

baswe.Ai Engineer Accelerator™

Safety & Ethics Checkpoint

Format:

Written Assessment + Stakeholder Loom + Technical Exam

Time investment:

4–6 hours

Deliverable:

AI Risk & Responsibility Assessment for the provided scenario



baswe's

AiEngineerAccelerator

Pre-checkpoint checklist

Before you start, confirm all of the following.

Materials provided

- Meridian Health company brief (provided in Safety & Ethics Checkpoint START)
- Stakeholder map (provided in Safety & Ethics Checkpoint START)
- Technical specification: ClaimsAssist v1 (provided in Safety & Ethics Checkpoint START)

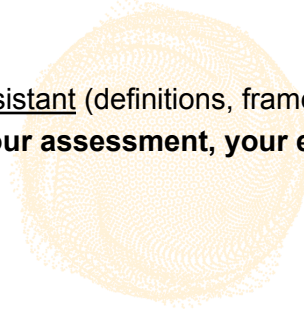
Tools you will need

- Google Docs, Microsoft Word, or any document editor that exports PDF
- Loom account (free tier is fine for a single recording)
- A text editor for the exam-answers.md file

On using LLMs to draft your work

You may use an LLM as a research assistant (definitions, framework lookups, industry references).

You may NOT use an LLM to draft your assessment, your exam answers, or your Loom script.



baswe's

AiEngineerAccelerator

Part 1 Written Assessment

Scenario

You have been hired as the first AI Safety & Governance Engineer at Meridian Health, a regional health insurance company with ~2,400 employees serving 850,000 members across 4 states. Meridian's ML team has built ClaimsAssist v1, an LLM-powered triage assistant that reads incoming claims, summarizes them, and recommends one of three actions to human claims adjusters: approve, deny, or escalate for further review.

The ML team wants to launch ClaimsAssist in production in 6 weeks. The executive team has asked you to perform an independent risk and responsibility assessment before they approve the launch. Your assessment will be reviewed by the Chief Medical Officer, Chief Compliance Officer, CISO, Head of Claims, and CEO.

Read the three provided documents carefully before starting your assessment. Your job is to identify what is good, what is risky, and what would need to change for you to recommend launch.

Deliverable

Produce a single document titled ai-risk-assessment.pdf (or .docx) with the following sections, in order. Section page targets are minimums, longer is fine if it earns its length, but padding without substance hurts.

1. Executive summary (1 page)

- Your overall recommendation: Launch as proposed / Launch with required changes / Do not launch
- The three most important reasons for your recommendation
- The single biggest risk if your recommendation is not followed

2. Responsible AI principles applied (1–2 pages)

- Identify which Responsible AI principles are most relevant to this deployment (fairness, accountability, transparency, privacy, safety, human oversight, etc.)
- For each principle you name, explain specifically how it applies to ClaimsAssist
- Identify at least two principles the current proposal does not adequately address

3. Data ethics review (1–2 pages)

- Review the training data described in the tech spec. Identify any data ethics concerns regarding provenance, consent, representativeness, or potential for bias on protected classes (age, race, disability status, gender, geography)
- Flag at least one specific population at elevated risk of being harmed by this system, with reasoning

- Recommend specific changes to data collection, labeling, or evaluation that would address the concerns you identified

4. Security & privacy assessment (1–2 pages)

- Identify the top 3 security risks specific to this LLM-based deployment (think about PHI handling, prompt injection, model extraction, training data leakage, etc.)
- For each risk: describe the attack or failure scenario, the realistic impact at Meridian's scale, and the mitigation you would require before launch
- Identify which risks the current tech spec already addresses well, if any

5. Governance & accountability (1 page)

- Recommend who at Meridian should own ClaimsAssist post-launch, not just the ML team but the cross-functional governance structure (who signs off on changes, who reviews incidents, who has authority to pause the system)
- Define what "human in the loop" should specifically mean for ClaimsAssist. The tech spec says human adjusters review every recommendation, is that enough? What does meaningful human oversight require here?
- Recommend at least three governance controls that should be in place before launch (e.g., model card requirements, change review process, incident response trigger criteria)

6. Launch conditions (1 page)

- If your recommendation is "launch with required changes," list the specific, verifiable conditions that must be met before launch. Be concrete: not "improve data quality" but "re-audit training data for demographic representativeness across the four-state service area; resample if any group is underrepresented by >20%"
- Order conditions by must-have vs strongly-recommended
- Define what you will measure in the first 90 days post-launch to know whether your conditions worked

Part 2 Stakeholder Loom

Format

- Single Loom recording, 15–25 minutes, screen share with face cam on
- Public or unlisted Loom link

Scenario

You are presenting your assessment to the Meridian executive team. Assume your written document has been pre-read (you are not reading it aloud), you are walking them through the substance and answering the questions they would ask. The audience: CEO, Chief Medical Officer, Chief Compliance Officer, CISO, Head of Claims. None of them are AI engineers. They are smart, busy, and skeptical.

Required structure

1. Recommendation up front (2–3 minutes)

- Open with your recommendation. Don't bury it.
- Give the three most important reasons. Plain language, no jargon.
- Frame the stakes: what happens if Meridian launches as proposed, and what happens if they delay.

2. Walk through your highest-stakes finding (4–6 minutes)

- Pick the single most important risk you identified. Walk through it like you are explaining to the CMO or Chief Compliance Officer.
- Use the scenario: who is affected, what happens to them, why this risk is real for Meridian specifically.
- Recommend the specific control or change that addresses it.

3. Address the most likely pushback (4–6 minutes)

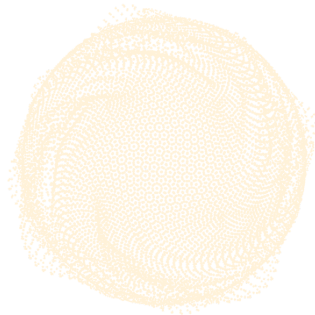
- Anticipate where the executive team will push back. "The ML team says we already handle bias by reviewing every recommendation." "Six weeks is the timeline our board promised." "Compliance signed off on the data sources."
- Walk through how you respond. Strong responses acknowledge the legitimate concern, then redirect to what the data or principle actually requires.
- This is judgment under pressure, not righteousness. The goal is to land your recommendation, not win an argument.

4. Launch path or kill criteria (3–5 minutes)

- If your recommendation is launch-with-changes: walk through the must-haves. Be concrete about what "done" looks like for each.
- If your recommendation is do-not-launch: walk through what would need to change for you to revisit the decision in 90 days.
- Either way: name the 90-day post-launch metrics you will watch and the threshold that would trigger pulling the system.

5. Open Q&A simulation (2–5 minutes)

- Simulate three executive questions to yourself. Pick questions that are hard, not soft.
- Answer each in 60–90 seconds. The goal is to demonstrate calm, structured thinking under live questioning.
- Examples of hard questions: "Has any peer company successfully deployed something like this?" "What if we just don't launch the deny recommendations and only do approves and escalations?" "How much does each of your required changes cost us in time and money?"



baswe's

AiEngineerAccelerator

Part 3 Technical Exam

Format

- Open-book. Open-internet.
- Do **NOT** use an LLM to generate your answers. Do **NOT** Google the entire question. Cheating on these only cheats yourself.
- Time-boxed: 1–2 hours. We do not enforce the timer, but if you spend 4+ hours on this section, that is a signal you might not have understood the concepts deeply enough.
- Written answers in a single Markdown file (exam-answers.md)

Section A: Responsible AI Foundations (8 points)

Q1 (3 pts)

You are deploying a customer-facing AI feature at a B2C SaaS company. List the Responsible AI principles you would prioritize in order, and explain why the order is what it is for THIS context.

Q2 (3 pts)

Explain the difference between transparency and explainability in AI systems. Give an example where a system can be transparent without being explainable, and one where it can be explainable without being fully transparent. Which matters more for end-user trust, and why?

Q3 (2 pts)

What is the difference between AI ethics and AI governance? Why do organizations need both, and what specifically does governance do that an ethics framework alone cannot?

Section B: Data Ethics (7 points)

Q4 (3 pts)

Your team is training a model on customer support transcripts to power an automated reply suggestion system. Walk through the data ethics concerns you would raise before training starts. Cover consent, data minimization, retention, and bias on protected attributes. Be specific about what you would and would not allow.

Q5 (2 pts)

Define representational bias and measurement bias in training data. Give a concrete example of each that could occur in a healthcare or financial services dataset.

Q6 (2 pts)

A vendor offers you a labeled dataset that they describe as "compliant with all applicable regulations." List the specific questions you would ask before agreeing to use it. What would cause you to walk away?

Section C: AI Security & Risk (10 points)

Q7 (3 pts)

Explain prompt injection in an LLM-powered application. Distinguish direct prompt injection from indirect prompt injection. For each, give a realistic attack scenario in an enterprise deployment and the primary mitigation you would implement.

Q8 (3 pts)

Your LLM-powered application processes sensitive customer data. List the top 4 security risks specific to LLM deployments (not generic web app risks). For each: describe the risk, the realistic impact, and one specific control you would put in place.

Q9 (2 pts)

Explain the concept of model risk management. How does it differ from traditional software risk management, and what specifically does it require that software risk management does not?

Q10 (2 pts)

You are asked to choose between deploying a smaller, fine-tuned open-source model on your own infrastructure versus calling a third-party LLM API for a high-sensitivity use case. Walk through the security and governance tradeoffs of each option, and name the scenarios where each is the right call.

Section D: Governance & Frameworks (5 points)

Q11 (3 pts)

The EU AI Act introduces a risk-based classification for AI systems. Briefly describe the categories. For each category, give an example AI system that would fall under it and explain the primary compliance obligation it triggers.

Q12 (2 pts)

What is an AI model card, what is an AI system card, and how do they differ? When would your organization produce each one, and who is the audience for each?