

A strong interview. Keep this level.

AI Engineer · City of Riverton

The City of Riverton is seeking an AI Engineer to build reliable, privacy-focused applications based on language models that assist residents and staff. The role involves taking these applications from prototype to production while ensuring explainability and quality.

2.4 / 3

average quality

4 / 9

requirements strongly covered in the interview

60%

answers with a result (R)

60%

Concrete examples

Statistics

STAR

S

60%

T

60%

A

60%

R

60%

Concrete examples in 60% of answers; results quantified in 40%.

Quality

level 0

0

level 1

1

level 2

1

level 3

3

Answer quality (0 = does not answer, 3 = concrete, structured, with result)

Requirement coverage

Requirement	In Documents	In Interview
A completed degree at bachelor level or higher in computer science, AI, data science or similar. required — you have this, but did not bring it up	strong	none
At least 3 years of experience with Python in production environments. required — you have this, but did not bring it up	strong	none
Demonstrable experience with applications based on language models (for example RAG). required	partial	strong
Experience with CI/CD, containers (Docker) and version control (Git). required	strong	strong
Good written and spoken English. required	strong	strong

REQUIREMENT	IN DOCUMENTS	IN INTERVIEW
Experience evaluating LLM output and building guardrails. nice to have	strong	strong
Knowledge of privacy law (GDPR) and working with personal data. nice to have — you have this, but did not bring it up	strong	none
Experience with open-source or self-hosted language models. nice to have — you have this, but did not bring it up	strong	none
Experience in the public sector or with complex stakeholders. nice to have — you have this, but did not bring it up	partial	none

Practice plan

- 1 Rework your answer about building trust with non-technical stakeholders into one concrete STAR story instead of general principles. Pick a specific moment (like the demo you gave in turn:d86f87bce07740b09c1c6588234a8e26) and lead with that from the start.**

Your initial answer on this topic scored highest quality but stayed abstract, and the interviewer had to explicitly ask for a concrete example before you delivered a strong story. Leading with specifics avoids that follow-up entirely.

Tip: Tell it as STAR

- 2 Explicitly mention your degree and your years of Python production experience somewhere in your answers, even briefly, since they weren't brought up at all.**

Both requirements are flagged as weak points with high severity (0.7) because they were never addressed, even though they're likely on your CV and relevant to eligibility for the role.

Tip: Address a gap honestly, then bridge

- 3 When telling stories, push harder on quantified results (numbers, percentages, time saved) since only 40% of your answers included quantification.**

Your CI/CD story (turn:f69fd4b984524c3d969483c5a2f2dc5a) and others have strong structure but often stop short of a measurable outcome, which would make the impact more convincing.

Tip: Tell it as STAR

- 4 Keep using a brief pause before answering tricky or unexpected questions, especially follow-ups that ask you to get more specific.**

The follow-up question about policy advisers shows interviewers will push for concrete detail; a short pause helps you retrieve a real example instead of defaulting to general statements.

Tip: Pause before you answer

- 5 Continue closing interviews with sharp, role-relevant questions like the ones you asked about evaluation ownership and go-live criteria.**

Your closing question scored a perfect 2.0 and showed genuine insight into production quality risks, which is a strong habit to keep reinforcing.

Tip: Close with a question of your own

Focus for next time

- Experience for 'A completed degree at bachelor level or higher in computer...' was not brought up (severity 0.7; weak)
- Experience for 'At least 3 years of experience with Python in production...' was not brought up (severity 0.7; weak)

Upload the report PDF with your next run to practise these again.

Answers

1

Intake account manager · Open the interview and understand motivation for the role and organisation

2.8 / 3

What draws you to this AI Engineer role at the City of Riverton, and how do you see yourself contributing to public-sector applications?

What draws me most is that the AI directly helps residents. At Harbour Logistics I built an assistant for planners, and the best part wasn't the technology. It was seeing people do their work faster and with more confidence. At a city, that impact is even more direct: if a staff member can help a resident quickly and correctly, the resident feels it right away.

I also really like that reliability, privacy and explainability come first here, because that's already how I work. Every answer from our assistant has to cite its source, and no release goes live without passing our evaluation set. I'd bring exactly that discipline. My public-sector experience is limited to an internship at a water authority, so I'd lean on policy advisers for the context. The part where you make AI measurable and trustworthy, that I can bring from day one.

quality: 2.8

✓ S

✓ T

✓ A

✓ R

✓ specific

✗ quantified

CV: consistent · Claude

✗ overclaim · Claude

hedging: 0.2

✗ resp_4 · Claude

STRONG

- You clearly connect your Harbour Logistics work to resident impact: 'it was seeing people do their work faster and with more confidence' your answer · CV role1 b1
- Good honesty about the limits of your public-sector experience, showing self-awareness your answer

ADD

- You didn't mention your GDPR/privacy work with a privacy officer at DataWorks, which is directly relevant to a public-sector employer's privacy concerns CV role2 b3
- Your meetup organising on open-source AI could show community engagement and communication skills relevant to explaining choices to non-technical stakeholders CV p8

EXPLORE

- Do you have an example, even from the water authority internship, where you had to translate a policy or operational question into a technical solution for non-technical stakeholders? CV role3 b1
- Can you describe a time you explained a technical choice (like a guardrail or model limitation) to someone without a technical background? your answer

BETTER OUTLINE

- Open with motivation: AI that helps residents directly, echoing the planner assistant impact you saw at Harbour Logistics your answer · CV role1 b1
- Add your GDPR/privacy officer collaboration at DataWorks as evidence you already work within data protection constraints relevant to government work CV role2 b3
- Acknowledge your public-sector experience is limited to your water authority internship, and that you'd rely on policy advisers for context your answer · CV role3 b1
- Close by connecting your ability to explain technical choices clearly, as shown by organising your open-source AI meetup, to translating between technical and non-technical colleagues CV p2 · CV p8

2

Technical lead · Explore hands-on experience building and productionising LLM-based applications

3.0 / 3

You mention taking applications to production with Docker and GitHub Actions on Kubernetes. Can you walk me through a specific instance where you handled versioning or CI/CD for an LLM-based application?

Sure. The clearest example is how we release changes to the customs assistant at Harbour Logistics. Early on, a change to the prompt or the retrieval setup could quietly make answers worse, and we'd only find out when a planner complained. I wanted every change to be checked before it reached users.

So I set up a GitHub Actions pipeline that builds the Docker image and then runs our evaluation set of 200 verified questions against it. If the share of wrong answers goes up, the release is blocked. Only when the set passes does the image get deployed to the Kubernetes cluster. To be honest, my Kubernetes knowledge is fairly basic; the platform team runs the cluster, and I work with them on the deployment side.

The result: no release goes live without passing the set, and that's a big part of how we brought wrong answers down from 18% to 6%.

quality: 3.0 ✓ S ✓ T ✓ A ✓ R ✓ specific ✓ quantified CV: consistent ✗ overclaim hedging: 0.2
✗ eis_1 ✓ eis_3 ✓ eis_4 ✓ eis_5 ✓ resp_1 · uncertain ✓ resp_2 ✗ wens_2

STRONG

- You gave a concrete, specific example: a GitHub Actions pipeline that builds the Docker image and gates deployment on an evaluation set of 200 questions before it reaches the Kubernetes cluster. your answer
- Nice honesty about acknowledging your Kubernetes knowledge is basic and that you collaborate with the platform team rather than overclaiming. your answer

ADD

- You didn't mention the guardrails you added (citations required, refuse when no source found) as part of the quality checks tied to this release process. CV role1 b3
- You didn't connect this to the concrete impact on planners, like the drop in average search time from 12 to 3 minutes. CV p4

EXPLORE

- Do you have an example of how you decided what threshold or rollback strategy to use when the wrong-answer rate went up during a release? your answer

- Do you have an example of versioning specific components separately, like the prompt versus the retrieval index, so you could tell which change caused a regression? *your answer*

BETTER OUTLINE

- Set the problem: unchecked changes to prompts or retrieval could silently degrade answers for the customs assistant at Harbour Logistics. *your answer*
- Describe the pipeline: GitHub Actions builds the Docker image, then runs the 200-question evaluation set before deployment to Kubernetes. *your answer · CV role1 b2 · CV role1 b4*
- Mention the guardrails (citations required, refuse when no source found) as another layer of quality control in the release process. *CV role1 b3*
- Be clear about ownership: you work with the platform team on Kubernetes since your knowledge there is more basic. *your answer · CV role1 b4*
- Close with impact: wrong answers dropped from 18% to 6%, and search time per question fell from 12 to 3 minutes. *your answer · CV p4*

3

Hiring manager · Probe quality measurement, guardrails against hallucinations, and translating needs from non-technical stakeholders

1.0 / 3

Tell me about a time you had to prove to non-technical stakeholders that an AI system's answers could be trusted, and how you approached measuring or improving that trust.

I think trust mostly comes from good communication. I always try to explain things clearly and involve stakeholders early, so they understand what the system does and what it doesn't do. I also try to be transparent about the limitations, because people trust you more when you're honest with them. And I make sure to keep them updated along the way, so there are no surprises at the end. In general, I've found that if you take people seriously and listen to their concerns, they're much more open to new technology.

quality: 1.0

X S

X T · Claude

X A · Claude

X R

X specific

X quantified

CV: not_in_cv · Claude

X overclaim

hedging: 1.1 · uncertain

X eis_2

X resp_3

X resp_4 · Claude

X wens_1

X wens_3

X wens_4

STRONG

- You emphasized being transparent about limitations and involving stakeholders early, which is a good foundation for building trust *your answer*

ADD

- You didn't mention the evaluation set of 200 questions you built to measure answer quality and track improvement over iterations *CV role1 b2*
- You didn't mention the concrete guardrails you added—requiring citations and refusing to answer when no source is found—which directly address hallucination risk *CV role1 b3*
- You didn't mention your work with planners at Harbour Logistics, where the RAG assistant reduced search time from 12 to 3 minutes, which would be a strong concrete example for non-technical stakeholders *CV role1 b1 · CV p4*

EXPLORE

- Do you have an example of how you specifically measured or quantified trust or accuracy for non-technical stakeholders, beyond general communication practices? *your answer*

- Did you present the evaluation results (e.g., the drop from 18% to 6% wrong answers) directly to non-technical stakeholders like planners or managers? CV role1 b2

BETTER OUTLINE

- Start with a concrete case: built a RAG assistant for planners answering customs questions, used by many stakeholders CV role1 b1
- Explain how you measured quality with an evaluation set of 200 verified questions, showing wrong answers dropped from 18% to 6% over three iterations CV role1 b2
- Describe the guardrails you introduced—mandatory citations and refusal when no source is found—to prevent hallucinations CV role1 b3
- Connect this to trust-building with stakeholders: showing measurable improvement and clear limitations helped planners rely on the system, cutting search time from 12 to 3 minutes CV role1 b1 · CV p4 · your answer

4

Hiring manager · Probe quality measurement, guardrails against hallucinations, and translating needs from non-technical stakeholders

2.9 / 3

You mentioned communicating with non-technical stakeholders to build trust, but your response focused on general principles rather than a specific instance. Can you describe a concrete situation where you had to explain a limitation of an AI system to a skeptical policy adviser and what the outcome was?

Sure. I haven't worked with policy advisers yet, but the closest situation was at Harbour Logistics, when I demoed the customs assistant to the management team. One of them was quite skeptical. During the demo the assistant answered one question with "I don't know," and he asked why we'd roll out something that can't answer everything.

I explained in plain words that this was deliberate. The assistant only answers when it can find the rule in our documents, and it always shows the source. If it can't find one, it says so, because a confident wrong answer on customs rules costs far more than a planner spending a few minutes checking. I backed that up with our evaluation results: wrong answers had come down from 18% to 6%, largely because of that refusal behaviour.

After that he was on board, and the "I don't know" behaviour stayed in as a feature.

quality: 2.9 ✓ S ✓ T ✓ A ✓ R ✓ specific ✓ quantified CV: consistent ✗ overclaim hedging: 0.1

✗ eis_2 ✓ resp_3 ✓ resp_4 ✓ wens_1 ✗ wens_3 ✗ wens_4 · Claude

STRONG

- Good use of a concrete example instead of general principles: the skeptical manager, the 'I don't know' moment, and the outcome that it stayed as a feature. your answer
- You backed up your explanation with a clear metric (wrong answers down from 18% to 6%), which is convincing evidence for a skeptical audience. your answer · CV role1 b2

ADD

- You didn't mention the citation requirement guardrail explicitly, which is part of what convinced the manager the system was trustworthy. CV role1 b3
- You didn't mention your experience working with a privacy officer under GDPR, which is another example of translating technical/compliance constraints for a non-technical stakeholder. CV role2 b3

EXPLORE

- Do you have an example where you explained an AI limitation specifically to someone in a policy or government role, rather than a management team, since the question asked about a policy adviser? your answer
- Was there any resistance or follow-up questions from the manager after you showed the evaluation numbers, and how did you handle pushback in the moment? your answer

BETTER OUTLINE

- Set the scene: skeptical manager during the customs assistant demo questioned why it said 'I don't know'. your answer
- Explain the guardrail design: answers only given with a source citation, refusal when no rule is found. CV role1 b3
- Give the reasoning: a wrong customs answer costs more than a planner double-checking. your answer
- Support with data: error rate dropped from 18% to 6% over iterations, largely due to refusal behavior. CV role1 b2 · your answer
- Outcome: manager was convinced, and the refusal behavior stayed as a permanent feature. your answer

5

Intake account manager · Close the interview and invite candidate questions

2.0 / 3

That covers what I wanted to ask — what questions do you have for us about the role or the team?

Thanks, I have two.

First: how do you decide when an AI application is good enough to put in front of residents? I'm curious whether you already have fixed criteria for that, or whether it's weighed per application.

And second: who in the team owns the evaluation of model output once something is live? In my experience, that's the moment quality can quietly slip if nobody is clearly responsible for it.

quality: 2.0 · Claude

X S · Claude

X T

X A · Claude

X R

X specific

X quantified

CV: consistent

X overclaim

hedging: 0.1

ADD

- You could have grounded your second question in your own experience of setting up an evaluation set and tracking wrong-answer rates over iterations, which would show this isn't just a theoretical concern for you. *CV role1 b2*

EXPLORE

- Do you have a concrete example of a time when evaluation ownership was unclear on one of your projects and quality started to slip, which you could reference if asked to elaborate on your second question? *your answer*

BETTER OUTLINE

- Ask how they decide when an AI application is ready for residents — fixed criteria or case-by-case. *your answer*
- Ask who owns ongoing evaluation of model output once live, noting that in your experience this is where quality can slip without clear ownership. *your answer*
- Tie it back to your guardrails work (citations required, refuse when no source found) as an example of building in accountability from the start. *CV role1 b3*

Models: Jev jev-1.13.0 · Claude claude-sonnet-5 · local qwen3.6:35b-a3b. This PDF contains report.json as an attachment; upload it for your next run.