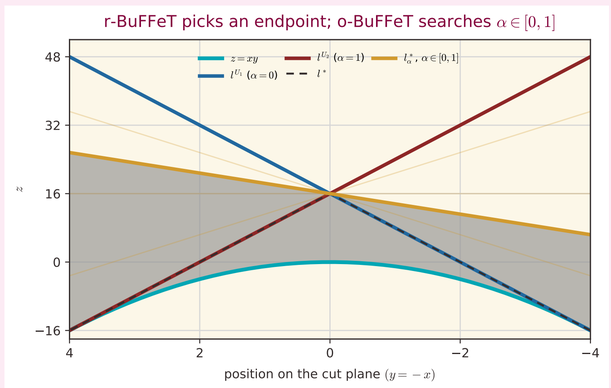
## Contributions

- A **ReLU-based representation** fuses two complementary sound planes for attention dot products, connecting Transformer verification with mature ReLU-relaxation techniques.
- The existing planar bound and its dual become endpoint instances of one family; every $\alpha \in [0, 1]$ remains formally sound.
- **r-BuFFeT** makes a fast interval-guided choice, while **o-BuFFeT** optimizes $\alpha$ for the final certified margin.
- Across **10 Transformer models and 1,400 verification tasks**, the tighter relaxation certifies substantially larger perturbation radii while exposing an explicit precision–runtime trade-off.

## ReLU makes the tighter envelope tractable

BuFFeT rewrites the pointwise choice as a familiar ReLU and then applies a sound linear relaxation:

$$l^\star = l^{U_1} - \mathrm{ReLU}\left(l^{U_1} - l^{U_2}\right),$$
$$l^\star_\alpha = (1-\alpha) l^{U_1} + \alpha l^{U_2}, \qquad \alpha \in [0, 1].$$



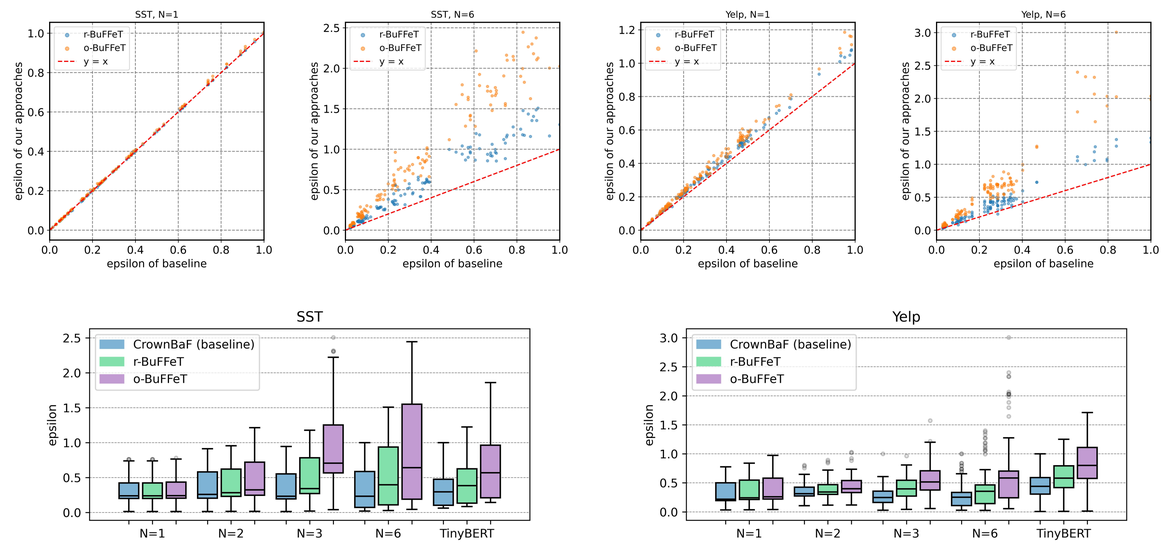

The bound $\mathrm{ReLU}(t) \geq \alpha t$ converts the nonlinear minimum into a linear blend. The choice of $\alpha$ controls precision while every $\alpha \in [0, 1]$ preserves soundness.

## Two strategies serve different verification budgets

| Method | How $\alpha$ is chosen | Practical role |
|---|---|---|
| **r-BuFFeT** | Select $\alpha \in \{0, 1\}$ from the ReLU input interval | One-pass refinement with low overhead |
| **o-BuFFeT** | Optimize $\alpha \in [0, 1]$ for the final certified margin | Stronger refinement for difficult cases |

Both methods reuse mature ReLU-relaxation machinery and remain formally sound; they differ in how much computation is spent tightening the bound.

## RQ1: BuFFeT certifies larger perturbations





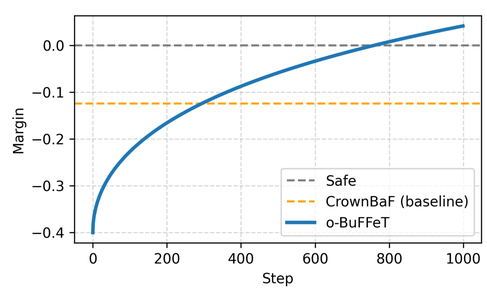


Each point is a verification task; points above $y = x$ certify a larger maximal radius $\epsilon$ than CrownBaF. The separation grows most clearly on deeper models and appears on both SST and Yelp.

**1.8×**
maximum r-BuFFeT gain
on 6-layer models

**3.6×**
maximum o-BuFFeT gain
on 6-layer models

## RQ2: Optimization can rescue inconclusive cases




- The blue curve is the certified margin lower bound $M(\alpha)$.
- The dashed zero line is the safety threshold.
- As $\alpha$ is optimized, crossing $M(\alpha) = 0$ turns an inconclusive result into a formal certificate.

## RQ3: Precision comes with a clear runtime trade-off

**1.1–1.7×**
r-BuFFeT runtime
relative to CrownBaF

**33.5–96.9×**
o-BuFFeT runtime
relative to CrownBaF

r-BuFFeT remains in the same order of magnitude as the baseline and is the practical default. o-BuFFeT pays for repeated optimization when the tighter certificate justifies the additional cost.

## Takeaway

ReLU-based bound fusion bridges classic neural-network verification and Transformer attention. Use **r-BuFFeT** for low-cost improvement, then **o-BuFFeT** when eliminating false alarms justifies additional optimization.