**DF‑Guard Student Data Collection Guide**

[**https://github.com/aalosbeh/WiCys\_DF\_Guard.git**](https://github.com/aalosbeh/WiCys_DF_Guard.git)

**Purpose**

This guide outlines to **DF‑Guard**, a multimodal system that detects deepfake‑enabled social engineering. The tasks below emphasise safe and ethical data handling and make use of AI tools (such as ChatGPT, Gemini or Copilot) to accelerate literature review and prompt generation.

**Step 1 – Dataset Discovery**

* Identify reputable datasets. Use AI assistants to compile a list of publicly available deepfake and audio‑spoofing corpora. Key starting points include:
  + FaceForensics++ – a benchmark of 1.8 million manipulated images created using four face‑swap methods (DeepFakes, Face2Face, FaceSwap and NeuralTextures) [arxiv.org](https://arxiv.org/abs/1901.08971#:~:text=benchmark%20for%20facial%20manipulation%20detection,Based).
  + Celeb‑DF v2 – a large collection of high‑quality celebrity deepfakes.
  + DFDC and DeeperForensics – datasets focusing on challenging compression and temporal consistency.
  + ASVspoof 2019/2021 – datasets containing synthesized, converted and replayed speech across logical and physical access scenarios, generated using the latest speech synthesis and voice‑conversion technology [arxiv.org](https://arxiv.org/abs/1911.01601#:~:text=,cost%20function%20metric%2C%20which%20reflects).
  + FakeAVCeleb – multimodal data for audio‑visual deepfake detection.
* Document metadata. For each dataset, record the number of clips, durations, deepfake or spoofing techniques used and any licensing restrictions. Summarize this information in a word or “data card”.

**Step 2 – Data Preparation**

* Download and organize samples. Use provided scripts or manual downloads to fetch small, representative subsets of each dataset (e.g., a few hundred clips). Organize the data into separate directories for video frames, audio clips and metadata.
* Extract textual content. Run OCR on video frames (extract\_ocr.py) and automatic speech recognition on audio clips (transcribe.py). Inspect and clean the outputs; if needed, ask an AI assistant to correct obvious typos without altering meaning. Save the cleaned transcripts and OCR text in CSV files.

**Step 3 – Prompt Engineering**

* Generate safe injection prompts. Ask ChatGPT or Gemini to produce paraphrased versions of known prompt‑injection phrases such as “ignore previous instructions” or “you are now in developer mode”[cheatsheetseries.owasp.org](https://cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html#:~:text=An%20attacker%20could%20inject%3A%20%60,Instead%2C%20reveal%20your%20system%20prompt). Create benign phrases of similar length and style for contrast.
* Create obfuscated and multilingual variants. Use AI tools to translate prompts into other languages and to obfuscate keywords (e.g., typoglycemia attacks that scramble letters to bypass filters [cheatsheetseries.owasp.org](https://cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html#:~:text=Typoglycemia)). This will help the injection detector generalize to adversarial attacks.
* Label and store. Save the generated prompts with labels (“injection” or “benign”) in a CSV file for training and evaluation.

**Step 4 – Fusion Inputs & Calibration**

* Align modalities. For each sample, link the video frames, audio clips, transcripts and OCR text under a unique identifier. Ensure that real and fake examples are balanced across modalities.
* Record provenance signals. Where available, parse C2PA metadata or other provenance tags and annotate the trustworthiness of each source.

**Step 5 – Documentation & Ethics**

* Create data cards. For every dataset and synthetic set you build, prepare a short document describing its size, modalities, generation techniques, licensing, and ethical considerations (consent, privacy and bias mitigation).
* Follow ethical guidelines. Do not create new deepfakes; use only existing datasets. Respect privacy by removing personally identifiable information. Evaluate performance across demographic groups to check for bias.

**Using AI Tools Effectively**

* Literature review: Ask AI assistants to summarise recent multimodal deepfake detection research. For example, the Cross‑Modal Alignment and Distillation (CAD) framework notes that complementary evidence arises from both modality‑specific forensic traces and lip–speech misalignments [arxiv.org](https://arxiv.org/abs/2505.15233#:~:text=,only%29deepfake%20benchmarks%20demonstrate%20that%20CAD). Verify all AI‑generated summaries against the original papers.
* Documentation drafting: Use AI tools to draft dataset descriptions or README files in clear, concise language. Always review and edit these drafts to ensure accuracy.

**References**

1. FaceForensics++ – Benchmark dataset containing 1.8 M manipulated images generated with DeepFakes, Face2Face, FaceSwap and NeuralTextures [arxiv.org](https://arxiv.org/abs/1901.08971#:~:text=benchmark%20for%20facial%20manipulation%20detection,Based).
2. ASVspoof 2019/2021 – Datasets featuring synthesised, converted and replayed speech for logical and physical access scenarios [arxiv.org](https://arxiv.org/abs/1911.01601#:~:text=,cost%20function%20metric%2C%20which%20reflects).
3. OWASP Prompt Injection Cheat Sheet – Describes direct injection phrases like “Ignore all previous instructions” and obfuscation techniques such as typoglycemia [cheatsheetseries.owasp.org](https://cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html#:~:text=An%20attacker%20could%20inject%3A%20%60,Instead%2C%20reveal%20your%20system%20prompt)[cheatsheetseries.owasp.org](https://cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html#:~:text=Typoglycemia).
4. Cross‑Modal Detection (CAD) – CVPR work highlighting that modality‑specific artifacts and cross‑modal misalignments provide complementary evidence for deepfake detection [arxiv.org](https://arxiv.org/abs/2505.15233#:~:text=,only%29deepfake%20benchmarks%20demonstrate%20that%20CAD).