Jon Ellis-Jones

Software Development Engineer · Conversational AI & Distributed Systems

London, UK / jon@ellis-jones.org / ellis-jones.org / GitHub / LinkedIn

Software engineer with two decades of experience building, shipping and scaling products used by millions of customers. Technical lead at Amazon for the past decade on conversational and generative AI, first for Alexa and now for Alexa for Shopping. Turn ambiguous product ideas into low-latency, cost-efficient production systems through designs that span organisations, and raise the engineering bar of the teams around me. BSc in Computer Science from the University of Leicester; member of the British Computer Society.

Experience

Senior Software Development Engineer (SDE III)

Amazon / London, UK / Nov 2016 – Present

Technical lead on Agentic AI for Alexa for Shopping (formerly Rufus), Amazon's generative AI shopping assistant, used by more than 400 million customers in the last 12 months, with customer use of its agentic features growing more than eightfold since the start of 2026. Previously led natural language understanding and contextual product Q&A work for Alexa, Amazon's voice assistant. Day to day, balance inference latency against answer quality and cost, validating changes through LLM evaluations, guardrails and A/B testing.

- Won first place at an internal hackathon with Scheduled Actions in 2025, then led its high-level design, technical strategy and delivery through to launch in 2026. It lets customers hand recurring and event-driven shopping tasks to the assistant: price drops, new releases, back-in-stock alerts and customer-defined schedules. Built with Kotlin, Amazon ECS and Amazon Bedrock.
- Owned critical-path engineering for Prime Day 2026 Deal Guides and Deal Alerts built on Scheduled Actions, including retry semantics for agent tools that change customer state and hardening the platform for spiky, high-volume traffic.
- Led the technical design of Auto Buy, an agentic feature that places orders automatically when a customer's target price is reached, coordinating partner teams across Alexa and ordering. Advised senior leadership on sequencing, and simplified the design by reusing the existing offline ordering pipeline.
- Steered engineering for bringing Alexa for Shopping inline to the product detail page, home page and cart. Authored the key technical requirements that shaped organisation-wide goals and prioritisation, and advised front-end leadership on widget-rendering strategy.
- Led engineering for the London team that productionised the first customer experience for Rufus, then unannounced, on Amazon's product detail pages in 2023, proving Amazon LLMs with real customers ahead of the app launch. Integrated with Rufus's RAG and tool-use architecture, shaping how it served shoppers on the page, and streamed gRPC output as server-rendered HTML at tens of thousands of requests per second.
- Authored the scaling strategy and two-year technical vision for Suggested Q&A ("Customers also ask Alexa") on Amazon Search, moving widget storage from Amazon OpenSearch Service to Amazon DynamoDB, replacing scheduled orchestration with event-driven pipelines and parallelising ingestion. Re-engineered its offline ranking pipelines for semantic and fuzzy keyword matching at scale, using multi-armed bandits to optimise question selection from customer engagement. Built with Scala, Apache Spark, AWS Glue and AWS Step Functions.
- Set the technical direction for contextual product Q&A on Alexa, solving cross-domain intent classification and entity resolution, including products mentioned earlier in, or implied by, a conversation. Designed tiered intent-classification workflows that combine pattern matching with machine-learned models at sub-40 ms p99 latency and low cost. Built with Java, Python, Amazon SageMaker and AWS Inferentia.
- Mentored engineers across the organisation, gave feedback on promotion documents, ran design reviews, interviewed engineers and managers, and represented the team at internal and external hiring events.

Software Development Engineer

Microsoft / London, UK & Tallinn, Estonia / Apr 2012 – Nov 2016

Worked across Skype's commerce teams in London, Tallinn and Prague, bringing structure and definition to thousands of calling and subscription products, enabling payments in geographies around the world, and driving technical innovation and improvement.

- Integrated third-party payment service providers, including PayPal, GlobalCollect, Yandex.Money and PayByCard, into the Skype ecosystem through a unified adapter middleware, managing customer payments, refunds, entitlements and reconciliation.

- Defined the commerce product model and its RESTful APIs for products, subscriptions, rates, promotions, offers and bundles, backed by versioned, transactional database reads and writes. Built with PHP, PostgreSQL (PL/pgSQL), Maven and Chef.
- Shipped Skype's Calling Rates experience, a low-latency single-page application written in JavaScript (ECMAScript) and PHPTAL with proprietary two-way data binding, helping customers worldwide choose the right calling product or subscription.
- Scaled the redemption service for Skype's Xbox One launch offer (six months of free group video calls and world calling minutes for Xbox Live Gold members in 13 launch territories), absorbing the holiday-season surge of gamers unboxing new consoles. Built with C#, ASP.NET Web API, Microsoft Azure and DocumentDB (now Azure Cosmos DB).

Senior Software Developer

Creative Insight / Birmingham, UK / Mar 2010 – Apr 2012

- Led a team of developers at a design agency, modernising how the studio worked by introducing version control, code reuse through an HMVC framework, and structured project scheduling.
- Developed an internal project management tool for tracking tasks and time that automated routine paperwork.

Web Developer

Kaweb / Birmingham, UK / Nov 2006 – Mar 2010

- Co-wrote a modular car insurance quotation engine that produced over 1.1 million quotes a month at peak, served through an XML API to price comparison sites including Confused.com, GoCompare and MoneySupermarket.
- Engineered the agency's fully featured, bespoke e-commerce platform: templatable and extensible, with SagePay payments, automated SEO for Google Shopping, and bulk product imports. Deployed for several high-value retail clients.

Volunteer

Volunteer

CoderDojo (Whitechapel E1) / Oct 2016

CoderDojo is a global movement of free, volunteer-led, community-based programming clubs for young people. At a Dojo, young people aged 7 to 17 learn to code, build websites, apps, programs and games, and explore technology in an informal, creative environment. In addition to learning to code, attendees meet like-minded people and are exposed to the possibilities of technology.

Education

Bachelor of Science (Hons), Computer Science

University of Leicester / 2002 – 2006 / Score: 2.1

Courses Dissertation: Social Networking and Degrees of Separation – The Mathematics of Social Networks (First)

Skills

Languages

Kotlin / Java / Scala / TypeScript / JavaScript / Python / PHP / C# / SQL

AI & Machine Learning

Agentic AI / Retrieval-augmented generation (RAG) / Agentic tool use / Multi-agent orchestration / Context optimisation / LLM evaluation / Guardrails / A/B testing / Inference latency vs quality and cost / Natural language understanding / Intent classification / Named entity recognition / Entity resolution / Multi-armed bandits / Amazon Bedrock / Amazon SageMaker / AWS Inferentia

Distributed Systems & AWS

Amazon ECS / AWS Lambda / AWS Step Functions / Amazon DynamoDB / AWS CloudFormation / gRPC / HTTP/2 / RESTful APIs

Data

Amazon OpenSearch Service (Elasticsearch) / PostgreSQL / Apache Spark / AWS Glue / Amazon Redshift

Platforms & Tools

AWS / Microsoft Azure / .NET (ASP.NET Web API) / Maven / Chef / CI/CD / Event-driven architecture