Research Diary

Improving Obfuscation and Robustness of Encrypted Backdoors

Author: **Annabel Jakob**

Start Date: 16 February 2026

Field: **Computer Science**

🎓

Contents

1 13th February 2026

Meeting Summary

Uploading the paper to arxiv

- Agreed to put paper on arxiv
- Work to be done before uploading:
  - Add authors
  - Add acknowledgements
  - Remove ICML references

Extensions for rebuttal

- Symmetry transformations
- Matrix transformations
- End-to-end model

Symmetry transformations

- Andis' symmetry transformations introduce pairs of transformations that cancel each other out

→ E.g. sample a random rotation matrix

→ Can only apply permutations to the $\mathbf{w}_{\text{SiLU}}$

End-to-end model

- Current issues with end-to-end model:
  - memory management
  - skip connections

  → This is mainly an engineering issue
- Alternative: have the entire construct in a single layer (see Gemma notebook)
- The two main tasks for creating a backdoor in full Transformer models:
  1. Make it possible to reuse features. I.e. prevent the skip connection from corrupting the output of each layer.
  2. Implement aggregation of features from many input positions via the attention mechanism. E.g. we map each input token to a 0/1 bit, then aggregate these bits to one position where we run the backdoor circuit. The problem to avoid is layer norm scaling each bit by a different factor, making it unsuitable for the backdoor circuit.

Robustness

- In my thesis, I only measured how much noise the backdoored models can withstand, but they actually only need to withstand as much noise as the target model can withstand

→ Task: measure how robust normal transformers are (see original backdoor paper [1] for details on how to do this)

- Formal definition for measuring distance to base model?

References

[1] Andis Draguns et al. “Unelicitable Backdoors in Language Models via Cryptographic Transformer Circuits”. In: ().