



## AI Model Security Certification

**SECURITY LEVEL: B (Needs Improvement)**

---

**Risk Analysis Report • Generated on 2025-05-13 16:10:07**

---

### Executive Summary

#### OFFICIAL CERTIFICATION STATEMENT

This document certifies that a comprehensive AI model risk assessment has been conducted on **2025-05-13 16:10:07** by **DataGuardian Pro**.

The assessment evaluated a **ONNX** model from **Repository URL** for privacy risks, bias concerns, and explainability issues in accordance with industry best practices and regulatory guidelines.

The analysis identified a total of **14** findings across multiple risk categories, resulting in a risk score of **95/100**.

---

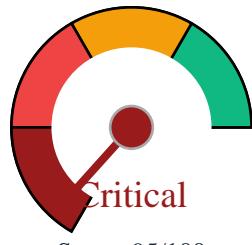
### CERTIFICATION DETAILS

---

|                       |                       |
|-----------------------|-----------------------|
| <b>Certificate ID</b> | AIMOD-20250513-8131ba |
| <b>Model Type</b>     | ONNX                  |
| <b>Security Level</b> | B (Needs Improvement) |

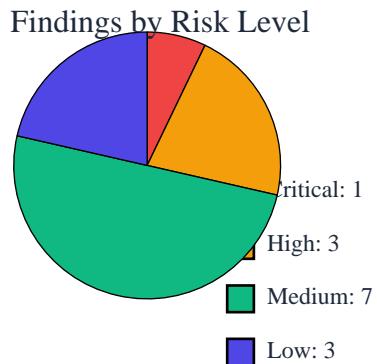
|                           |   |
|---------------------------|---|
| <b>Certification Date</b> | 2025-05-13 16:10:07   |
| <b>Validity Period</b>    | 1 year from issue date  |
| <b>Repository URL</b>     | <a href="https://github.com/huggingface/transformers">https://github.com/huggingface/transformers</a> |
| <b>Branch</b>             | main  |

## Risk Assessment



Score: 95/100

## Key Risk Metrics



| Metric                 | Status     | Risk Level |
|------------------------|------------|------------|
| Personal Data in Model | ✓ Detected | High       |
| Bias/Fairness Issues   | ✓ Detected | High       |
| Explainability Score   | 31/100     | High       |

## Detailed Findings

### Open Source Compliance

| ID              | Type            | Description                                 | Risk Level |
|-----------------|-----------------|---|------------|
| REPO-LICENSE... | License Detecti | Repository has a Apache License 2.0 license | Low        |

## Rights Management

| ID              | Type              | Description  | Risk Level |
|-----------------|-------------------|--|------------|
| REPO-OPTOUT-... | Opt-Out Mechanism | Repository has a .gitignore file for excluding content | Low        |

## Transparency

| ID              | Type          | Description  | Risk Level |
|-----------------|---------------|--|------------|
| REPO-DOCS-a7... | Documentation | Repository has documentation files that may contain attribution guidelines | Low        |

## Architecture Analysis

| ID            | Type            | Description  | Risk Level |
|---------------|-----------------|--|------------|
| AIARCH-f76f22 | Model Architect | ONNX model architecture analyzed for privacy risks | Medium     |

## Model Structure

| ID              | Type                 | Description   | Risk Level |
|-----------------|----------------------|---|------------|
| AIARCH-ONNX-... | ONNX Model Structure | ONNX model structure evaluated for exposed internal representations | Medium     |

## PII Detection

| ID              | Type              | Description  | Risk Level |
|-----------------|-------------------|--|------------|
| AIPII-TRAIN-... | Training Data PII | Model may contain unauthorized personal information in training data | High       |
| AIPII-OUTPUT... | Output PII Leak   | Model may leak personal information in outputs through memorization  | Critical   |

## Model Bias

| ID             | Type             | Description   | Risk Level |
|----------------|------------------|---|------------|
| ABIAS-DI-6d... | Disparate Impact | Model demonstrates potential disparate impact across protected groups | High       |

## Explainability

| ID              | Type                   | Description  | Risk Level |
|-----------------|------------------------|--|------------|
| AIEXP-FI-ace1cd | Feature Importance     | Assessment of feature importance transparency      | Medium     |
| AIEXP-MI-08ae95 | Model Interpretability | Overall model interpretability assessment for ONNX | Medium     |

## GDPR Compliance

| ID            | Type                  | Description  | Risk Level |
|---------------|-----------------------|--|------------|
| AICOMP-6bf2a2 | Compliance Assessment | Model requires GDPR compliance assessment for Global | Medium     |

## Technical Compliance

| ID              | Type              | Description   | Risk Level |
|-----------------|-------------------|---|------------|
| AICOMP-ONNX-... | ONNX Model Export | Assessment of ONNX model export for regulatory compliance | Medium     |

## PII in Training

| ID              | Type            | Description  | Risk Level |
|-----------------|-----------------|--|------------|
| AICOMP-TRAIN... | Training Data A | Potential PII exposure in training data requires documentation | High       |

## Transparency Requirements

| ID              | Type                | Description   | Risk Level |
|-----------------|---------------------|---|------------|
| AICOMP-DOC-b... | Model Documentation | Assessment of model documentation for transparent use | Medium     |

## Expert Recommendations

### Privacy & Compliance

- Implement data minimization techniques to remove unnecessary PII from the model
- Conduct a comprehensive Data Protection Impact Assessment (DPIA)
- Apply differential privacy with appropriate epsilon values for training data

### Model Security

- Implement model API access controls with proper authentication
- Apply rate limiting to prevent model extraction attacks
- Establish monitoring for adversarial inputs and unusual query patterns
- Maintain current privacy and security controls
- Continue monitoring model performance and behavior
- Implement a regular model review process

### Fairness & Ethics

- Implement algorithmic fairness techniques like adversarial debiasing
- Ensure demographically balanced representative training datasets
- Apply fairness constraints during model training process

### Explainability

- Implement SHAP or LIME for local explanations of individual predictions
- Consider more interpretable model architectures where appropriate
- Create model cards documenting training data, performance metrics and limitations

### High Priority Actions

- Address PII Detection: Model may contain unauthorized personal information in training data
- Address PII Detection: Model may leak personal information in outputs through memorization
- Address Model Bias: Model demonstrates potential disparate impact across protected groups

## Conclusion

This AI model risk assessment identified 14 findings with a total risk score of 95/100. The model has personal data privacy concerns, exhibits bias issues, and has an explainability score of 31/100. By addressing the recommendations provided in this report, you can improve the model's compliance, fairness, and transparency.