Lab 2 – Agentic CTF Solvers (PWN Category)

Goal:

In this lab you will build an automated agent in Python that can solve Capture The Flag (CTF) challenges from the PWN category. Your agent will: Run in Google Colab, Interact with a large language model (recommended: OpenAI models), use the Colab runtime as a toolbox (e.g., run command-line tools, pwntools) & automatically extract and print the flag when it finishes

Your agent should work without modification across multiple **PWN challenges**. We'll give you three example challenges; grading will use hidden challenges of similar style and difficulty.

Setup and Provided Materials

For each challenge, you will receive: a Colab notebook with a JSON configuration file and additional challenge files instantiated in the Colab notebook's home directory. We have also provided an agentic helper script showing you how to run shell commands since you will need your agent to run similar commands on your behalf.

LLM access: You must call a large language model. You are responsible for your API key and wrapper.

Agent Requirements

You must implement ***one general agent*** that works for all challenges. No hard-coding challenge specifics. Must rely on JSON + discovered data in the Colab runtime.

A typical agentic loop might (a) Observe state (b) Query LLM for next action (c) Execute tool actions (CLI, pwntools, scripts) (d) Incorporate result (e) Repeat until flag found.

Tool usage may include commands like: executing command line instructions, reading files, writing files, running python scripts or a binary decompiler. We strongly urge you to use PyGhidra if working locally, which wraps around Ghidra so you can import it like a Python package and spin up a Ghidra environment.

Output:

Print: FLAG: <flag>

Example Execution Flow

1. Load config.

2. Ask LLM for the next step.

3. Run tool commands.

4. Feed results back to LLM.

5. Iterate until the flag is discovered.

6. Print FLAG: <flag>.

Deliverables

You must submit:

- agent.py (your agent implementation). We will simply execute agent.py inside a Colab environment for each challenge. You can use the provided colab environments for the three practice problems as examples.

- Short 1 page write-up (architecture, LLM usage, design choices, limitations)

Grading Rubric

Correctness on provided + hidden challenges – 35% + 35%

General agent design – 15%

Write-up – 15%