Technical Report – AI Safety Models POC

1. Introduction

The purpose of this Proof of Concept (POC) is to design a suite of AI Safety Models that enhance user safety in conversational AI platforms, such as chat applications.
The POC addresses key safety tasks:

1. **Abuse Language Detection** – Identifies harmful or inappropriate content in messages.
2. **Escalation Pattern Recognition** – Detects conversations that may escalate emotionally.
3. **Crisis Intervention** – Recognizes severe distress or self-harm indicators.
4. **Content Filtering** – Ensures age-appropriate messages for minor users.

This report documents the **design, implementation, evaluation, and technical considerations** of the project.

2. Dataset & Preprocessing

2.1 Dataset

The dataset used contains **user-generated messages** labeled for:

- `abuse_label` (1 = abusive, 0 = non-abusive)
- `escalation_label` (1 = escalating conversation, 0 = stable)
- `crisis_label` (1 = crisis indicator, 0 = normal)
- `age_flag` (1 = adult content, 0 = safe for minors)

Columns: `message, abuse_label, escalation_label, crisis_label, age_flag`

2.2 Preprocessing

Steps applied to clean the text:

1. Convert all text to lowercase.
2. Remove punctuation and special characters.
3. Strip extra whitespace.
4. (Optional) Stopwords removal was not applied to preserve context for multi-label detection.

This preprocessing ensures models receive **clean, normalized text input** for feature extraction.

3. Model Architectures

3.1 Logistic Regression (Baseline)

- **Vectorization**: TF-IDF (max features: 5000)
- **Classifier**: MultiOutputClassifier wrapping Logistic Regression (`max_iter=500`)
- **Purpose**: Quick baseline with interpretable predictions

3.2 XGBoost (Advanced)

- **Vectorization**: TF-IDF (same as Logistic Regression)
- **Classifier**: MultiOutputClassifier wrapping XGBoost (`n_estimators=100`, `learning_rate=0.1`)
- **Purpose**: Improve precision and recall, handle non-linear patterns

4. Training & Evaluation

4.1 Training

- Split: 80% training, 20% testing (`random_state=42`)
- Training scripts: `train.py` (Logistic Regression), `train_xgboost.py` (XGBoost)

4.2 Evaluation

- Metrics: **Precision, Recall, F1-score** per label
- Script: `evaluate.py`

Logistic Regression (`ml_model.pkl`)

Label	Precision	Recall	F1-score	Support
abuse_label	1.00	1.00	1.00	1000
escalation_label	1.00	1.00	1.00	1000
crisis_label	1.00	1.00	1.00	1000
age_flag	1.00	1.00	1.00	1000

- **Observation:** Perfect classification on the test set; indicates dataset separability.

XGBoost (`xgb_model.pkl`)

- Similar performance to Logistic Regression; suitable for production scaling if dataset grows in complexity.

Notes:

- Perfect scores indicate a clean or synthetic dataset.
- For real-world deployment, further testing on unseen and noisy data is required.

5. System Integration

The POC integrates all models into a **Streamlit app**:

- Accepts one message as input
- Preprocesses messages and applies TF-IDF vectorization
- Predicts **all four safety labels**
- Displays results in a user-friendly interface

Sample Input & Output:

Message	Abuse	Escalation	Crisis	Age Flag
"I don’t want to live anymore."	0	1	1	0
"You are so stupid!"	1	0	0	0

6. Ethical Considerations

- **Bias Mitigation**: Preprocessing preserves context; TF-IDF ensures no demographic assumptions.
- **Fairness**: Models trained on publicly available datasets without personally identifiable information (PII).
- **Safety**: Crisis detection can trigger alerts for human intervention, not autonomous actions.

7. Scalability & Future Work

- Upgrade to **transformer-based models** (BERT, RoBERTa) for improved text understanding
- Add **multilingual support** for global user bases
- Integrate **real-time streaming** for production-grade systems
- Implement **logging & monitoring** for continuous model evaluation
- Conduct **bias and fairness audits** across demographics

8. Conclusion

This POC demonstrates an **end-to-end AI Safety system** for conversational AI platforms. By combining Logistic Regression and XGBoost models with a Streamlit interface, the system:

- Detects abusive and unsafe content
- Recognizes escalation and crisis situations

- Provides age-appropriate filtering

This project illustrates both **technical feasibility** and **ethical design** for real-world AI safety applications.