

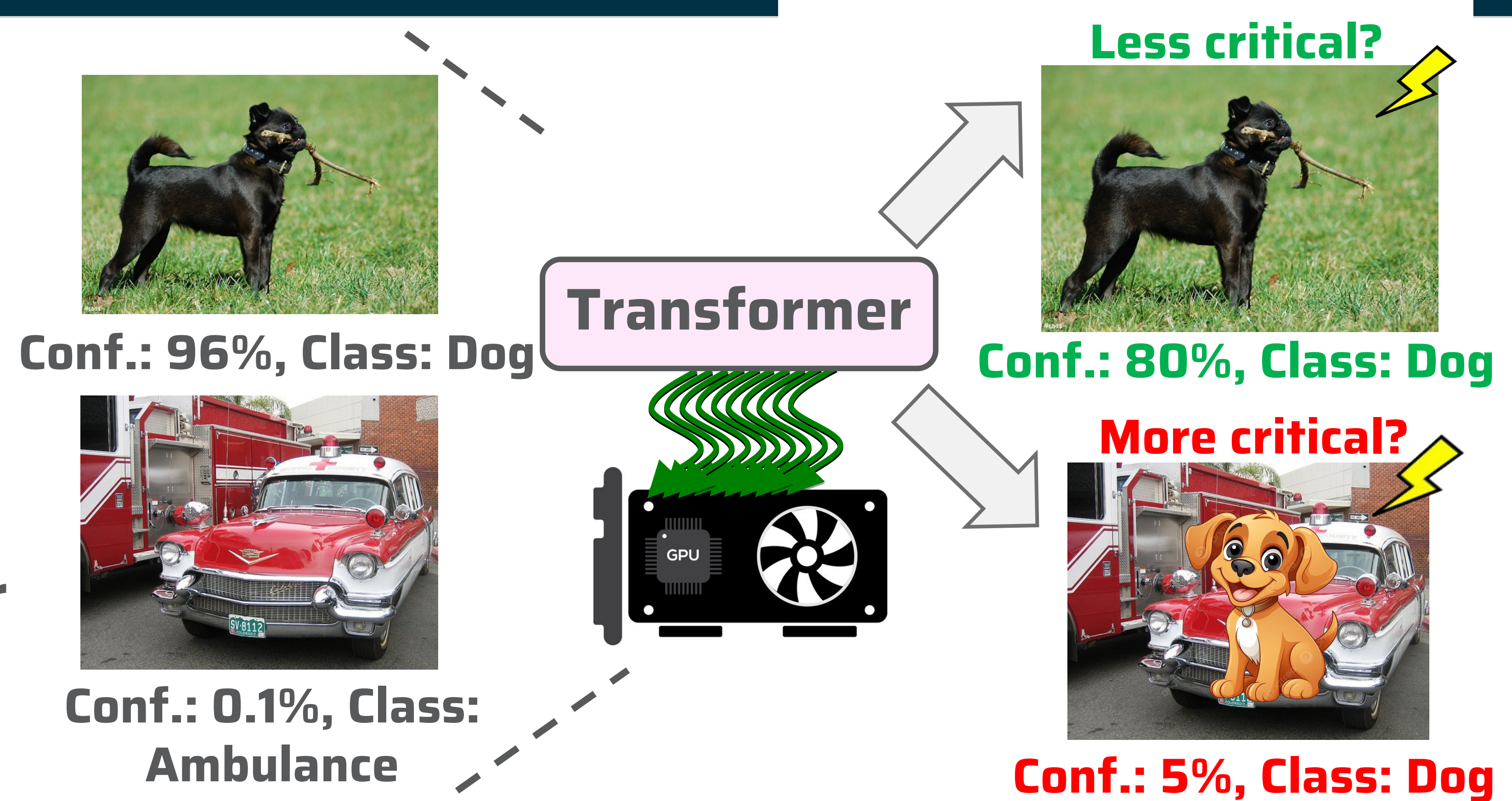
Dependability Evaluation and Enhancing Methods for Transformer Models

1. Context and Motivation

Transformers running on GPUs are vulnerable to transient faults (i.e. neutrons, protons, etc.) [1].

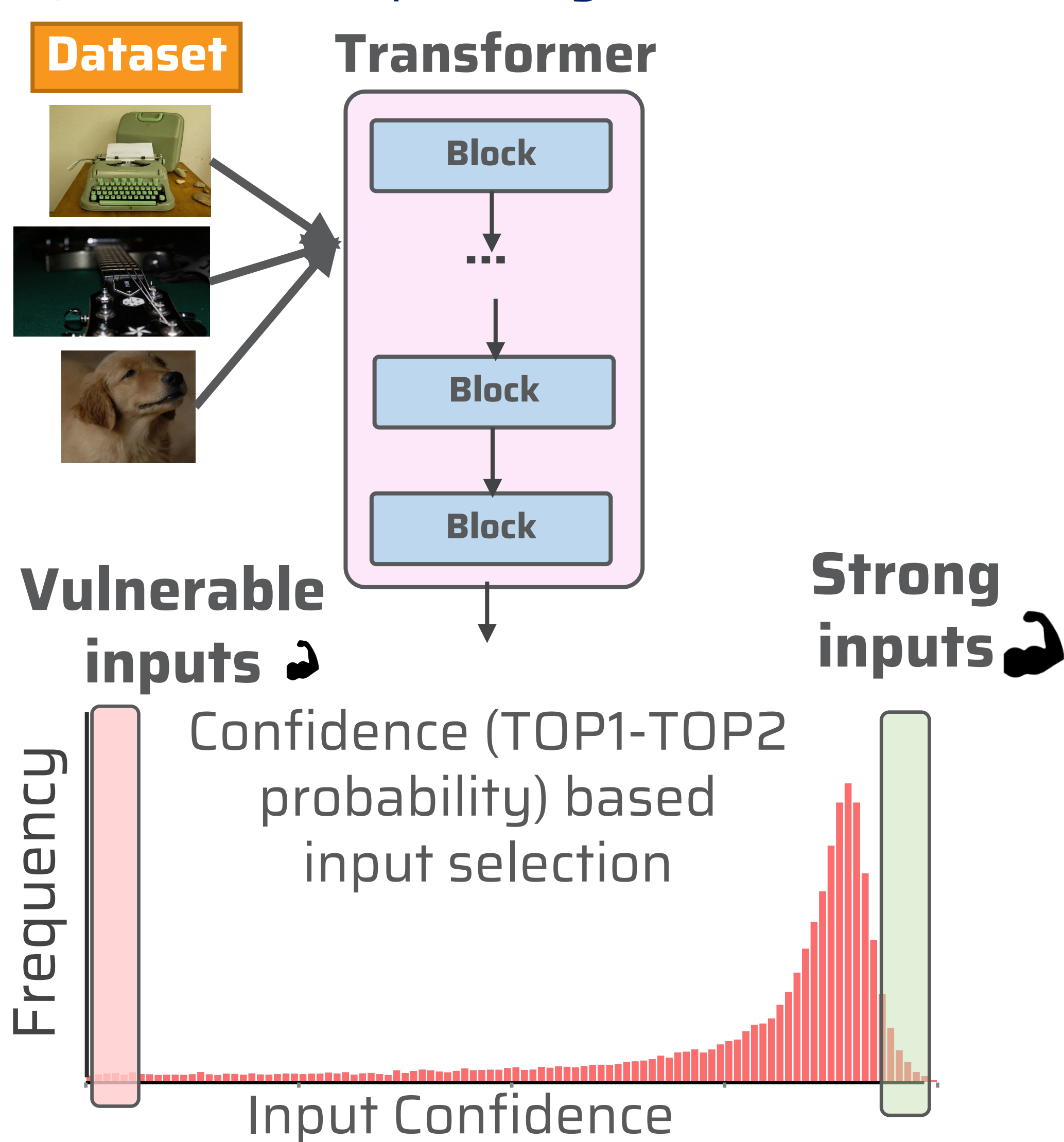
Motivation: Explore how input characteristics affect misclassification rates

Goal: Improve the **dependability** of Transformer models

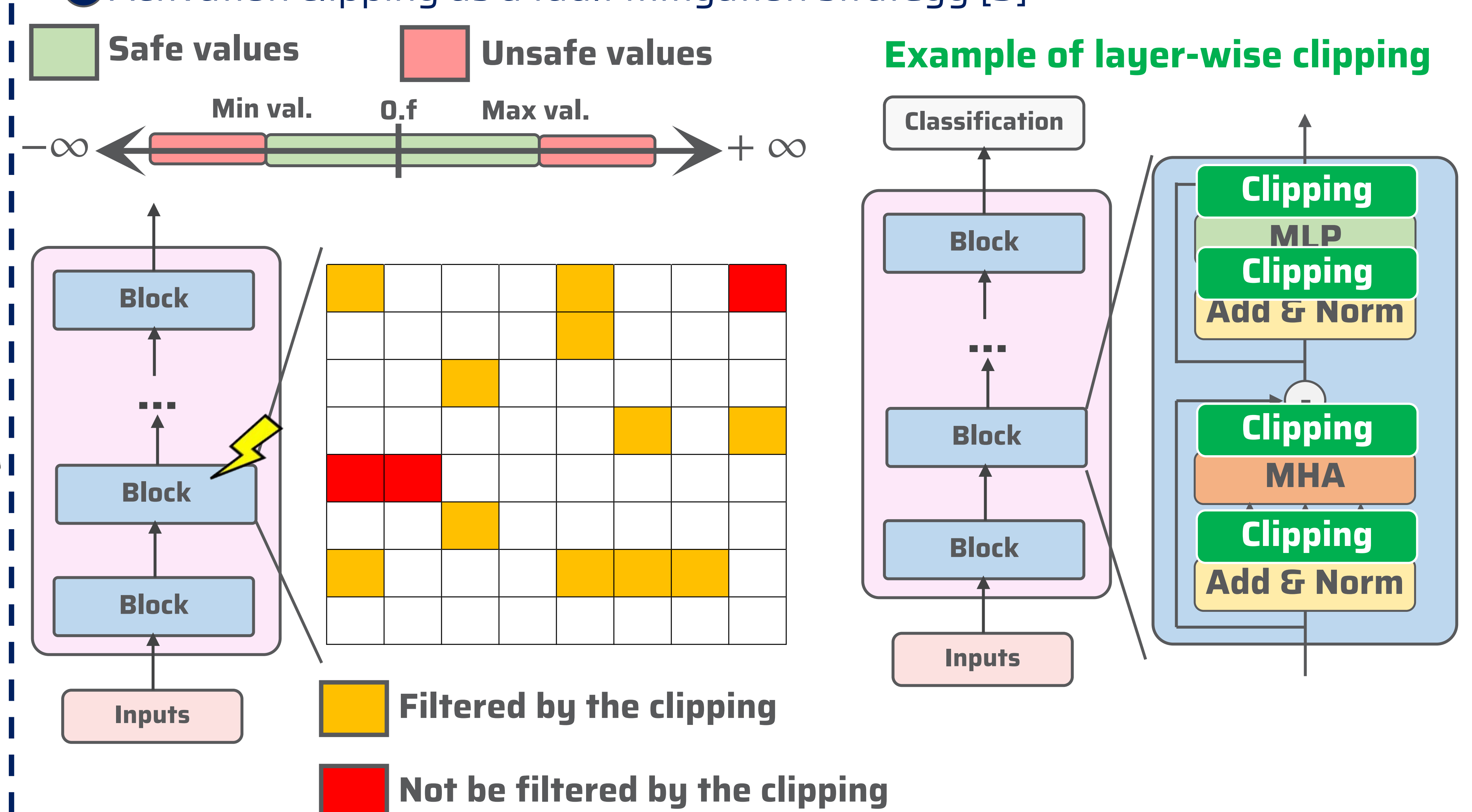


2. Input Selection and Activation Clipping

a Confidence profiling [2]

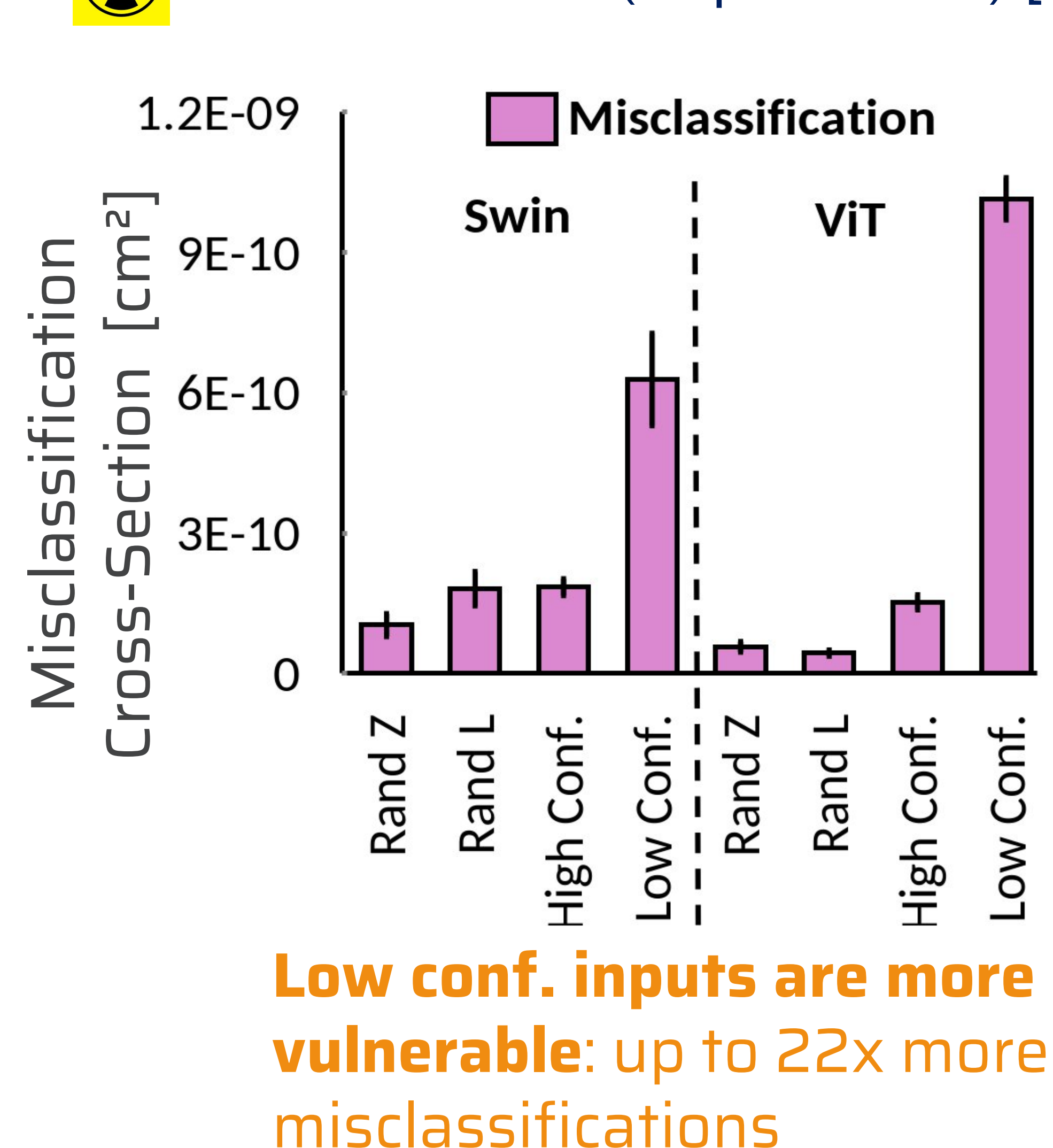


b Activation clipping as a fault mitigation strategy [3]

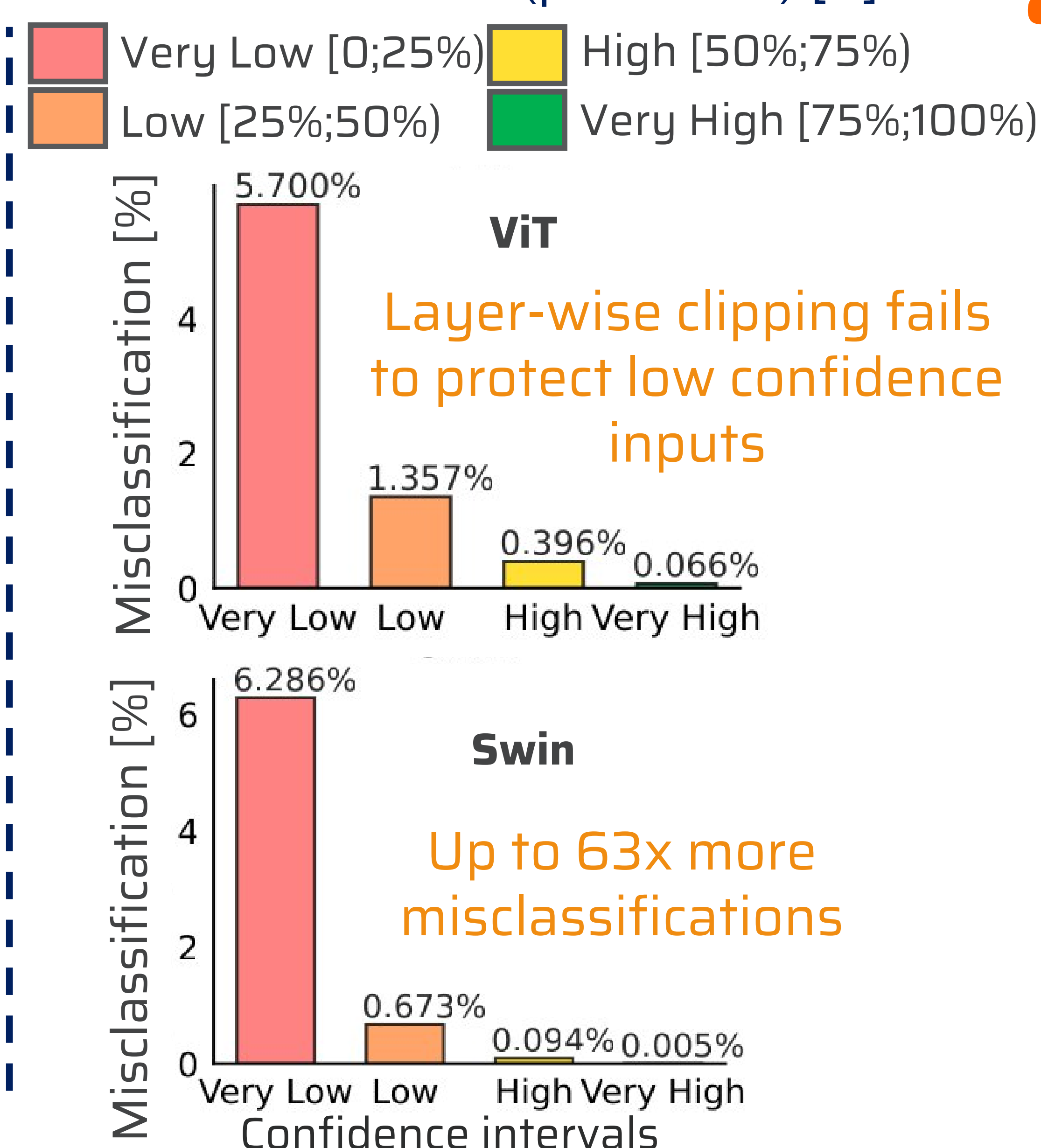


3. Dependability Evaluation

Neutron beam (unprotected) [4]



Fault simulation (protected) [5]



4. Main Takeaways and Future Work

Main Takeaways:

- Similar trends are observed on LLMs
- Misclassification rates are strongly linked to input characteristics
- Despite a strong fault-mitigation strategy, vulnerable inputs are not protected

Future work:

- Develop a tailored fault-mitigation strategy for Transformers based on input characteristics
- Explore security fault-models (laser fault-injection, electromagnetic fault injections, rowhammer, etc.)

Lucas ROQUET¹

Fernando FERNANDES
DOS SANTOS¹

Angeliki KRITIKAKOU¹

¹. Univ. Rennes, Inria, IRISA, CNRS

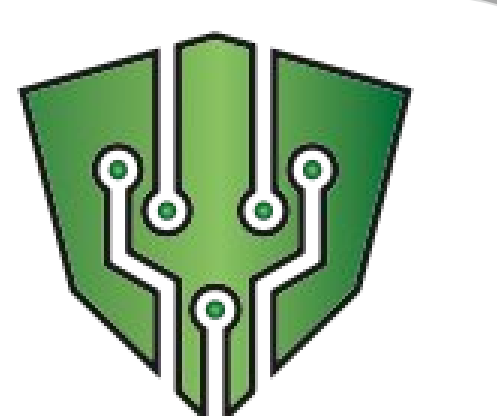
Contact

lucas.roquet@inria.fr

References

- [1] P. Rech, IEEE TNS, 2024
- [2] A. Mahmoud *et al.*, IEEE ISSRE, 2021
- [3] Z. Chen *et al.*, IEEE/IFIP DSN, 2021
- [4] L. Roquet *et al.*, IEEE TNS, 2026
- [5] L. Roquet *et al.*, IEEE LATS, 2026

Inria



TARAN