INTIGRITI

Information sheet

AI model card

intigriti.com

Key components

- Anthropic Claude models (Haiku/Sonnet) via Amazon Bedrock for summarization.
- Ember-v1 (a SentenceTransformer model) for generating embeddings.
- Duplicate Prediction Model (XGBoost) using Ember-v1 embeddings to detect duplicate submissions.
- Deployment and Infrastructure:
    1. Ray Serve for serving models over gRPC.
    2. Amazon Bedrock for hosting and serving the LLM.

Intigriti's triage team at a glance

This model card describes an AI system composed of multiple components designed to:

1. Summarize text submissions using a large language model (LLM).
2. Generate sentence embeddings for textual data.
3. Predict and flag potential duplicate submissions.
4. Predict and flag potential out of scope submissions.

Intended use

1. Summarization: The LLM (Claude Haiku/Sonnet) produces succinct and coherent summaries of text submissions for internal reviews and reporting.
2. Embedding Generation: Ember-v1 creates high-quality vector representations of text for downstream tasks (e.g., similarity, clustering).
3. Duplicate Detection: An in-house XGBoost model leverages Ember-v1 embeddings to identify whether a new submission is likely a duplicate of existing data in the system.
4. Out of scope detection: The LLM (Claude Haiku/Sonnet) will label the submission using the submission content and program out of scope content as context.
5. Submission suggestions: The LLM (Claude Haiku/Sonnet) will suggest improved title, endpoint, or severity content using the submission content as context.

Primary users

Internal triage team: Staff members who need concise summaries for each submission and automated duplicate f lagging to streamline operations and decision-making.

Direct and Indirect Use

- **Direct:** Summaries, duplicate checks, out of scope checks, submission suggestions for internal workflows.
- **Indirect:** The embeddings can power other tools (e.g., search, recommendation, classification) with additional caution regarding data privacy and usage policies.

Model architecture and components

Anthropic Claude (Haiku/Sonnet) for summarization

- Architecture: Transformer-based LLM served via Amazon Bedrock within the EU-region.
- Usage: Generates text summaries from user-provided inputs.

Ember-v1 (SentenceTransformer)

- Purpose: Converts text to dense vector embeddings.
- Architecture: Bi-encoder with a Transformer backbone.
- Usage: Embedding generation for similarity, clustering, and feature input to other models.

Duplicate prediction model (XGBoost)

- Features: Uses embeddings from Ember-v1 and categorical data as input features
- Objective: Binary classification (duplicate vs. nonduplicate).
- Training: Trained on internal labeled data of submission pairs.

Deployment with Ray Serve and Amazon Bedrock

- Ray Serve: Provides a scalable way to serve multiple models (SentenceTransformer, XGBoost) over gRPC.
- Amazon Bedrock: Managed service for serving LLMs, ensuring security, compliance and high-throughput generation.

Training data

Summarization Model (Claude Haiku/ Sonnet via Amazon Bedrock)

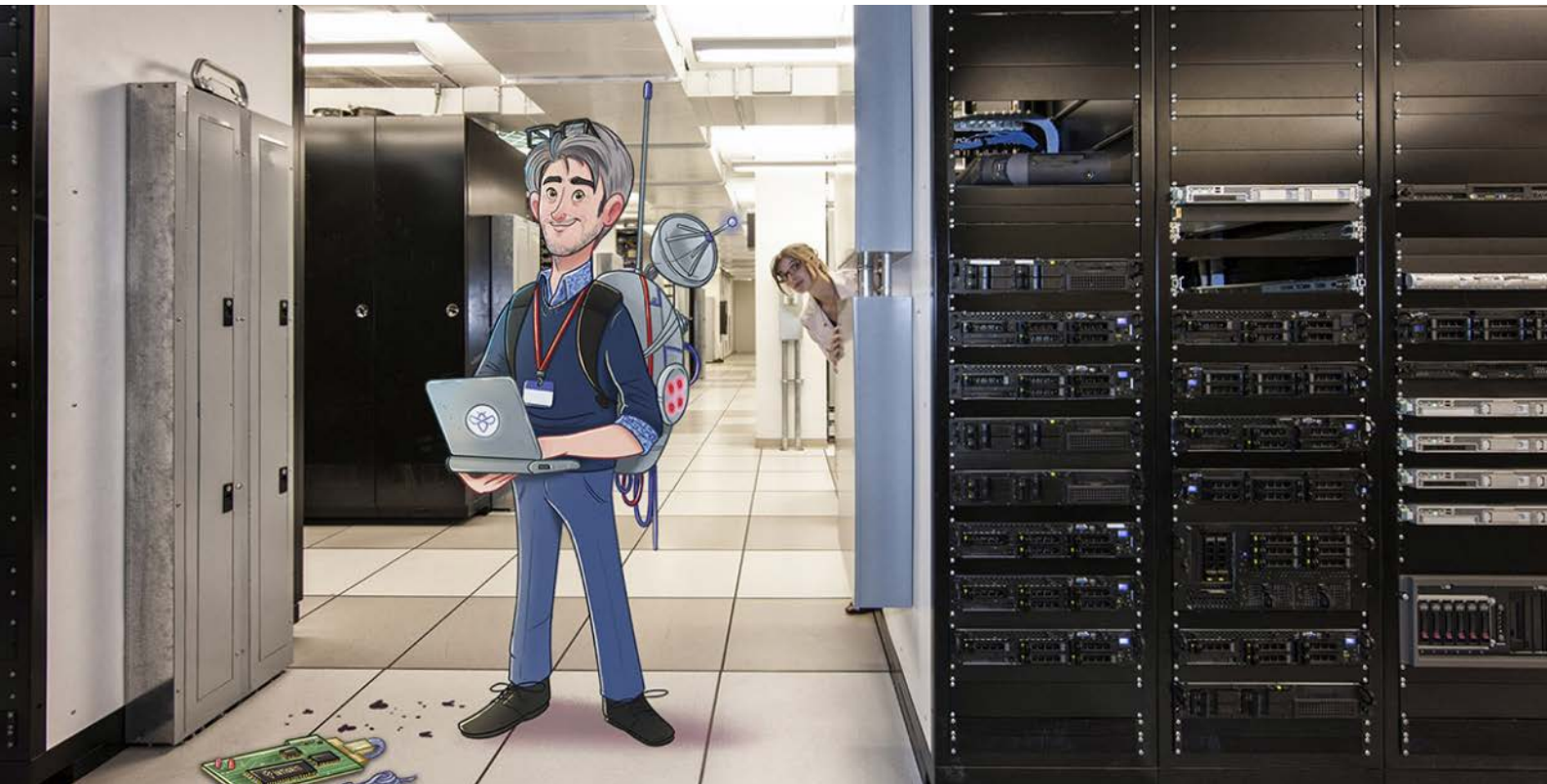
Pre-trained by Anthropic on a broad dataset of text.

This model has not been fine-tuned. Ember-v1

- Pre-trained on a large corpus of general text for robust embedding generation.
- Additional adaptation datasets or domain-specific finetuning are internal and confidential.

Duplicate prediction model (XGBoost)

- Labeled pairs of submissions (duplicate vs. nonduplicate).
- Internal data from past submissions, quality-checked for correctness.



Performance metrics

- **Summarization quality:** Evaluated by human reviewers and standard metrics (ROUGE, BLEU) on internal test sets.
- **Embedding quality:** Assessed by clustering purity and similarity benchmarks.
- **Duplicate Prediction:**
    1. **Accuracy:** 80% of correct duplicate vs. nonduplicate classifications. b.
    2. **Precision / Recall / F1-Score:** Reflects the trade-off between false positives and false negatives.
- Out of scope prediction:
    3. **Accuracy:** 75% of correct out of scope vs. non-out of scope classifications.
    4. **Precision / Recall / F1-Score:** Reflects the trade-off between false positives and false negatives.

Ethical considerations

Data privacy

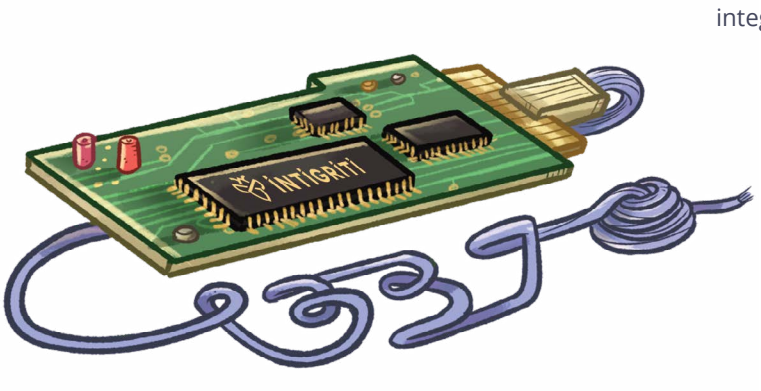
- All submission data is stored securely and processed following internal privacy guidelines.
- Aggregations and embeddings are used to minimize direct exposure of sensitive text content.

Bias and fairness

The model inherits biases from training data. Efforts are made to reduce and monitor biases through internal testing and human review.

Human oversight

Summaries and duplicate flags are used as decision aids; final decisions remain with qualified staff.



Limitations

Summarization fidelity

- The LLM may omit critical details or introduce minor inaccuracies.
- Human verification is recommended for high-stakes contexts.

Embedding generalizability

- Ember-v1 embeddings may vary in quality for highly specialized or niche text domains not well-represented in training data.

Duplicate prediction boundaries

- The duplicate prediction model depends on embedding accuracy; near-duplicates or paraphrased content might reduce performance.

Resource requirements

- LLM inference via Amazon Bedrock is cost-efficient but still requires monitoring for usage and optimization.

Ethical considerations

- **Scheduled retraining:** XGBoost duplicate prediction model is retrained every 6 months (or as needed) with new labeled data.
- **Monitoring:** We track performance on a rolling basis, with alerting systems for anomalies in performance metrics.
- **Versioning:** Each major update to the summarization, embedding, or duplicate prediction model is documented.

Contact and support

- **Technical support:** For questions, support, or incident reporting, please contact Support Team.
- **Documentation:** Further technical details and integration guides are available upon request.

About Intigriti

Global crowdsourced security provider trusted by the world's largest organizations

Intigriti's bug bounty platform provides continuous, real-world security testing to help companies protect their assets and their brand. Our community of ethical hackers challenge our customers' security against realistic threats: we test in precisely the same way malicious hackers do.

How vulnerability management works with Intigriti

1. Researcher tests and searches for a vulnerability
2. Researcher submits a report via Intigriti
3. Intigriti's triage begins communication with researcher
4. Intigriti's triage team applies quality assurance steps
5. In-scope, unique and well-written reports are submitted to client
6. Client accepts report, and payment is automatically processed

125.000+ researchers

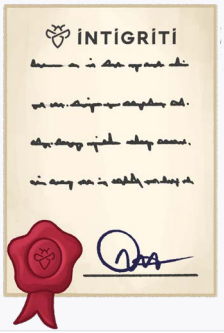
More than 125.000 security researchers use Intigriti to hunt for bugs — and we're growing!

400+ live bug bounty programs

Companies of all sizes, and across multiple industries, trust Intigriti to launch their bug bounty program.

GDPR compliant

We ensure compliance with the highest security and data security standards.



Strong global presence

Intigriti has a strong global presence. In terms of hacker pay-outs, the 10 best performing countries are globally represented in North America, Europe and Asia. In 2025, vulnerabilities were submitted from more than 180 countries.

A vulnerability reported and fixed is one less opportunity for a cybercriminal to exploit. Ready to talk about launching your first bug bounty program? We're here to help you launch successfully.

REQUEST A DEMO
intigriti.com/demo

VISIT THE WEBSITE
intigriti.com

GET IN TOUCH
hello@intigriti.com

YOU'RE IN GOOD COMPANY

onzo Dropbox Nestlé Personio NVIDIA intel Red Bull Microsoft UBISOFT