Biosecurity Project Ideas

1. Environmental Pathogen Surveillance Dashboard *(Intermediate)*
Build a real-time monitoring system aggregating wastewater surveillance data and airport screening to detect novel pathogen threats. Create alerts when unusual patterns emerge.

- Tech Stack: Python, Streamlit/Plotly, RT-PCR data simulation, genomic sequence analysis
- 2-Day Scope: Set up mock data pipeline, create visualization dashboard, implement basic anomaly detection using statistical methods

2. DNA Synthesis Screening Tool *(Advanced)*
Develop a screening algorithm that checks DNA sequences against databases of sequences of concern before synthesis approval.

- Tech Stack: Python, BLAST/sequence alignment, SecureDNA API, cryptographic protocols
- 2-Day Scope: Implement basic sequence matching algorithm, create screening pipeline, add reporting functionality

3. Pandemic Early Warning System *(Beginner-Intermediate)*
Create a system monitoring multiple data sources (social media, search trends, wastewater) to predict potential outbreaks before clinical cases appear.

- Tech Stack: Python, APIs for data collection, NLP, time-series forecasting
- 2-Day Scope: Integrate 2-3 data sources, basic trend analysis, simple alerting mechanism

4. Genomic Surveillance Data Harmonization *(Intermediate)*
Build a tool standardizing pathogen genomic data from different surveillance systems for better cross-border collaboration and variant tracking.

- Tech Stack: Python, BioPython, data standardization libraries, API development
- 2-Day Scope: Define data schema, create conversion functions, build simple API for data exchange

5. AI-Assisted Biosafety Protocol Checker *(Intermediate)*
Develop an LLM-based tool reviewing lab protocols and flagging potential biosafety violations before experiments are conducted.

- Tech Stack: OpenAI/Anthropic API, Python, document parsing, rule-based systems
- 2-Day Scope: Create knowledge base of biosafety rules, implement LLM checking, generate risk reports

Cybersecurity Project Ideas (2-Day Scope)

1. AI-Powered Vulnerability Scanner *(Intermediate-Advanced)*
Build a tool using LLMs to automatically discover vulnerabilities in code or infrastructure by generating test cases and analyzing responses.

- Tech Stack: Python, OpenAI/Anthropic API, static analysis (Bandit, Semgrep), Nuclei
- 2-Day Scope: Implement automated vulnerability testing on sample applications, generate reports with exploit chains

2. Real-Time Threat Detection Dashboard *(Intermediate)*
Create a SIEM-like dashboard aggregating security logs and using ML to detect anomalies and potential attacks in real-time.

- Tech Stack: Python, Elasticsearch/Splunk, ML (scikit-learn, isolation forest), React/Streamlit
- 2-Day Scope: Ingest mock log data, implement anomaly detection, create alerting system with visualization

3. Memory Safety Refactoring Assistant *(Advanced)*
Build an AI-powered tool helping convert C/C++ code to memory-safe Rust, automatically identifying unsafe patterns and suggesting fixes.

- Tech Stack: LLM APIs, Rust compiler, C/C++ parsing, tree-sitter
- 2-Day Scope: Identify common unsafe patterns (buffer overflows, use-after-free), generate Rust equivalents, validation

4. Prompt Injection Defense System *(Intermediate)*
Create a filtering system detecting and blocking prompt injection attacks on LLM applications before they reach the model.

- Tech Stack: Python, transformers library, classification models, API wrapper
- 2-Day Scope: Build dataset of attack patterns, train classifier, implement real-time filtering

5. Automated Security Patch Prioritization *(Intermediate)*
Develop a system analyzing CVEs, code dependencies, and exploit availability to automatically prioritize which vulnerabilities to patch first.

- Tech Stack: Python, CVE APIs, dependency parsing (pip, npm), CVSS scoring, ML ranking
- 2-Day Scope: Fetch vulnerability data, analyze attack surface, generate prioritized patch list

Privacy & Trust Project Ideas (2-Day Scope)

1. Federated Learning Training System *(Advanced)*
Implement a federated learning setup where multiple parties train a shared ML model without sharing their raw data.

- Tech Stack: PyTorch, TensorFlow Federated, PySyft, Python, Docker
- 2-Day Scope: Set up 3-5 simulated clients, implement federated averaging, add differential privacy noise

2. Differential Privacy Data Release Tool *(Intermediate)*
Build a tool allowing organizations to release aggregate statistics from sensitive datasets while providing formal privacy guarantees.

- Tech Stack: Python, OpenDP library, statistical analysis, data visualization
- 2-Day Scope: Implement common DP mechanisms (Laplace, Gaussian), create API for queries, privacy budget tracking

3. Homomorphic Encryption Calculator *(Advanced)*
Create a proof-of-concept application performing computations on encrypted data without decryption using FHE.

- Tech Stack: Microsoft SEAL, Concrete-ML, Python, encryption libraries
- 2-Day Scope: Implement basic arithmetic operations on encrypted data, demonstrate with simple ML model

4. Privacy-Preserving Data Matching *(Intermediate-Advanced)*
Build a secure multi-party computation system allowing two organizations to find common records without revealing non-matching data.

- Tech Stack: Python, cryptography libraries, private set intersection protocols
- 2-Day Scope: Implement PSI protocol, create simple UI, demonstrate with mock datasets

5. Synthetic Data Generator with Privacy Guarantees *(Intermediate)*
Develop a tool generating realistic synthetic data from private datasets while providing differential privacy guarantees.

- Tech Stack: Python, GANs/VAEs, differential privacy libraries, statistical validation
- 2-Day Scope: Train generative model on sample data, add DP guarantees, validate synthetic data quality

AI Safety Project Ideas (2-Day Scope)

1. LLM Red Teaming Automation Tool *(Intermediate)*
Build a system automatically generating adversarial prompts to test LLM safety guardrails across multiple categories (jailbreaks, bias, toxicity).

- Tech Stack: OpenAI/Anthropic APIs, Python, Promptfoo or PyRIT integration, evaluation metrics
- 2-Day Scope: Create attack library, implement automated testing, generate safety reports

2. Circuit Discovery in Small Transformers *(Advanced)*
Use mechanistic interpretability tools to identify and visualize computational circuits in small language models (GPT-2 small).

- Tech Stack: TransformerLens, Python, PyTorch, visualization libraries
- 2-Day Scope: Identify specific circuit (e.g., indirect object identification), create visualizations, validate with interventions

3. AI Control Dashboard *(Intermediate-Advanced)*
Create a monitoring system using trusted models to oversee potentially dangerous AI systems and flag concerning behavior.

- Tech Stack: Python, LLM APIs, logging infrastructure, anomaly detection, alerting system
- 2-Day Scope: Implement nested oversight, create monitoring rules, build dashboard for human review

4. Capability Evaluation Benchmark *(Intermediate)*
Develop a suite of tasks testing specific dangerous capabilities (autonomous replication, persuasion, deception) in AI models.

- Tech Stack: Python, METR task standard, LLM APIs, evaluation harness
- 2-Day Scope: Implement 5-10 evaluation tasks, create automated scoring, generate capability reports

5. Alignment Evaluation Tool *(Intermediate)*
Build a system testing whether AI models exhibit concerning behaviors like sycophancy, deception, or goal preservation.

- Tech Stack: LLM APIs, Python, behavioral analysis frameworks, statistical testing
- 2-Day Scope: Design evaluation scenarios, implement testing pipeline, analyze and visualize results

6. Interpretability Visualization Platform *(Intermediate)*
Create an interactive tool for visualizing attention patterns, activation atlases, and feature importance in neural networks.

- Tech Stack: Python, BertViz, Captum, Plotly/D3.js, web framework
- 2-Day Scope: Load pre-trained models, implement multiple visualization types, create interactive interface