Dylan Bouchard, PhD

Boston, MA — (207) 695-7013 — dbouchard92@gmail.com — LinkedIn — GitHub

SUMMARY

Applied Research Scientist with 8+ years of experience in AI/ML. Specializing in turning AI safety research into production tools, with five first-author publications, including TMLR and JMLR. Built and open-sourced toolkits adopted by 300+ internal data scientists and thousands of external users.

EXPERIENCE

CVS Health, Principal Applied Scientist *01/2022 - Present*

Technical Leadership Overview

- Spearheaded the company's **first AI research program**. Authored five first-author papers and four research-backed AI safety toolkits (UQLM, LangFair, red teaming, ML fairness).
- Open-sourced **UQLM** and **LangFair**, CVS Health's most widely adopted open-source libraries based on GitHub stars and PyPI downloads. Recruited and led an internal team of 10 maintainers.
- Rolled out AI safety toolkits to **300+ data scientists** across CVS Health. Achieved internal policy alignment and replaced **$2–3M/year** in vendor spend across the program.
- Partnered with **C-suite** leaders and secured buy-in to publish research and open-source toolkits.
- Recognized as the **enterprise SME** for AI safety. Delivered trainings for **1000+ colleagues** on implementation and best practices.
- **Promoted** from lead-level IC (01/2022-03/2024) to principal-level IC (03/2024-present).

LLM Uncertainty Quantification (UQ) & Hallucination Detection

- Developed **UQLM**, a UQ-based hallucination detection toolkit. Released as open-source Python package, earning **1K+ stars**, 30K downloads in 7 months, and recognition from LangChain. Published (first-author) software paper in **JMLR**.
- Designed a **novel hallucination detection algorithm** using a tunable UQ ensemble, achieving **state-of-the-art** performance across 6 benchmarks. Published (first-author) **TMLR** paper.
- Led a first-author comparative study on UQ for long-form LLM outputs, introducing a taxonomy for claim decomposition, scoring, and aggregation that unifies and extends prior methods.
- Developed novel UQ methods for **code generation** based on code-adapted semantic entropy, achieving **state-of-the-art** performance in predicting code correctness on Python and SQL.

LLM Bias & Fairness

- Developed **LangFair**, an open-source Python package for LLM bias/fairness (**35K+ downloads**; 250+ GitHub stars). Integrated into **LangChain ecosystem**, expanding its industry adoption.
- Published (first-author) LangFair software paper in **Journal of Open Source Software**.
- Pioneered a **novel framework** for defining LLM bias and fairness assessments, introducing several new metrics. Solo-authored research paper on the framework (arXiv preprint).
- Led collaboration with Coalition for Health AI (CHAI) leadership to incorporate LangFair framework and toolkit into **CHAI's Responsible AI best practices**.

LLM Red Teaming

- Developed internal Python library of **adversarial prompting** techniques and defenses (response filtering, PII detection, prompt injection detection), reducing manual effort.
- Identified **high-risk vulnerabilities** in multiple LLM applications pre-deployment.

Machine Learning Fairness

- Developed internal Python library for **streamlined ML fairness**. Provided pre-packaged SQL queries to fetch protected attribute data, customized model bias testing, and 17 bias mitigators.
- Operationalized fairness testing for **100+** models and guided bias mitigation for **15+** models.
- Created three **novel ML fairness mitigators**: disparate impact calibrator, fair clustering, and generalized threshold optimization postprocessor, addressing bias mitigation gaps.

ZS Associates, Associate Consultant *12/2020 - 01/2022*

- Quantified marketing impact on rare disease treatment rates using marketing mix modeling. Optimized marketing portfolio, collaborating with leadership to understand business constraints.

Independent Consultant, Research Economist *08/2018 - 12/2020*

- Led interdisciplinary research projects in collaboration with academic researchers and industry experts on demand for agricultural products, deriving insights with generalized linear models.

North Carolina State University, Researcher & Instructor *08/2016 - 07/2020*

- Identified empirical evidence of alleged chicken price-fixing using advanced time series econometrics, including asymmetric error-correction models, to measure price transmission dynamics.
- Independently developed and taught intermediate-level undergraduate courses.

EDUCATION

North Carolina State University, Ph.D. Economics (Specialization: Econometrics) *2016-2020*
University of Maine, M.A. Economics *2014-2016*
University of Maine, B.S. Economics *2011-2014*

SKILLS

Languages & Frameworks: Python, LangChain, Hugging Face Transformers, scikit-learn, PyTorch
Tooling & DevOps: Git, GitHub Actions, Poetry, Ruff, pre-commit, Sphinx
Expertise: LLM Evaluation & Safety, Uncertainty Quantification, Hallucination Detection, Bias & Fairness, Red Teaming, Responsible AI

RECENT PUBLICATIONS

- **D. Bouchard** and M. S. Chauhan. Uncertainty Quantification for Language Models: A Suite of Black-Box, White-Box, LLM Judge, and Ensemble Scorers. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=W0Fspd4lq5`
- **D. Bouchard**, M. S. Chauhan, D. Skarbrevik, H.-K. Ra, V. Bajaj, and Z. Ahmad. UQLM: A Python Package for Uncertainty Quantification in Large Language Models. *Journal of Machine Learning Research*, 27(13):1–10, 2026b. URL `http://jmlr.org/papers/v27/25-1557.html`
- **D. Bouchard**, M. S. Chauhan, V. Bajaj, and D. Skarbrevik. Fine-Grained Uncertainty Quantification for Long-Form Language Model Outputs: A Comparative Study, 2026a. URL `https://arxiv.org/abs/2602.17431`. *Under review*
- **D. Bouchard**, M. S. Chauhan, D. Skarbrevik, V. Bajaj, and Z. Ahmad. LangFair: A Python Package for Assessing Bias and Fairness in Large Language Model Use Cases. *Journal of Open Source Software*, 10(105):7570, 2025. doi: 10.21105/joss.07570
- **D. Bouchard**. An Actionable Framework for Assessing Bias and Fairness in Large Language Model Use Cases, 2025. URL `https://arxiv.org/abs/2407.10853`. *Under review*
- Full publication list available on Google Scholar page

TALKS AND SERVICE

- "Uncertainty Quantification for Language Models: Standardizing and Evaluating Black-Box, White-Box, LLM Judge, and Ensemble Scorers," NeurIPS 2025 LLM Evaluation Workshop
- Oral presentation "UQLM: Detecting LLM Hallucinations with Uncertainty Quantification in Python," PyData Global, 2025
- Invited talk "UQLM: A Toolkit for LLM Hallucination Detection Using Uncertainty Quantification," AI Alliance Trust and Safety Working Group, Jul 2025
- Reviewer service: NeurIPS, ICLR, ACM TIST, ACM CHI, Journal of Open Source Software