HAYDEN ARNOLD

Scottsdale, AZ • (+1) 480-297-6607 • haydenaa7@gmail.com
haydenarnold.github.io • linkedin.com/in/hayden-arnold • github.com/haydenarnold

Education

Master of Science in Computational Social Science June 2026
Korea Advanced Institute of Science and Technology (KAIST) *Daejeon, South Korea*

- Cumulative GPA: 4.2/4.3
- Advised by Prof. Taegyoon Kim

Bachelor of Science in Computer Science May 2024
University of Arizona *Tucson, Arizona*

- Cumulative GPA: 3.83/4, *magna cum laude* distinction

Honors Bachelor of Arts in Political Science, Foreign Affairs Emphasis May 2024
University of Arizona *Tucson, Arizona*

- Minors in Korean and Spanish Language
- Summer 2023 study abroad at Yonsei University in Seoul, South Korea
- Cumulative GPA: 3.83/4, *magna cum laude* distinction
- Published honors thesis advised by Prof. Yotam Shmargad

Work Experience

AI Evaluation Expert (Contract) 03/2024 – Present
DataAnnotation.tech *Remote*

- Designed and executed structured **LLM evaluation protocols** across 300+ programming and reasoning prompts, performing pairwise model comparison and qualitative error analysis
- Engineered **adversarial and edge-case prompts** to systematically surface failure modes in code generation, logical reasoning, and instruction following
- Developed detailed evaluation rubrics and delivered structured feedback identifying hallucinations, unsafe outputs, and reasoning inconsistencies
- Contributed to alignment and safety evaluation of frontier LLMs through structured red teaming, adversarial prompt design, and behavioral failure analysis

Data Analyst Intern 01/2024 – 05/2024
Pima County Democratic Party *Tucson, Arizona*

- Designed automated data pipelines to aggregate and clean multi-source voter data across **266 precincts** and 10+ issue categories
- Developed predictive models to forecast election outcomes, achieving **80% accuracy** against certified results
- Performed feature engineering and cross-validation to improve model robustness and reduce overfitting
- Communicated statistical findings to non-technical stakeholders through structured reports and data visualizations

Senior Undergraduate Teaching Assistant 01/2023 – 05/2024
University of Arizona Department of Computer Science *Tucson, Arizona*

- Led and coordinated a team of **10+ TAs** supporting a **500+ student** CS course, ensuring consistent grading standards and instructional quality
- Redesigned course materials and assessment rubrics to improve clarity and reduce grading variance across sections
- Implemented workflow improvements to streamline assignment distribution, feedback cycles, and staff communication

Undergraduate Teaching Assistant 08/2021 – 12/2022
University of Arizona Department of Computer Science *Tucson, Arizona*

- Instructed and mentored 20+ students per semester in Python-based programming coursework and algorithmic problem solving
- Led supplemental instruction sessions for 100+ students, focusing on debugging strategies and computational reasoning
- Provided structured feedback on code quality, correctness, and edge-case handling

Publications

Arnold, H. (co-author), Nguyen, Q.M., & Kim, T. (2026). Large Language Models are Not Always Faithful in Explaining Their Political Decision Making. Under review at *EMNLP 2026*.

- Extends concept of **faithfulness** from alignment research to the political decision making context
- Operationalizes and compares faithfulness in three distinct evaluation settings: news credibility, source framing, and hate speech detection
- Evaluates consistency under **manipulated chain-of-thought (CoT) and personas**

Arnold, H., Kim, S., & Kim, T. (2026). How Does Toxicity Emerge in Political Livestreams? Large-Scale Evidence from Twitch. Under review at *Political Communication*.

- Collected and processed **116M Twitch chat messages**; identified 471 political livestreamers via fine-tuned **RoBERTa** classifier
- Fine-tuned transformer models for large-scale toxicity scoring and stream-level interaction modeling
- Applied **System GMM** to estimate temporal and causal drivers of toxicity dynamics

Arnold, H. (co-author), Choi, J., Kim, T., & Jang, J. (2026). Evidence-Based Language Increases Persuasion in Online Debates: Insights from 1.6M Reddit Discussions (2013–2024). Under review at *Nature Human Behavior*.

- Analyzed **1.6M Reddit discussions** to quantify evidence- vs intuition-based persuasion signals
- Built semantic embedding pipeline (**Word2Vec**) with cosine similarity–based feature extraction
- Applied **BERTopic** and logistic regression to model topic-level heterogeneity and persuasion outcomes

Yu, X., **Arnold, H.**, Su, B., & Blanco, E. (2025). Measuring and Forecasting Conversation Incivility. In *Proceedings of International AAAI Conference on Web & Social Media (ICWSM)*.

- Operationalized incivility as a supervised prediction task using behavior-aware conversational features
- Implemented temporal forecasting models to predict future incivility from early-stage dynamics
- Demonstrated performance gains over content-only baselines using standard classification metrics

Arnold, H. & Shmargad Y. (2024). Statistical Trends in Online Anti-Feminism and the 2022 South Korean Presidential Election. *University of Arizona W.A. Franke Honors College*.

- Collected **1,321 videos** and **41,541 comments** via YouTube Data API; integrated **Perspective API** for multi-dimensional toxicity scoring
- Conducted time-series regression analysis to quantify temporal sentiment shifts across six incivility dimensions

Presentations

KAIST x Yonsei University 1st Inaugural Conference on Computational Social Science (2025)

- Presented *How Does Toxicity Emerge in Political Livestreams? Large-Scale Evidence from Twitch.*

Projects

Arnold, H. PoliNet: Political Actor Intelligence Network (2026)

- Built an end-to-end **political intelligence platform** that scrapes live news, extracts entities and relationships, and models a dynamic **political actor knowledge graph** updated in real time
- Engineered a full NLP pipeline using **spaCy NER** and **HuggingFace REBEL** for relation extraction, feeding structured triples into a **NetworkX + PyTorch Geometric** graph
- Trained **GraphSAGE and Heterogeneous Graph Transformer (HGT)** models with a link prediction objective to learn latent political relationships and generate actor embeddings
- Developed a **FastAPI backend** and **Next.js dashboard** featuring a natural language query engine (**FAISS** similarity search), actor profiles, finance explorer, and community detection

Arnold, H., Tsurho, K. FraudAI: Graph Neural Networks for Credit Card Fraud Detection (2025)

- Built fraud detection pipeline on **1.85M transactions** (0.52% fraud) with 57 engineered features and temporal train/val/test splits
- Constructed heterogeneous bipartite graph (1,692 nodes, 1.85M edges) and implemented **GCN, GraphSAGE, GAT** in PyTorch Geometric
- Benchmarked against Random Forest and XGBoost; ensembles outperformed GNNs (**F1=0.373 vs. 0.126**), revealing limitations under extreme imbalance (1:176)

Arnold, H. Law School Admissions Calculator (2025) — GitHub Link | Website Link

- Developed a predictive web application **forecasting U.S. law school admissions outcomes** using historical admissions datasets
- Engineered features and trained classification models using **Python, NumPy, Pandas, and Scikit-learn**
- Built and deployed a **Flask backend** and **interactive frontend on Vercel**, enabling user inputs and real-time predictions
- Performed model evaluation and calibration using cross-validation and accuracy/error metrics for improved reliability

Arnold, H. U.S. Presidential Election Forecasting Model (2023) — GitHub Link

- Constructed an end-to-end ML pipeline to **predict 2024 U.S. presidential election results** from historical demographic and voting data
- Cleaned and preprocessed multi-source datasets using **Pandas and NumPy** to build robust training features
- Trained and compared multiple predictive models (e.g., logistic regression, random forest) using **Scikit-learn**
- Visualized results and insights using **Matplotlib** to communicate performance and underlying voting trends

Skills

Concepts: Supervised Fine-Tuning, Model Alignment, Behavioral Benchmarking, Adversarial Prompting, Red Teaming, Failure Mode Analysis, Representation Probing, Activation Steering, Model Reliability & Safety Evaluation, Regression Analysis, Statistical Inference & Hypothesis Testing, Causal Inference
Languages: Python, R, SQL, Java, C, JavaScript, HTML/CSS
ML/NLP: PyTorch, TensorFlow, Scikit-learn
Data & Tools: Pandas, NumPy, Matplotlib, Seaborn

Language Certifications

TOPIK (한국어능력시험): 5급 (유창) — Held 2025/2/22, 문서확인번호: 1834-9334-2108-6281