Adi Singhal

+1(929)4003809 • adis@nyu.edu • linkedin.com/in/adi-singhal • github.com/adityaarunsinghal

"Extensive experience in software dev and research science;
Most recently focused on RAG and Agentic AI at Amazon."

EDUCATION

B.A. Data Science, New York University**,** GPA 3.98 Aug 2018 – May 2022

• Advanced Statistics for Research • Computer Systems Org (Assembly) • Reinforcement Learning • Natural Language Processing • Responsible Data Science (Ethical AI) • High Performance Computing (HPC) • Causal Inference

B.A. Honors. Psychology, New York University**,** GPA 4.0 Aug 2018 – May 2022

• Lab in Industrial/Organizational Psychology • Fundamentals of Decision-Making • Human Motivation & Volition
• Computational Cognitive Neuroscience (also M.S. Data Science)

WORK EXPERIENCE

Software Engineer (SDE2), Amazon Web Services, New York July 2022 - Present

- 3+ years at Amazon driving transformative AI/ML initiatives across healthcare and enterprise platforms
- Primary dev-tooling expert on a team of 40+ engineers. Aug 2025– I automated on-call with a custom agent using Strands SDK and MCP on our weekly ticket queue of 200+ Sev-2 trouble tickets. This *homebrew*, named OncallCompanion, has now evolved; used by 650+ engineers, reducing countless hours in Mean Time To Resolution. Went on to build a volunteers community of 25+ engineers globally
- Built and marketed a Slack Helper AI Agent named BedrockBot as a side project; now used by 25k+ global users within Amazon and serves as the primary way to connect team-context to Slack channels
- Managing new model integrations in our product=*AWS Bedrock Knowledge Bases*; Got a head-start on all releases from Meta (Llama), Anthropic (Claude) and other providers; Made an automated Retrieval-Augmented-Generation benchmarking pipeline for end-to-end performance + regression testing of these new models, reducing model-support time from 5 days to 4 hours (hailed the Day 0 Effort). Led architectural design and implementation of next-gen reasoning models with 60-minute processing capabilities and 128k token context support, positioning Amazon Q at the forefront of enterprise AI
- Over the years did lots of diverse cross-team work; created the health.amazon.com subdomain, integrated One Medical sign-ups with Amazon and founded the Health Services *Applied Science* Club. Pioneered Amazon Health's first conversational AI system (Project Macaw) in 3 weeks, achieved 57% accuracy improvement over baseline in diagnosis prediction and presenting directly to SVP leadership, Neil Lindsay
- Architected enterprise-grade AWS PrivateLink integration enabling secure private connectivity for Fortune 500 compliance requirements with 99.99% availability SLA for AWS's Managed Vector Store
- Led frontend architecture for Prime healthcare benefits integration (Project Hornbill) reviewed by current CEO Andy Jassy, ensuring 100% accessibility compliance for all users
- Mentored 3 interns with 100% return offer rate, established best practices adopted across multiple teams

NLP Engineer, Kasisto Inc, New York May 2021 - 2022

- Boosted company's intent detection model by 7% F1 (near-human) in 3 client applications: JPMC, TD Bank, WestPac
- Asked by leads to report directly to the CTO of the company, Sasha Caskey, to improve model accuracy
- Wrote Python programs to push 125+ experiments to the company's HPC and presented the new pipeline and findings to the entire company; Maintained meticulous documentation of all experiments, visualized findings and wrote reports for the Data Science team
- Built a scalable framework for automatic model improvements by folding in gold standard data (patent pending)

PROJECTS AND RESEARCH

NYU Agentic AI Workshop for Alumni Fall 2025

- Conducted a 4-week Agentic AI workshop with MCP workshop for NYU graduate students and alumni, with a classroom style, python-notebook based interactive tutorial with a final show and tell
- Built production-ready Model Context Protocol (MCP) server for intelligent news aggregation and personalized content discovery using vector similarity search and multi-source integration
- Designed an educational template for MCP server development with extensible architecture supporting new sources, handlers, and AI-powered content analysis features. Collaborated with one of the core contributors of MCP's Python SDK and released the final "learning" module to the public on HackerNews (YC)
- Website: http://adityasinghal.com/agentic-ai-workshop
- GitHub: adityaarunsinghal/agentic-ai-workshop-2025

Meta's AI Lab Researchers – Computational Simulation of Perceived Difficulty in Games Oct 2021 - May 2022

- Under guidance from FAIR researchers Drs. Brenden Lake, Todd Gureckis, and Guy Davidson, I created a novel 2D physics environment in Unity to study human game difficulty perception and playability modeling
- Implemented a Domain Specific Language (DSL) for formal game representation, enabling automated game setup, scoring, and procedural content generation research
- Developed OpenAI Gym-compatible framework enabling both human interaction and RL agent training, with comprehensive state tracking of ball trajectories, object collisions, and player actions across 64,000+ automated gameplay scenarios; Conducted human subjects research with 12 participants across 5 game categories
- Discovered strong negative correlation (-0.7) between pre-play human difficulty ratings and computational playability scores, demonstrating humans' intuitive ability to assess game difficulty from visual inspection alone
- AI-driven playability metrics; procedural content generation; human-AI collaboration in game design
- GitHub: adityaarunsinghal/temporal-goals-in-games

IBM – Domain Adaptation Spring 2020 – Summer 2021

- Developed a hierarchical LLM fine-tuning approach using domain-specific Named Entity Recognition (NER) as auxiliary task to improve RoBERTa performance on zero-shot question answering across Movies, News, Biomedical, and COVID-19 domains
- Outperformed baseline RoBERTa models in 3 of 4 tested domains (F1 scores: Movies 67.99 vs 67.09, Bio 58.86 vs 57.97, COVID 42.66 vs 42.05) without access to target domain QA training data
- Demonstrated that domain-specific supervised auxiliary tasks transfer more effectively than Domain Adaptive Pretraining (DAPT) for question answering, with DAPT underperforming baselines in 3/4 domains tested
- Catastrophic forgetting patterns; transfer learning efficacy; sequential fine-tuning strategies; low-resource NLP
- GitHub: adityaarunsinghal/Domain-Adaptation

Facebook's Parl.ai – Empathetic Chatbots Fall 2020 - Present

- Created synthetic dataset "Empathic Conversations in the Wild" with 30k automatically curated empathetic conversations from Reddit, achieving 5.7% precision filtering using fine-tuned RoBERTa classifier
- Fine-tuned Facebook's BlenderBot (9.4B parameters) on HPC using Singularity containers, discovering empathy-relevance trade-offs where Reddit-trained model generated more informative but less empathetic responses (p<0.05)
- Validated automated scoring pipeline with high human-annotation correlation (r=0.623, p<4e-9) and identified key limitations of forum-based data for empathy training
- GitHub: empathic-conversations-chatbot: A new Parlai Task

ACHIEVEMENTS

- Amazon **"Deliver Results" award** for making a company-wide Slack AI Agent 2024
- **Published academic papers** in collaboration with Dr. Lyle Ungar (UPenn) and IBM 2022
- Inducted to **Phi Beta Kappa** as a Junior 2021
- Awarded the 2021 **NYU Student Leadership Award** for community engagement 2021
- **Dean's Research Fund** recipient for Affective Computing research 2020
