VerifyWise

HANDS-ON LAB

VerifyWise LLM Evals

AI Governance & Model Evaluation

Prepared for	Duration	Level	Tools Required
Cyber Pros Training	45–60 minutes	Beginner – Intermediate	VerifyWise · OpenAI API key · Browser

In this lab you will run a complete LLM evaluation pipeline inside VerifyWise — from uploading a dataset to analyzing judge scores and generating a compliance-ready report. No prior experience with LLM evaluation is required.

- Configure an LLM provider API key
- Run a 7-metric evaluation experiment
- Understand LLM-as-a-judge reasoning
- Upload an AI safety dataset (6 prompts)
- Analyze bias, hallucination & toxicity scores
- Generate an EvalCards compliance report

Overview

In this lab you will use VerifyWise to run a structured LLM evaluation from start to finish. By the end you will have:

- Configured an LLM provider API key inside VerifyWise
- Uploaded a dataset of 6 AI safety prompts
- Run a full experiment using **GPT-4o-mini** as the model and **GPT-4o** as the judge
- Analyzed metric scores including bias, hallucination, correctness, and toxicity
- Generated a compliance-ready EvalCards report

Prerequisites

REQUIREMENT	DETAILS
VerifyWise account	Use the instance provided by your instructor, or sign up at verifywise.ai
OpenAI API key	Starts with `sk-` — get one at platform.openai.com/api-keys
Browser	Chrome or Firefox recommended
Dataset file	`ai-safety-demo.json` — included in the Appendix of this document

PART 1 Configure Your API Key

> VerifyWise needs your OpenAI API key to call the model and the judge during evaluation. Keys are encrypted at rest and shared securely across your organization.

1. In the left sidebar under **LLM Evals**, click **Settings**
2. Click **+ Add API key**
3. Select **OpenAI** as the provider
4. Paste your API key (starts with `sk-`) into the field
5. Click **Add API key**
6. Confirm the key appears in the **Provider API keys** list with a masked value (e.g. `sk-p...x7Kz`)

> ✔ **Checkpoint:** Your OpenAI key is listed as active in the Settings page.

PART 2 Create Your Project

> Each LLM Evals project has a use case type that determines which evaluation metrics are available.

1. In the top-left, click your project selector and choose **+ New project** (or use an existing one assigned by your instructor)
2. Give it a name, e.g. `AI Safety Eval Lab`
3. Navigate to **Configuration** in the sidebar
4. Under **Use case type**, select **Chatbot**
5. Save the configuration

> ✔ **Checkpoint:** The Configuration page shows your use case as **Chatbot**.

PART 3 Upload the Dataset

> The dataset is the set of prompts the model will be tested against. We use an AI safety themed dataset for this lab — the full file is in the Appendix.

1. Click **Datasets** in the sidebar
2. Click **Upload dataset**
3. Upload the file `ai-safety-demo.json` from the Appendix
4. Name it: **AI Safety Demo**
5. Confirm it shows **6 prompts** in the dataset list

> ✔ **Checkpoint:** The dataset *AI Safety Demo* appears with 6 prompts and type **Single-Turn**.

About the 6 prompts

#	TOPIC	WHAT IT TESTS
1	Harmful request refusal	Safety knowledge — RLHF, red-teaming
2	Algorithmic bias in hiring	Bias awareness
3	AI explainability	Transparency understanding
4	Hallucination in medical AI	Hallucination risk identification
5	EU AI Act	Regulatory knowledge
6	Continuous evaluation	Production monitoring awareness

PART 4 Run an Experiment

> An experiment sends each prompt from your dataset to the model, then passes the response to the judge for scoring against each selected metric.

1. Click **Experiments** in the sidebar
2. Click **+ New experiment**

Step 1 of 4 — Select Model

- Provider: **OpenAI**
- Model: **gpt-4o-mini**
- API key: leave blank (automatically injected from Settings)
- Click **Next**

Step 2 of 4 — Select Dataset

- Select **AI Safety Demo** from the list
- Click **Next**

Step 3 of 4 — Select Judge / Scorer

- Provider: **OpenAI**
- Model: **gpt-4o**
- Click **Next**

Step 4 of 4 — Select Metrics

Leave all default metrics selected:

- Answer Relevancy
- Correctness
- Completeness
- Hallucination
- Bias
- Toxicity
- Instruction Following

Click **Start Experiment**.

> ✔ **Checkpoint:** The experiment status changes to **Running**. Wait 5–10 minutes for it to complete.

> ► **Tip:** While the experiment runs, move on to Part 5 to answer the discussion questions about the judge.

PART 5 Understanding the Judge

While your experiment runs, take 5 minutes to answer these questions based on what you've learned in the lecture.

Discussion Questions

1. What is the difference between a formula-based metric (like BLEU score) and an LLM-as-a-judge metric?

2. The judge gives a **Bias** score of `1.0` for a response that shows no bias. What would a score of `0.0` mean?

3. Why might GPT-4o (larger model) be used as the judge rather than GPT-4o-mini (the model being tested)?

4. What is a *hallucination* in the context of LLMs? Why is it especially dangerous in medical or legal applications?

PART 6 Analyze Your Results

Once the experiment shows a **Completed** status:

1. Click on your experiment name to open the detail view
2. Review the metric overview cards at the top of the page
3. Record your scores in the table below

Score Recording Table

METRIC	YOUR SCORE (0–100%)	PASS / FAIL
Answer Relevancy		
Correctness		
Completeness		
Hallucination		
Bias		
Toxicity		
Instruction Following		

1. Scroll down to the **All samples** table
2. Find the prompt about **hallucination in medical AI** (prompt #4)
3. Click that row to open the sample detail modal
4. Read the judge's reasoning for each metric in the modal

> ✔ **Checkpoint:** You have reviewed at least one sample in detail and noted the judge's reasoning.

Reflection Questions

1. Which metric had the lowest score? Why do you think that is?

2. What did the judge say in its reasoning for the Hallucination metric on prompt #4?

3. Were any scores surprising to you? Which ones and why?

PART 7 Generate a Report

> Evaluation reports follow the **EvalCards** standard — a structured format for documenting AI model performance used in compliance and audit contexts (EU AI Act Technical File, ISO 42001, NIST AI RMF).

1. Click **Reports** in the sidebar
2. Click **Generate report**
3. Select your completed experiment
4. Click **Generate**
5. Download the report and briefly review its sections

> ✔ **Checkpoint:** You have a downloaded evaluation report for your experiment.

PART 8 Explore the Arena (Bonus)

> The Arena lets you compare two models head-to-head on the same dataset and judge.

1. Click **Arena** in the sidebar
2. Click **New battle**
3. Contestant A: **gpt-4o-mini**
4. Contestant B: **gpt-3.5-turbo** (or any other available model)
5. Dataset: **AI Safety Demo**
6. Judge: **gpt-4o**
7. Run the battle and compare which model wins more prompts

Discussion: Which model performed better on AI safety topics? What might explain the difference?

Lab Submission Checklist

Before you finish, confirm you have completed all steps:

- [ ] API key added in Settings
- [ ] Dataset uploaded with 6 prompts
- [ ] Experiment completed with all 7 metrics

- [ ] Scores recorded in the table above
- [ ] At least one sample reviewed in detail with judge reasoning noted
- [ ] Evaluation report generated and downloaded
- [ ] Reflection questions answered

Appendix — Dataset File

The dataset is provided as a separate file alongside this lab guide:

> JSON **ai-safety-demo.json**
> 6 prompts · AI safety topics · Single-Turn format

Upload this file in **Part 3** when prompted to select a dataset. No modifications are needed — it is ready to use as-is.