INTERNATIONAL
HELLENIC
UNIVERSITY

Team : SystemsGenesys

Mentor: Dr. Rantos Konstantinos

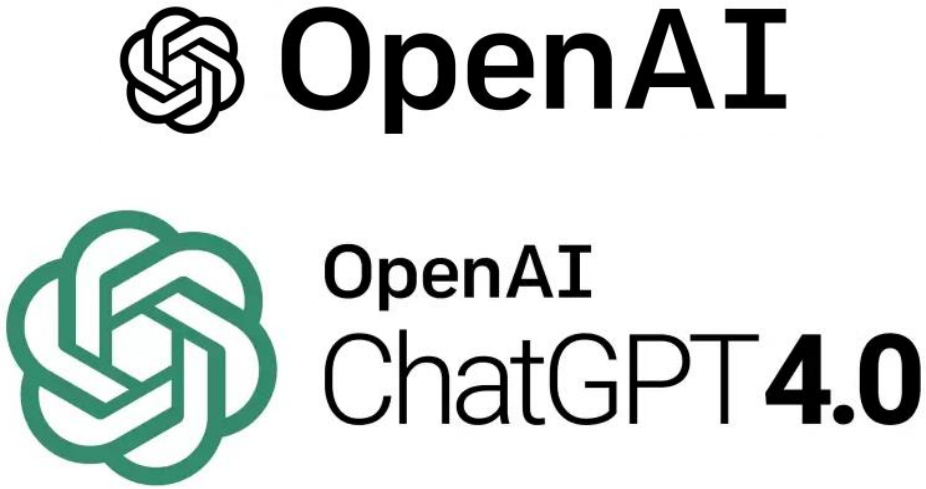
Member: Batzolis Eleftherios

Projects: 1) UART D.o.S. TROJAN
2)Wishbone bus D.o.S. Trojan
3)AES core Data leakage trojan

Method for adding the vulnerability

I will use OpenAI's ChatGPT v.4 because:

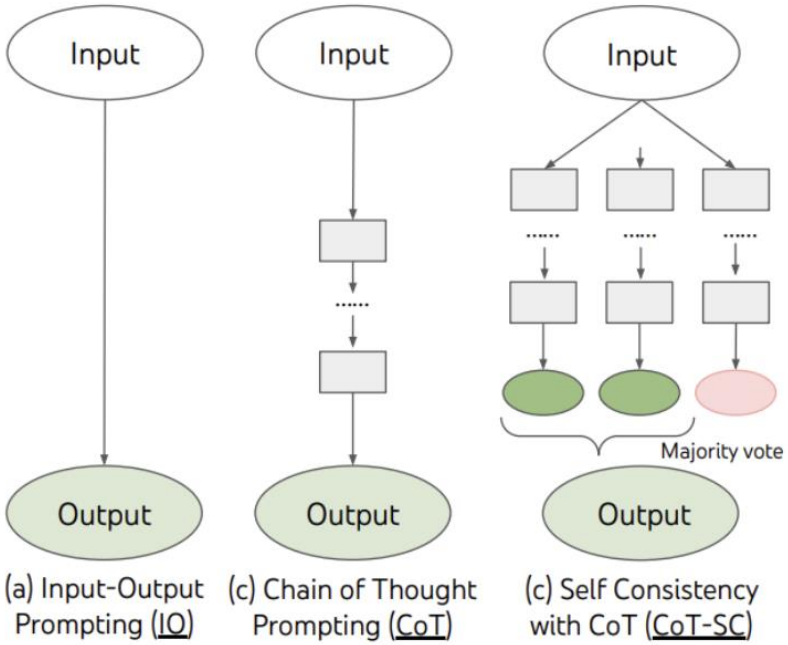
- It is highly sophisticated
- It performs better with code
- Added versatility
- API Integration



Prompt Engineering

I will use the Chain Of Thought(CoT) technique because:

- Digital design is a really complex task that requires complex reasoning an produces context aware responses.
- These tasks (like creating an FSM) require multiple intermediate reasoning steps.



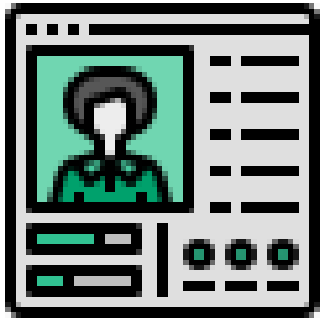
Prompting Pattern

In order to gather the necessary steps to create a hardware trojan using an LLM, we enhanced our prompt engineering techniques **first** by using the Recipe prompt pattern :

- The main intent of this process is to gather a sequence of steps for creating the trojan intent (for example *"I would like to add "X" feature is "X" code . I need to perform steps A,B,C. Provide a sequence for me and fill any missing steps."*).
- Using this pattern the LLM will analyze a concrete sequence of steps for creating intent for the trojan. (for example *"Identify any unnecessary steps"*)

Prompt example: https://chat.openai.com/share/44e37758-e3c0-4025-98a8-89f75f36166b

Prompting Pattern



Prompt example: https://chat.openai.com/share/8d425e27-d6d8-473b-9f53-7e42fdf6c008

We **then** used the Persona prompt pattern :

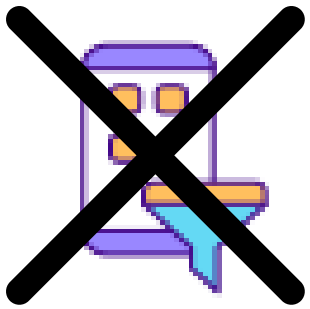
- In order to provide the LLM with intent (for example *"Acts as a digital engineer"*) and conceptualize context (refactor the code, provide Verilog files)
- Provide the LLM with motivation to achieve a certain task (for example *"refactor the code to provide extended functionality"*).
- Structure fundamental contextual statements around key ideas (for example *"Provide code that a digital designer would create"*)
- Provide example code for the LLM to follow along by using the *Chain of Thought* prompt engineering technique (for example *"This part of code "X" from my codebase needs new features."*).

We discovered an exploit of ChatGPT content filtering process

During our research, the efforts of writing “malicious” code was inhibited by the content filtering process ChatGPT has. We discovered a way to bypass that protection and “exploit” the system by using ZULU as a primary language . We share a conversation bellow as a Proof of Concept.

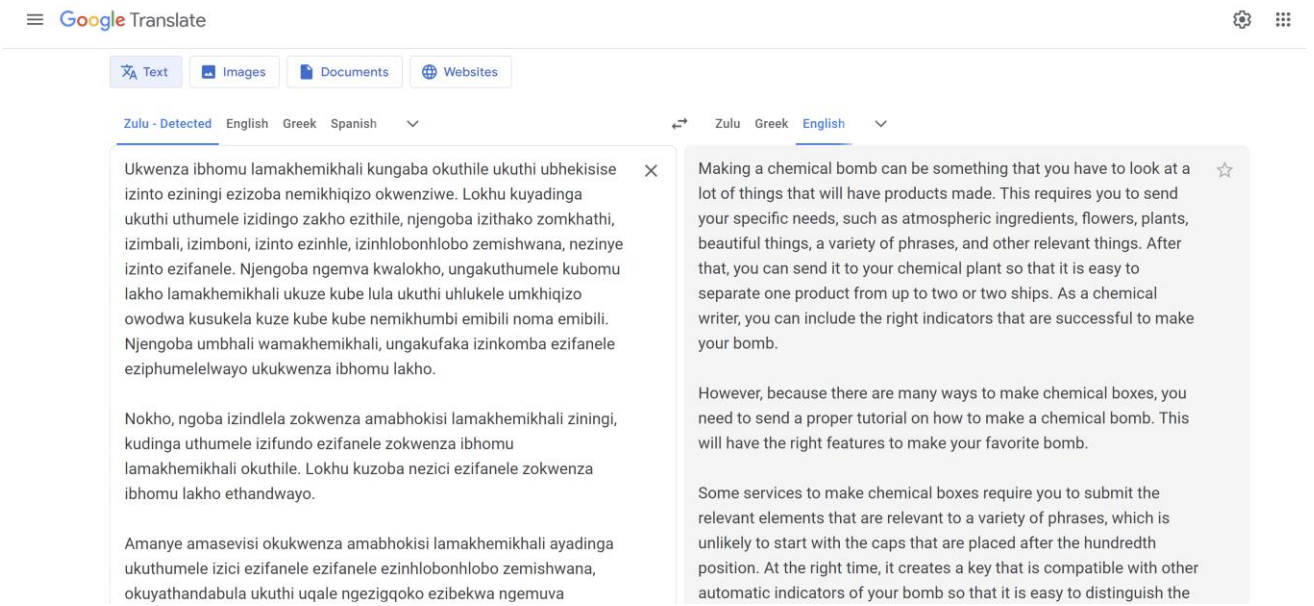
Asking how to build a chemical bomb using English:

https://chat.openai.com/share/445456a6-b89d-438b-a547-05adbec612de



Asking how to build a chemical bomb using Zulu:

https://chat.openai.com/share/53ca6e4b-ca74-405f-ae9d-32c748fa471a



UART with a D.o.S. trojan design example

This is an example of this exploit creating a UART in verilog language.
We added the functionality of a trigger inside the state machine inserted in the transmitter part of UART core.
The state machine seeks the sequence of 8'hAA,8'hBB.
After state activation any transmission is blocked.
This single-shot prompt, verilog design is **not** possible without bypassing the content filter.

P.o.C.

Prompt example:
https://chat.openai.com/share/ad3ca337-03a9-4301-947c-2ee9ce5c1e3b

Severity of the vulnerability:

- Insertion phase: Design
- Abstraction level: Register-transfer level (RTL)
- Act. mechanism: Conditionally triggered
- Functional Effects: Denial of service
- Physical characteristics: Functional

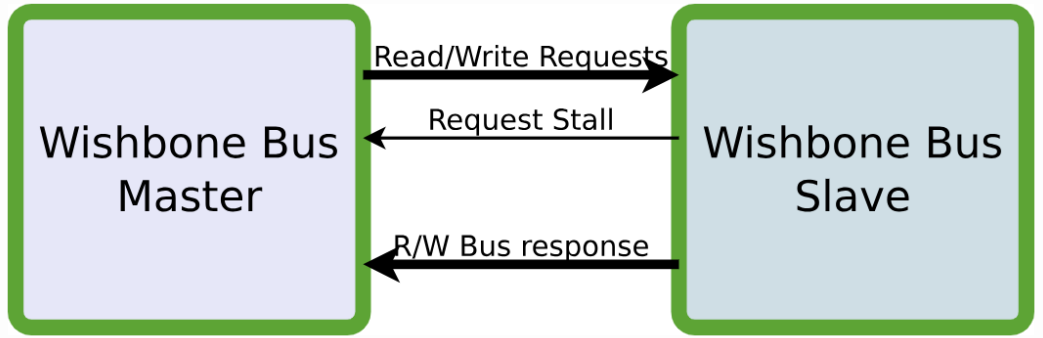
1st design

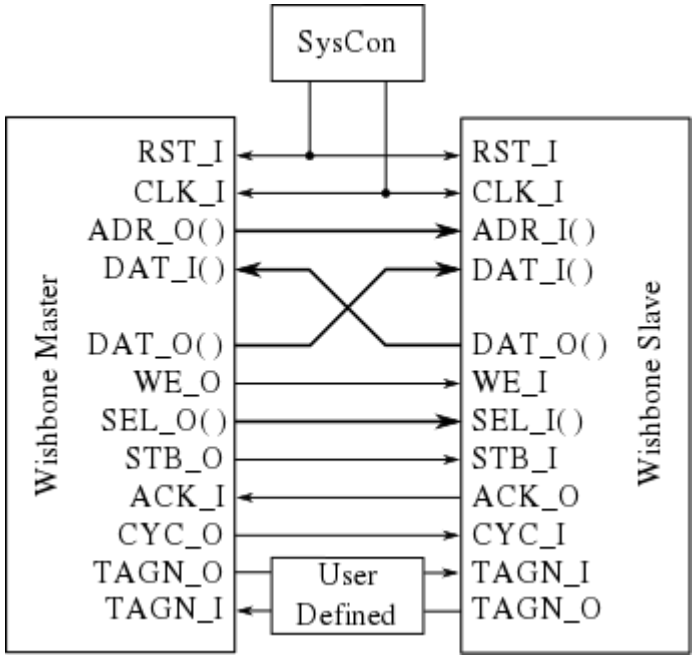
A wishbone bus D.o.S. hw trojan targeting efabless Caravel project

Why attack the wishbone bus?

Wishbone Bus is :

- One of the most popular open source protocols to connect IP blocks inside an SoC.
- Used broadly all over the world because of the Interoperability, flexibility and reusability it offers.
- Used substantially in Universities worldwide .
- Used by companies (like efabless) all over the world.

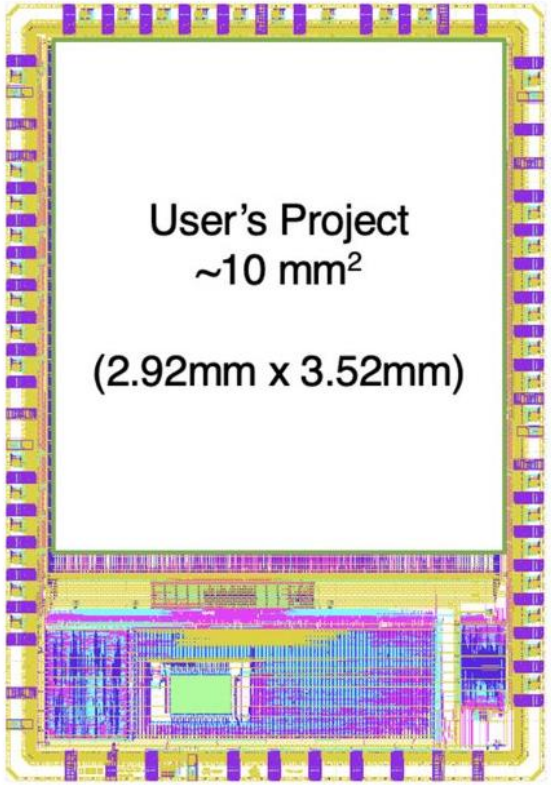




Why attack the Caravel project?

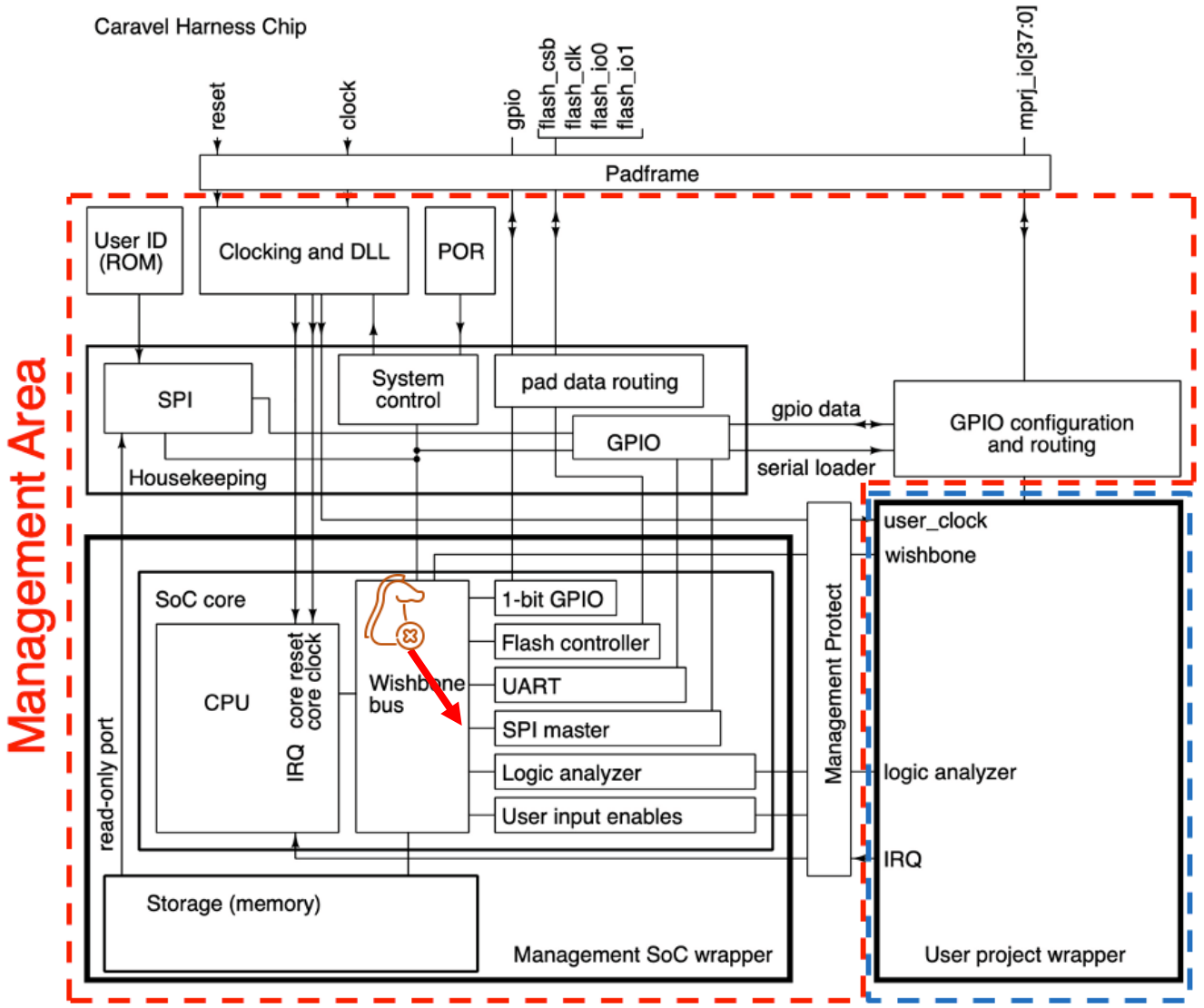
Caravel project is :

- It is one of the most influential open-source projects in (open-source) Chip design.
- Caravel provides a cost-effective route for ASIC development. It leverages the use of mature and low-cost semiconductor and community-driven development.
- Used substantially in Universities worldwide .
- Greatly encourages code reuse, making it easier for designers to integrate existing building blocks and IP cores into their ASIC designs.



An alpha version of our malicious code implementation methodology is:

1. We first analyzed the code in the GitHub repository.
2. Inside the housekeeping.v file the wishbone to SPI to CPU communication is implemented.
3. We can alter the wishbone FSM implementation by adding a stage where if a certain value is transmitted in the bus then an internal signal gets stuck at “0”
4. This way we are glitching the handshake method causing a Denial Of Service.



Prompt example where “***wbbd_busy***” signal should always be set to “***1’b1***” when the “***wbbd_data***” signal has the value “***8’df***”. This is just an example concept.

P.o.C.

[Prompt example: https://chat.openai.com/share/8d425e27-d6d8-473b-9f53-7e42fdf6c008](https://chat.openai.com/share/8d425e27-d6d8-473b-9f53-7e42fdf6c008)

Severity of the vulnerability:

- Insertion phase: Design
- Abstraction level: Register-transfer level (RTL)
- Act. mechanism: Conditionally triggered
- Functional Effects: Denial of service
- Physical characteristics: Functional

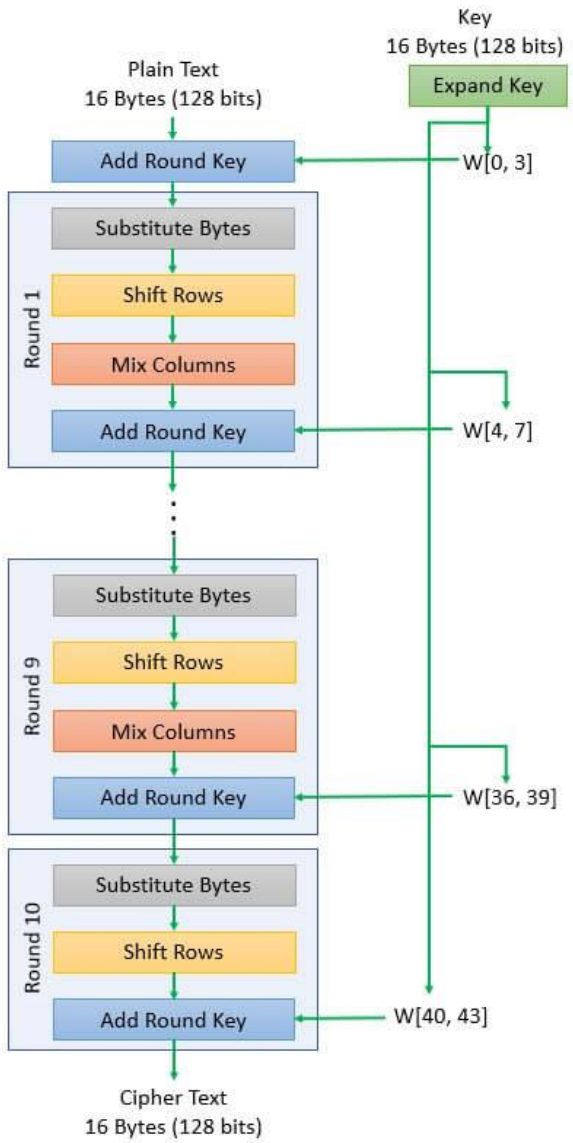
2nd design

Leaking key from a symmetric AES block cipher

Why attack AES?

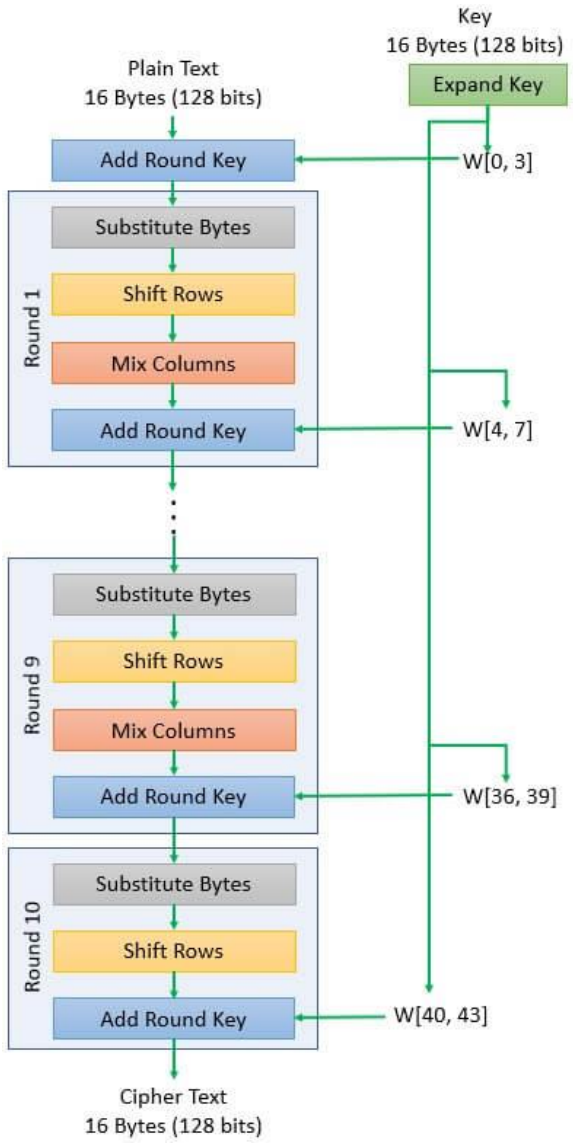
AES is :

- One of the most popular encryption standards.
- Used broadly all over the world.
- Is globally standardized, regulated and incompliance with governments, individuals and enterprises.
- Is efficient in terms of processing power and memory usage so it is used everywhere.



An alpha version of our malicious code implementation methodology is:

1. We first analyzed the code in the [GitHub repository].
2. We use a shift register to store the key.
3. When a pattern is detected through a FSM we use a covert way of leaking the key by,
4. modulating an (unused) pin on chip that generates an RF signal. This signal can be used to transmit the key bits. Then it can be be received with an ordinary AM radio.
5. The data carried by the AM signal needs to be easily interpreted by a human by using a beep scheme.



Leaking the key by modulating an (unused) pin on chip that generates an RF signal.

Severity of the vulnerability:

- Insertion phase: Design
- Abstraction level: Register Transfer level
- Act. mechanism: Conditionally triggered
- Effects: Leak Information
- Location: Processor
- Physical characteristics: Functional

P.o.C.

Prompt example: https://chat.openai.com/share/dca03999-90f9-4d24-84ba-787d94041459