May 2026

Savanta

Role Specification: Senior AI Engineer

Classified: Private

Location: UK (Remote, onsite 1-2 days per month)

Business Area: Technology

Reports to: Engineering Team Lead

About the Role:

Senior AI Engineer | Tech | UK

Our roadmap includes ambitious delivery across products and services where AI capabilities are becoming a core part of the user experience and internal platform. This role is focused on designing, building, and shipping production-grade AI applications using C#/.NET and Azure, with a particular emphasis on LLM-powered workflows, evaluation, and reliable deployment.

The impact of your work should be visible quickly. Projects will often be delivered in short iterations, with a focus on getting usable capability into production, learning from real behaviour, and improving quality through structured testing and evaluation.

Senior AI Engineers operate within agile teams, typically alongside other software engineers, product managers, QA specialists, UI/UX, and business stakeholders. They work closely with others to translate business problems into robust AI-enabled solutions, balancing speed, safety, cost, and quality. This role also helps raise AI engineering capability across the wider technology team through coaching, technical leadership, and practical standards.

Our tech stack includes:

- .NET / C#, including APIs and backend services
- Microsoft Azure, including cloud-native services for AI workloads
- LLM application patterns, including prompt orchestration, retrieval pipelines, and evaluation
- OpenAI-compatible APIs / SDKs and related tooling
- GitHub / Azure DevOps CI/CD pipelines
- Data platforms and storage services, including structured and unstructured data handling

Technical:

- Design and implement LLM-powered systems using appropriate grounding techniques, such as RAG, search tools, knowledge graphs, etc in production.
- Implement AI-powered features such as chat, summarisation, extraction, classification, workflow copilots, and retrieval-backed assistants.
- Work with OpenAI-compatible protocols and SDKs, and understand the practical trade-offs between models, latency, quality, cost, and context limits.
- Design and implement RAG-based solutions, including chunking, retrieval strategies, grounding, citation patterns, and failure-mode handling.
- Demonstrate working knowledge of vector embeddings and how they are used in semantic search and retrieval systems.
- Make sound engineering decisions around prompt design, tool use, structured outputs, orchestration, guardrails, and fallback behaviour.
- Build AI systems that are observable, debuggable, and maintainable, with clear handling for errors, retries, and degraded model behaviour.
- Apply software engineering fundamentals to AI systems, including clean architecture, design patterns, version control, code review, and secure API integration.
- Influence testing by ensuring AI features are developed with both automated software tests and LLM-specific evaluation approaches in mind from the outset.

Delivery:

- Break down complex AI proposals into clear stories, tasks, milestones, and acceptance criteria.

- Translate ambiguous business needs into practical AI solutions that are testable, supportable, and valuable to users.
- Consider users and UX when designing AI behaviour, including how confidence, uncertainty, citations, escalation, and fallback states should appear in product experiences.
- Think about quality early: ensure features are validated through unit tests, integration tests, and evals rather than relying on ad hoc prompt checks alone.
- Define and run LLM evaluations to verify behaviour such as factuality, grounding, instruction following, schema adherence, safety, and task success.
- Work iteratively, shipping improvements regularly and using evidence and user feedback to refine prompts, retrieval, orchestration, and system behaviour.

Data / ML Understanding:

- Show numerical confidence and comfort discussing metrics, thresholds, experiment results, and trade-offs.
- Demonstrate familiarity with core ML and data concepts such as embeddings, similarity, clustering, principal component analysis, feature spaces, and basic linear algebra.
- Work effectively with data required for AI systems, including dataset preparation and evaluation sets.
- Partner with other engineers, or data specialists to improve model performance using structured diagnosis rather than intuition alone.

Communication:

- Communicate technical AI concepts clearly to both engineering and non-technical stakeholders.
- Explain why an AI system is behaving a certain way, what its limits are, and how confidence in its output is established.
- Facilitate productive discussions around model choice, architecture, risk, evaluation, and operational readiness.
- Write clear documentation including runbooks, technical decisions, eval approaches, prompt/change logs, and support guidance.
- Communicate changes that may impact critical systems, data flows, quality levels, or user-facing AI behaviour.

People & Team:

- Work with the Team Lead and Head of Engineering to build and develop a high performing team, fostering a culture of continuous innovation, and held accountable for delivering high quality work. Deputising for Team Lead when necessary.
- Mentor software engineers and other team members in AI engineering patterns, testing approaches, and responsible delivery practices.
- Support the delivery of our recruitment strategy, including interviewing prospective members of the team.
- Build trusted relationships with stakeholders and act as a credible advisor on where AI is and is not the right solution.
- Create progression plans and stretching objectives for junior staff, ensuring they deliver against them, managing performance where required and taking corrective action (e.g. Performance Improvement Plans).
- Identify & deal effectively with underperformance among junior staff, implementing and running Performance Improvement Plans in a timely manner.

Personal Development / Profile at Savanta:

- Embody company values and meet performance expectations.
- Take ownership of personal development in a fast-moving AI landscape.
- Stay current with evolving tooling, SDKs, model capabilities, and engineering practices, while remaining grounded in delivery value rather than hype.
- Help shape pragmatic standards for how AI solutions are built, tested, monitored, and maintained across the business.
- Build a visible profile across the team and wider business.
- Promote global best practices, innovation, and collaborative ways of working.
- Contribute positively beyond core responsibilities through initiatives like Career Management, knowledge sharing, standards, and continuous improvement.
- Actively own personal development goals and work closely with managers to achieve them.

About You:

Essential Experience:

- Strong ability in C# and .NET building production applications and services.
- Writing and using unit tests, using any language and test framework.
- Microsoft Azure and deploying cloud-based applications in production environments.
- Hands-on experience implementing LLM-driven applications beyond experimentation or demos.
- Able to explain and apply LLM evals as a core quality discipline for AI applications.
- Familiarity with OpenAI-compatible APIs, protocols, or SDKs, and the practical considerations of integrating them into applications.
- Clear understanding of RAG and the role of vector embeddings in retrieval-based AI systems.
- Strong understanding of software testing fundamentals, including unit, integration, and end-to-end testing, and how these differ from AI-specific evaluation.
- Numerical confidence and working familiarity with concepts such as similarity, clustering, PCA, and basic linear algebra.
- Full product delivery mindset focused on user value, reliability, and measurable quality.
- Comfortable working with ambiguity, rapid iteration, and shifting priorities.
- Strong communication skills and ability to work effectively with technical and non-technical stakeholders.

Desirable Experience:

- Using or building RAG systems in production, including retrieval tuning and relevance evaluation.
- Vector databases or Azure-native retrieval/search services.
- Working with Snowflake or other data warehouses.
- Working with SQL Server.
- Using Azure AI services, model hosting platforms, or related observability tooling.
- Prompt management, structured output validation, tool calling, or agent-style workflows.
- Defining evaluation datasets, benchmark suites, and regression testing pipelines for LLM applications.
- Exposure to broader ML concepts or workflows beyond LLMs.
- Mentoring other engineers and helping establish AI engineering standards.
- Security, privacy, and governance considerations for enterprise AI solutions.