# Andrey Shor

Andrusha.shor@gmail.com · linkedin.com/in/andrey-shor · github.com/Andrusha-Shor

*Research Engineer translating AI research into production systems at enterprise scale — 3 publications, 4 platforms, 15,000+ users.*

## EDUCATION

**M.S. Artificial Intelligence** | University of Texas at Austin — *2025 – 2027 (expected)*

**B.S. Computer Science** | Purdue University • *Minor: Psychology* — *2018 – 2022*

## WORK EXPERIENCE

**Research Engineer** | Burns & McDonnell — *May 2024 – Present*

**Focus:** *Agent Orchestration, Retrieval-Augmented Generation, AI Safety, Interpretability & Reasoning*

- Drove adoption of AI-native engineering practices across Burns & McDonnell's Technology Solutions Group, influencing the team's technical strategy toward research-informed AI development across all three platforms below.

### *Axiom — Harness Engineering Platform*

**Tech:** Python, FastAPI, Vertex AI, Cloud Run, ADK, MCP, A2A

- **Architected a multi-agent orchestration platform** supporting any Vertex AI model and coding harness (OpenCode, Claude Code) with full lineage and provenance tracking.
- **Designed a context engineering layer with MCPs, Skills, and A2A protocols** so admins define guidance structures and non-technical end users interact through a brief agent → planner agent → specialized agent teams pipeline.
- **Platform generates documents, images, data files, and deployable web applications** for proof-of-concept workflows across a 50–100 user beta group.

### *Document Processing Platform & SDK*

**Tech:** Python, FastAPI, Gemini API, Cloud Run, Docker

- **Achieved 85–90% extraction accuracy across 60 business rules, reducing manual review time by ~90% and scaling to 1,500–3,000 users.** Built a microservice for certificate-of-insurance-to-subcontract comparisons with rule-specific preprocessing and structured prompts routed to Gemini.
- **Applied RLVR's strict verifier paradigm to design deterministic verification functions** sourced from SMEs, enabling automated pass/fail validation of each extraction against ground-truth business rules before human approval.
- **Extracted core logic into a reusable document processing SDK** rolled out to citizen developers enterprise-wide. Users provide a prompt, JSON schema, and optional verification function; the SDK handles preprocessing, LLM calls, and validation.

### *Experience IQ — Enterprise NL-to-SQL Agent*

**Tech:** Python, GCP Spanner, BigQuery, Vertex AI Agent Engine, Gemini Enterprise, Looker, Cloud Workflows, DSPy

- **Improved system quality from 72% to 80%, reliability from 75% to 94%, and reduced experience search time by ~85%; deployed to 15,000 employees.** Built an LLM agent translating natural language into SQL over a 110-table Spanner database via intent routing, dynamic schema retrieval, and SME-curated few-shot curriculum.
- **Ran systematic evaluation cycles: categorized failure modes by root cause via trace analysis,** hypothesized upstream fixes, and tested interventions across data dictionary refinements and prompt-level guardrails—improving SQL formatting accuracy from 80% to 95% over four iterations.
- **Automated curriculum curation with an LLM-as-a-Judge evaluator** trained via grid search over K-shot configurations and GEPA prompt optimization (DSPy). Deployed on a daily batch cadence, maintaining >70% judge-SME agreement. Presenting at Google Cloud Next 2026.
- **Gave product owners real-time signal on agent drift** via a Looker LLMOps dashboard surfacing query volume, guardrail pass rates, reliability, token usage, and per-query debug traces.

**Data Engineer** | 1898 & Co. — *Jan 2023 – Apr 2024*

**Focus:** *ETL Pipelines, Data Visualization, Azure Cloud*

- **Reduced manual data integration time by ~70% for utility-sector clients** by building a Flask REST API deployed on Azure Serverless Functions to automate data migration from SharePoint to Oracle APEX.
- **Delivered project health and scheduling dashboards to 10–15 Project Managers** by building PowerBI visualizations of financial, scheduling, and operational KPIs, enabling data-driven resource allocation and client reporting.

## RESEARCH EXPERIENCE

**Independent Researcher** | University of Virginia (Aidong Zhang Group) *Jan 2024 – Present*

- **Published at IEEE Big Data 2024.** Finetuned language models via PEFT and prompt-tuning for Social Determinants of Health extraction in MIMIC-IV. Augmented dataset with disease-specific clinical literature to improve performance.

**Undergrad Research Assistant** | Purdue University (Benotman Lab) *Aug 2022 – Dec 2022*

- **Published at IEEE FIE 2024.** Built a clustering tool for ER diagram submissions using t-SNE and silhouette scoring to surface patterns and improve grading consistency.

**Undergrad Research Assistant** | Cal State Northridge *May 2021 – Dec 2022*

- Designed grid-search experiments on 2,000+ randomized DFA to uncover structural trends in random language complexity under Meng-che Ho.

**Undergrad Research Assistant** | Purdue University (Goldwasser Lab) *May 2021 – Dec 2021*

- **Published at NAACL 2022.** Analyzed temporal trends in 2.7M COVID-19 tweets and validated BERT sentiment classifiers for a vaccine debate framework.

## PUBLICATIONS

L. Gong, **A. Shor**, A. Zhang, K. Jha. "Context-specific feature augmentation for improving SDoH extraction." *IEEE Big Data*, 2024.

S. Thadani, **A. Shor**, S. Ahn, L. Gong, A. Alawini, H. Benotman. "Clustering ER diagrams for feedback quality in large DB courses." *IEEE FIE*, 2024.

M. L. Pacheco, T. Islam, M. Mahajan, **A. Shor**, M. Yin, L. Ungar, D. Goldwasser. "A holistic framework for analyzing the COVID-19 vaccine debate." *NAACL-HLT*, 2022, pp. 5821–5839.

H. Viswanath, M. A. Rahman, A. Vyas, **A. Shor**, et al. "Neural operator: Is data all you need to model the world?" *arXiv:2301.13331*, 2023.

## INVITED TALKS

**Speaker**, Google Cloud Next 2026 — Experience IQ: Enterprise NL-to-SQL Agent *2026*

**Speaker**, Google Dev Days — AI Applications *Dec 2025*

**Guest Lecturer**, Purdue University CS473 — Agentic Systems in Production *Nov 2025*

**Guest Lecturer**, Purdue University CS473 — Retrieval Augmented Generation *Nov 2024*

## TEACHING & MENTORING

**Mentored 2 interns** at Burns & McDonnell (2023–2024): Seth Deegan (M.S., Purdue) and Amogh Garuda Dwajan (B.Sc., UConn).

**Undergraduate TA**, Purdue — CS390 Deep Learning, CS252 Systems Programming, ECE264 Advanced C *2020 – 2022*

## SKILLS

**AI/NLP:** LLM Agents, RAG, Controlled Generation, NL-to-SQL, LLM Evaluation, Prompt Optimization (DSPy/GEPA), Multi-Agent Orchestration (ADK, MCP, A2A), Interpretability, Reasoning

**Platforms & Infrastructure:** Vertex AI (Agent Engine, Gemini API, Model Garden, Embeddings), GCP (Spanner, BigQuery, Cloud Run, Cloud Workflows), Azure (Serverless Functions), Looker, Docker, CI/CD

**Frameworks:** PyTorch, TensorFlow, HuggingFace Transformers, FastAPI, Flask, Scikit-Learn, NLTK

**Languages:** Python, SQL (GoogleSQL, PostgreSQL), JavaScript, MATLAB, C