Mr. Brian Formento

Phone: (+65) 8931 8337 | Email: brian.formento@u.nus.edu | Linkedin: linkedin.com/in/brianformento | Github: Aniloid2
Country: Singapore | Address: 1P, Pine Grove | City: Singapore | Postal Code: 590001

Education

National University of Singapore — Singapore
PhD in Computer Science, GPA: 4.63/5 — *Aug 2020 – Present*

- Received the Singapore International Student Award (SINGA) scholarship.
- Research focus on machine learning and adversarial robustness in natural language processing
- Supervised by: Prof See-Kiong Ng (NUS), Dr. Chuan Sheng Foo and Dr. Chen Zhenghua (A*Star).

Southampton University — Southampton, UK
MEng (1st Class Hons) Electronic Engineering with AI — *Sep 2015 – Jul 2019*

Key Modules: Computer Vision, Evolution of Complexity, Advanced Programming, Machine Learning, Computational Biology.

Co-organised and co-run 5 speaker events and 2 workshops throughout the year 16/17 with the soton entrepreneurial society Fish on Toast

Scouted and fast-forwarded through the society's incubator 3 teams, one won £25k from 3 investors in the 2017 university's dragons den.

Pitched a start-up concept (GCX) together with two all-star students in 2016 and won £3k.

Pitched another start-up concept (GrabAPint) together with the same students in 2016 at the university's dragons den organized by Future Worlds.

Experience

Research Associate — July 2025 – Present
Department of Computing, Imperial College London — *London*

- Supervised by Prof. Alessio Lomuscio
- Currently working on a project that combines NLP, reinforcement learning (GRPO/DAPO), and adversarial robustness.

Researcher — Aug 2020 – July 2025
*Institute for Infocomm Research, A*Star* — *Singapore*

- Research on adversarial robustness in LLMs as part of the PhD program
- Published 4 first authorship papers
- Demonstrated how modern LLMs in natural language processing exhibit similar vulnerabilities to their earlier encoding counterparts by introducing the concept of confidence elicitation attacks (ICLR 2025).
- Explored adversarial training in natural language processing using gradient and heuristic guidance in both the discrete token space and the continuous embedding space (NAACL 2024).
- Focused on adversarial attacks utilizing character and word-level perturbations in natural language processing (IJCNN 2021, EACL Findings 2023).

Researcher — Aug 2019 – Aug 2020
Institute of Data Science, National University of Singapore — *Singapore*

- Delivered 2 projects while doing research in computer vision for medical imaging.
- Developed EyeCam, A technology to help medical practitioners detect chronic kidney disease using retinal fundus photography. It uses ResNet and GradCam to highlight important input features. The front end Web UI is in ReactJS. It has been developed for hospitals and polyclinics in Singapore.
- Built a business insight technology with python to analyse the spending behaviour of 1.5M Singaporeans for EzLink and developed a collaborative filtering-based recommender system to match users with merchants.
- Supervised by: Prof Wynne Hsu & Prof Mong Lee.

Intern - Signal Processing — Jul 2017 – Sep 2017
Roke Manor Research — *United Kingdom*

Working with Matlab, C, and Python, for the implementation of a researched SP algorithm.

Intern - Electronic Engineer — Jul 2016 – Sep 2016
L3 ASV — *United Kingdom*

Designed a PCB, reduced costs by 63% per item.

Skills

Programming Languages:

- 6 years of paid experience with the Python programming language
- 6 years of paid development experience in Unix/Linux and Windows environments, both on local PC and cloud/server platforms.

Machine Learning & Deep Learning:

- 6 years of paid development experience with PyTorch, Numpy, Pandas, Matplotlib

Tools & Frameworks:

- 5 years of paid experience with TextAttack, Huggingface Transformers, Git, LaTex

Computer Science:

- 5 years of paid experience with Natural Language Processing (NLP), Adversarial Training and Adversarial Attacks
- 1 year of paid experience in Computer Vision

Research & Soft Skills:

- Technical Presentations, Scientific Creativity, Collaborative Research

Languages

English & Italian: Native

Hackathons

Attended 11, Finalist in 3.

Dreadnode and GovTech AI capture the flag Singapore
26/09/2024

Proposed to three other PhD students to participate in an AI Capture the Flag event in Singapore. Our team placed 23rd out of 500+ international teams.

Aria and Personal Timeline Workshop San Francisco, USA
Meta Menlo Park *06/12/2023*

Flown in to Participate in a 3-day workshop where we developed an AR & AI-augmented fashion app.

LLM Bio Hackathon 2023 Singapore
Gene Chaser Yacht *20/07/2023*

Developed a tech app to train healthcare practitioners in underdeveloped countries to recognize markers of cancerous diseases.

Publications

Brian Formento, Chuan Sheng Foo, See-Kiong Ng. Confidence Elicitation: A New Attack Vector for Large Language Models. International Conference on Learning Representations (**ICLR 2025**). Paper: OpenReview.net. GitHub: CEAttacks

Description: Exploring how new emergent properties of uncertainty estimation in LLMs can be used to craft adversarial examples.

Brian Formento, Wenjie Feng, Chuan Sheng Foo, Luu Anh Tuan, See-Kiong Ng. SemRoDe: Macro Adversarial Training to Learn Representations That are Robust to Word-Level Attacks. North American Chapter of the Association for Computational Linguistics (**NAACL 2024, Main Track, Oral**). Paper: ACLAnthology.org. GitHub: SemRoDe.

Description: Applying distribution matching with MMD, CORAL, or optimal transport (SinkHorn) to BERT to align the base and adversarial distributions in the feature space to learn robust representations.

Brian Formento, Chuan Sheng Foo, Luu Anh Tuan, See-Kiong Ng. Using Punctuation as an Adversarial Attack on Deep Learning Based NLP Systems: An Empirical Study. European Chapter of the Association for Computational Linguistics (**EACL 2023 Findings**). Paper: ACLAnthology.org. GitHub: EmpiricalPunctuationAttacks.

Description: A paper investigating the use of punctuation as an attack vector in deep learning NLP systems.

Brian Formento, See-Kiong Ng, Chuan Sheng Foo. Special Symbol Attacks on NLP. IEEE International Joint Conference on Neural Networks (**IJCNN 2021, Oral**), IEEExplore.org.

Description: An algorithm to discover and exploit special symbols in NLP models such as BERT.

Awards

DAAD AINet Fellowship 2024 – Security in AI (Germany): www.daad.de/en/the-daad/postdocnet/fellows/fellows/

Presentations

Adversarial robustness in deep learning NLP systems (As part of DAAD 2024):

1. LMU, Germany, Prof. Volker Tresp.

2. RUB, Germany, Prof. Ivan Habernal.

Community Service

1. Teaching Assistant (2023): I spent a semester creating and marking assignments for the Text Mining CS5246 module.

2. NUS MComp Application Reviewing (2022): Reviewed 30 applications to the school of computer science for admission purposes.

3. Mentoring Undergrads (2022): Built a statistical data cleaning pipeline based on the ActiveClean model with 4 undergrad students.

4. Supervising Undergrads (2022/2023): Employed two undergrads under the institute of data science (IDS) at NUS, which I was supervising while they worked part-time. We were building a customer feedback classifier and summarizer for Changi Airport Group (CAG).

5. Stack Overflow: Reached over 120k people through my questions and answers with a reputation of approximately 780. stackoverflow.com.