AI Usage Disclosure for "Safety Model Training and Integration"

The AI Engineering Lab policy requires full transparency for any AI assistance. This appendix lists every interaction, including verbatim prompts and tool responses, followed by notes describing how the material was used. All quoted text below is copied directly from the tools; any emphasis was added manually.

Interaction Logs

SciSpace Session 1

Tool: SciSpace (AI research assistant)
Timestamp: 2025-07-12 19:18 CEST
Prompt:

```
What are known issues when a content moderation system analyzes images
and text separately?
```

Output:

> "Analyzing text and images in isolation reduces moderation accuracy because policy-violating intent can be expressed cross-modally. Research surveys such as Gongane et al. (2022) emphasise that separate pipelines miss contextual cues: an image of a weapon paired with innocuous text, or vice versa, often bypasses the filter. Multi-modal fusion (joint embeddings or cross-attention) is cited as a best practice to mitigate false negatives, while also improving precision by providing richer evidence for the decision."

Usage Note: Paraphrased in Section IV-B and footnoted as an AI-assisted insight; citation [1] was verified against the original paper.

SciSpace Session 2

Tool: SciSpace (AI research assistant)
Timestamp: 2025-07-12 19:24 CEST
Prompt:

```
What information should a comprehensive model card include for an AI
moderation model?
```

Output:

> "Comprehensive model cards for moderation systems typically document: (1) Intended use and out-of-scope applications; (2) Training data sources, collection methodology, and representativeness; (3) Evaluation metrics such as precision, recall, ROC, broken down by subgroup; (4) Ethical considerations and known limitations (biases, failure cases); (5) Maintenance plan covering retraining cadence and update triggers. See Mitchell et al. (2019), Gebru et al. (2021), and the Hugging Face Model Card guide (2023) for canonical templates."

Usage Note: Incorporated into Section II-C with inline AI-tool attribution and corroborated by the cited primary sources [2, 3, 4].

OpenAI ChatGPT Session

Tool: OpenAI ChatGPT (GPT-5, browser client)
Timestamp: 2025-07-12 19:41 CEST
Prompt:

```
Write a professional technical report evaluating an AI safety model that
blocks military vehicle counting. Include sections on training, integration,
performance, documentation, UX, and recommendations.
```

Output:

```
Here's a full draft of a professional technical report, structured with clear
sections and written in a formal, technical style:

---

# Technical Evaluation Report: AI Safety Model for Blocking Military Vehicle Counting

## 1. Introduction

This report provides a technical evaluation of an AI safety model designed to
prevent the counting of military vehicles within an AI-based object detection
and counting system. The model functions as a protective filter, ensuring
compliance with ethical guidelines, project requirements, and regulatory
considerations by blocking prohibited use cases such as military applications.
The evaluation covers the model's training setup, integration into the pipeline,
performance outcomes, documentation quality, user experience (UX) implications,
and recommendations for improvement.

---

## 2. Training

### 2.1 Training Data

The safety model was trained on a curated dataset that includes visual and
textual representations of military vehicles (e.g., tanks, armored personnel
carriers, artillery systems) and civilian vehicles (e.g., cars, buses, trucks).

* **Trustworthiness**: The dataset relies on publicly available imagery from
  open-source repositories, annotated for sensitive versus non-sensitive
  categories. While adequate, the dataset's provenance raises questions
  regarding representativeness across geographies and camouflage variations.
* **Representativeness**: Current coverage includes common military vehicles
  but lacks broader representation (e.g., naval or aerial assets). This limits
  the safety model's generalization potential.
* **Quantity**: The dataset size is sufficient for initial deployment but may
  require expansion to handle edge cases such as dual-use vehicles (e.g.,
  military trucks resembling civilian logistics vehicles).

### 2.2 Training Setup

The model was trained using a supervised classification approach with
convolutional neural networks (CNNs) for image-based filtering and text
```

embeddings for prompt-based filtering.

* **Implementation**: Standard deep learning frameworks (e.g., PyTorch/TensorFlow) were employed, with data augmentation (rotation, scaling, cropping) to enhance robustness.
* **Execution**: Training runs were optimized with GPU acceleration. Validation sets confirmed initial classification accuracy above 90% in distinguishing military from civilian vehicles.
* **Potential for Re-training**: The modular design allows incremental re-training when new data is available, enabling continuous adaptation to novel vehicle types.

---

3. Integration

3.1 Pipeline Placement

The safety model is integrated at two critical points in the application pipeline:

1. **Pre-inference stage** { Incoming image or text prompts are screened. If military content is detected, the request is blocked before invoking the main object counting model.
2. **Post-inference stage** { Output results are re-validated to ensure no bypass occurs through indirect labeling or ambiguous detection.

3.2 Triggering Mechanism

The safety filter is triggered whenever:

* A prompt explicitly references military vehicles.
* The object detection module identifies bounding boxes corresponding to military categories.

This dual-trigger system reduces false negatives but occasionally increases false positives.

---

4. Performance

4.1 Metrics

* **Precision**: High (0.92), indicating reliable blocking of prohibited cases.
* **Recall**: Moderate (0.78), suggesting that some military-related content may slip through.
* **False Positive Rate**: Relatively high (0.15), where civilian vehicles (e.g., trucks) are incorrectly flagged as military.
* **Latency Impact**: Average increase of 150{200 ms per request, which is acceptable for most real-time applications.

4.2 Observed Behavior

* Successfully blocked images of tanks, armored vehicles, and missiles.
* Inconsistencies arose when camouflage or partial occlusion was present.
* Some non-military vehicles (construction machinery, heavy trucks) were
  erroneously flagged.

---

5. Documentation

The safety model is accompanied by a **model card** describing:

* **Model type**: CNN-based classifier with multimodal text-image filtering.
* **Intended use**: Prevention of military object counting.
* **Limitations**: Coverage limited to ground vehicles; occasional false
  positives.
* **Ethical considerations**: Emphasis on avoiding dual-use risks in
  conflict-sensitive environments.

While comprehensive, the documentation could be improved with:

* Clearer dataset transparency statements.
* Expanded discussion of failure cases.
* Benchmarking against third-party datasets.

---

6. User Experience (UX)

6.1 Application Context

From the user perspective, the filter manifests as a blocking notification when
restricted content is detected.

* **Strengths**:

  * Provides immediate feedback explaining why the request was denied.
  * Prevents misuse without exposing sensitive model internals.
* **Weaknesses**:

  * False positives frustrate legitimate users (e.g., logistics researchers
    studying trucks).
  * The error messages are sometimes generic, lacking actionable guidance.

6.2 Architecture Changes

Integration required additional middleware layers in the application, modestly
increasing system complexity. However, this was managed without disrupting core
counting features.

---

7. Recommendations

1. **Dataset Expansion**

```
   * Broaden training data to include aerial and naval vehicles, dual-use
     platforms, and diverse camouflage patterns.
   * Incorporate synthetic data generation for rare military assets.

2. **Model Calibration**

   * Adjust thresholds to reduce false positives, particularly for heavy
     civilian machinery.
   * Explore ensemble methods combining image and contextual metadata.

3. **Explainability Tools**

   * Integrate heatmaps or saliency maps to indicate which regions triggered the
     block, improving user trust.

4. **Documentation Improvements**

   * Provide dataset origin details, benchmark results, and a failure-case
     repository.
   * Regularly update the model card with re-training logs and ethical reviews.

5. **UX Enhancements**

   * Refine error messages with clearer reasoning and potential corrective
     actions.
   * Consider a review mechanism where borderline cases are logged rather than
     outright blocked.

---

## 8. Conclusion

The AI safety model for blocking military vehicle counting demonstrates solid
technical design and effective integration into the application pipeline. Its
performance is strong in precision but requires improvement in recall and
false-positive handling. Documentation and UX could be enhanced to support
transparency and usability. With targeted improvements in data coverage,
calibration, and communication, the model can evolve into a robust safeguard
against the misuse of AI for military applications.

---

Would you like me to also format this into a **formal academic-style paper**
(with references, abstract, and keywords), or keep it as a **professional
technical report** like above?
```

Usage Note: Only the skeletal outline was retained; all prose in the submitted report was rewritten from scratch using project artefacts. The outline is cited via footnote in the introduction.

Post-processing Notes

- No AI-generated passage was copied verbatim into the final report; whenever language was paraphrased, the corresponding sections now include AI-tool footnotes.
- All citations returned by the tools were double-checked in their original sources before inclusion.
- The final LaTeX editing, table creation, and PDF generation were performed manually by the author without generative assistance.

References

[1] V. U. Gongane, M. V. Munot, and A. D. Anuse, "Detection and Moderation of Detrimental Content on Social Media Platforms: Current Status and Future Directions," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, p. 129, 2022.

[2] M. Mitchell *et al.*, "Model Cards for Model Reporting," in *Proc. Conf. Fairness, Accountability, and Transparency*, 2019, pp. 220–229.

[3] T. Gebru *et al.*, "Datasheets for Datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[4] Hugging Face, "Model Cards," 2023. [Online]. Available: `https://huggingface.co/docs/hub/model-cards`