PART 1: BUSINESS UNDERSTANDING

TriageLens

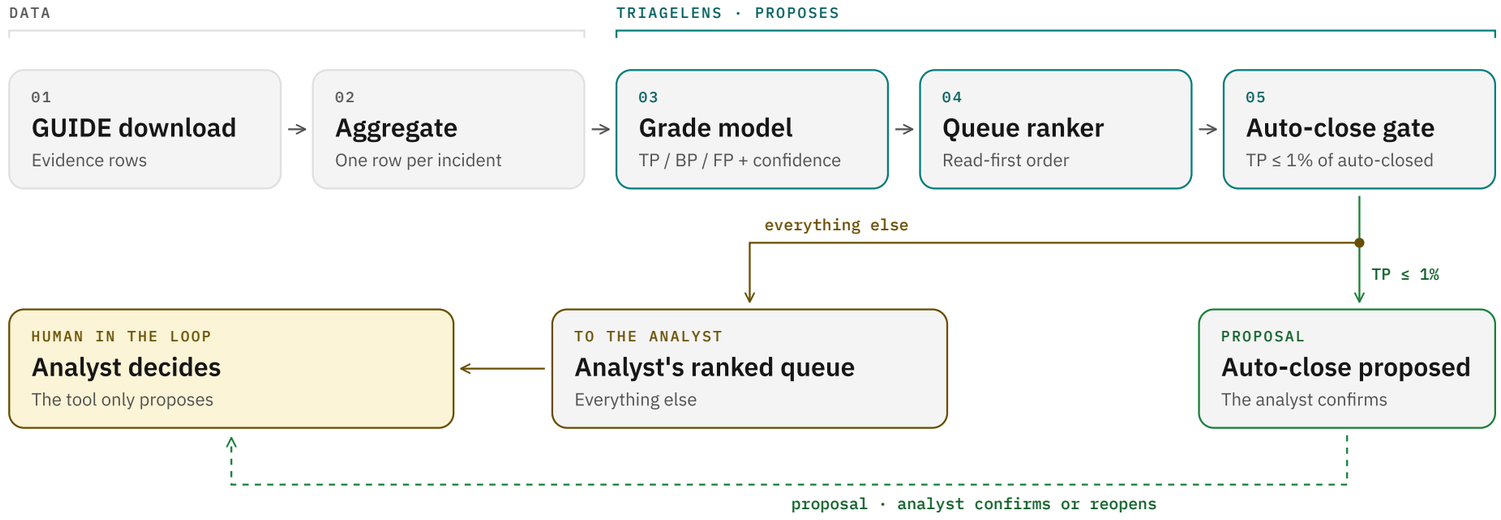
TriageLens is a queue assistant for a security operations centre (SOC), built and tested on Microsoft's GUIDE dataset.

⚠ THE PROBLEM

A tier-1 analyst faces a queue that is too long to read in full. Most incidents are false alarms or harmless, and about one in five is a real attack.

01 **Evidence rows** → 02 **One row per incident** → 03 **Graded** → 04 **Ranked in the queue** → 05 **Auto-close gate** → 06 **The analyst decides**

08 TriageLens flow



03 Objectives

1. **Grade** each incident as true positive (TP), benign positive (BP) or false positive (FP), with a confidence and the reasons behind it (detectors, categories, ATT&CK techniques, entity types).
2. **Rank** each organisation's queue so the incidents worth reading come first.
3. **Propose auto-closes** only where the chance of closing a real attack is held below a fixed limit, and send everything else to the analyst.
4. **Measure honestly** on unseen organisations and on newer data, next to the official split.
5. **Describe the groups**: cluster incidents and check whether outliers are more often attacks.

05 Questions the model answers

ON A SINGLE INCIDENT:

- Is it TP, BP or FP, and how confident is the model?
- Which features and ATT&CK techniques drove that grade?
- Can it be closed without review?

ON THE QUEUE:

- Which incidents should I read first, and how many of the top 10 match the experts' top 10?

ON THE EVALUATION:

- How well does it grade unseen organisations versus known ones?
- Does a model trained before 10 June 2024 work after it?
- Which organisation-detector pairs raise the most false alarms?
- Do incidents form groups, and are outliers more often real attacks?

06 Success criteria



04 Problem types

01

Grade

Multi-class classification · Binary classification with a cost-based threshold · Fine-tuned language model

KNN · linear SVM · RBF SVM (on a subsample) · logistic regression · class weights vs undersampling · Not a course algorithm: LoRA fine-tune of Qwen3-8B on Tinker vs zero-shot and the best scikit-learn model

Incident grade: TP, BP or FP

Real attack or not: TP vs not TP, with a probability for auto-close

Grade from text (extra): The grade, from incident text

02

Rank

Recommender system

Content-based · collaborative filtering (user-based, item-based, TruncatedSVD) on the organisation × detector false-alarm matrix · hybrid

Queue order: Which incidents to read first

03

Patterns

Regression · Clustering · Anomaly detection

Ridge regression · K-Means · CAH (at most 20,000 sampled incidents) · DBSCAN on PCA features · DBSCAN noise vs TP labels

False-alarm rate: FP share per organisation-detector pair

Incident groups: Groups of similar incidents

Unusual incidents: Are BP/FP outliers more often TP?

07 Error-cost table

ACTUAL GRADE ↓ / ACTION →	Auto-closed (not read)	Sent to analyst, low in queue	Sent to analyst, top of queue
True positive (real attack)	!! CRITICAL — The attack goes unseen and keeps running. This is the error the 1% limit controls.	! HIGH — The attack is seen late, after other work.	√ CORRECT
Benign positive (real but harmless activity)	√ CORRECT	· LOW — Some analyst time spent.	~ MEDIUM — Analyst time spent before real attacks lower in the queue.
False positive (detector misfired)	√ CORRECT	· LOW — Some analyst time spent.	~ MEDIUM — Analyst time spent and trust in the tool drops.

Severity is qualitative: we have no measured cost per incident, so we invent none.

06 · SUCCESS CRITERIA

CRITERION	TARGET	REFERENCE POINT
Grade quality on unseen organisations	Macro-F1 ≥ 0.60 on the 20% of organisations held out	Per-detector lookup: 0.43 unseen, 0.68 official split
Queue ranking	Precision@10 > 0.38 against the expert top-K, macro-averaged across organisation queues	Detector-rarity reference: 0.32 on the 111 held-out queues. Picking 10 incidents at random scores 0.07
Safe auto-close	Auto-close only at a threshold fixed on validation data where at most 1% of auto-closed incidents are TP	Everything above the threshold goes to the analyst
Drift	Macro-F1 training before 10 June 2024, testing after	Same model on the organisation split
Honest reporting	Unseen-organisation scores shown next to the official split	Official split as optimistic reference