What Good Looks Like

Worked Exemplar (Gold Standard)

> **Purpose of this exemplar**
> This example represents a **high-quality, defensible outcome** for the AI Assurance Capstone exercise.
> It is intentionally **not perfect or overly complex**. It demonstrates **judgement, restraint, and lifecycle thinking**.

Use Case Chosen

Customer Complaint Severity & Regulatory Risk Scoring (NLP)

1. Decision & Scope

Decision Being Supported

> Which customer complaints must be escalated within 24 hours due to regulatory, conduct, or reputational risk?

Decision Owner (Role)

> Head of Customer Experience (with Compliance concurrence)

What Changes If This Works

- High-risk complaints are escalated earlier
- Regulatory breaches are less likely
- Audit sampling becomes risk-based instead of random

If We Do Nothing

- Escalation relies on manual reading and subjective judgement
- High-risk complaints may be identified too late

- Inconsistent treatment across branches and channels

Why this is strong

- One clear decision
- Clear operational impact
- Clear downside of inaction

---

2. Minimum Defensible Use Case (MDUC)

MDUC Level Selected

> **Level 1 – ML scoring with human review**

Why We Are Not Starting at Level 2

- Complaint risk has direct regulatory consequences
- False positives and false negatives must be understood first
- Human judgement must remain explicit during early use

Stop Criteria

- Explainability is insufficient for reviewers
- False negatives exceed agreed tolerance
- Escalation rationale cannot be defended

Why this is strong

- Conservative starting point
- Clear justification
- Explicit stop conditions

---

3. Minimum Data Required

Data Element	Source System	Type	Historical Depth	Key Risk
Complaint text	CRM	Unstructured	12 months	Inconsistent wording
Product type	CRM	Structured	Current	Misclassification
Channel	CRM	Structured	Current	Missing values
Resolution outcome	CRM	Structured	12 months	Label inconsistency

Explicit Exclusions

- Social media data (too noisy for pilot)
- Call audio (phase 2 only)

Why this is strong

- Minimal but sufficient data
- Explicit exclusions (very strong signal of discipline)
- Clear data quality risks

---

4. ML Lifecycle Thinking

4.1 Problem Framing

- Scope limited to **severity prioritisation**, not sentiment analysis
- Output is a **risk score**, not an automated decision

4.2 Data Preparation

- Completeness checks by channel
- Manual review of labels used for training
- Bias review across product categories

4.3 Model Development

- Baseline: keyword rules
- Model: logistic regression (explainable)

- No complex embeddings or deep models in phase 1

4.4 Validation & Testing

- Precision and recall prioritised over overall accuracy
- False negatives analysed separately
- Segment testing by product and channel

4.5 Deployment & Use

- Risk score displayed to case handler
- Mandatory escalation reason captured
- Overrides logged and reviewable

4.6 Monitoring & Change

- Weekly drift checks on language patterns
- Monthly review of false negatives
- Kill-switch triggered by unexplained score shifts

Why this is strong - End-to-end lifecycle coverage
- No “magic” steps
- Monitoring treated as a control, not an afterthought

---

5. Controls & Evidence

Key Controls

- Human-in-the-loop escalation
- Explainability available for every score
- Override tracking and review

Evidence Expected

- Model documentation
- Validation metrics
- Override logs
- Monthly monitoring report

Explainability Approach

- Top contributing words and phrases shown to reviewers

Why this is strong - Evidence is concrete and reviewable
- Explainability is operational, not theoretical

---

6. Governance & Approvals

Stage	Approver	Evidence
Pilot start	CX + Compliance	Use case note
Continue pilot	Risk Committee	Validation report
Expand usage	Executive delegate	Monitoring results
Model change	Model governance forum	Change log

Accountability Question

Who signs if this goes wrong?

> Head of Customer Experience

Why this is strong - Clear ownership
- No shared ambiguity

---

7. Risk Judgement & Trade-Offs

Top Risks

1. False negatives on serious complaints
2. Bias by product or language
3. Reviewer over-reliance on scores

What We Explicitly Allow to Fail Early

> False positives — they slow us down but do not harm customers

What We Do Not Allow

> Silent misses or unexplained model behaviour

Why this is strong - Differentiates learning failures from control failures
- Demonstrates mature judgement

---

8. Final Recommendation

Proceed / Pause / Stop?

> **Proceed with pilot**

Rationale

- Clear value
- Defensible controls
- Conservative automation posture

Why This Example Scores Highly

This approach: - Starts **small but meaningful**
- Keeps humans **explicitly accountable**
- Thinks across the **entire ML lifecycle**
- Treats monitoring as assurance
- Makes trade-offs explicit

> *It is not the smartest AI.*
> *It is the safest AI that can learn quickly.*

Facilitator Note

Use this exemplar: - **After** peer scoring, not before
- To explain why some approaches scored higher
- To reinforce that “good” is achievable and practical

This is the level of thinking we are aiming for —
defensible, disciplined, and scalable.