EMILY CHENG

emilyshana.cheng@upf.edu

EDUCATION

Universitat Pompeu Fabra September 2022-2027
PhD Candidate in Linguistics
Linguistic complexity and its representation in biological and artificial brains
Supervisor: Marco Baroni

Massachusetts Institute of Technology January 2022
Master of Engineering in Computer Science (2022) GPA: 4.7/5.0
Supervisors: Boris Katz and Andrei Barbu
Bachelor of Science in Computer Science and Engineering (2020)
Bachelor of Science in Mathematics (2020)

RESEARCH SUMMARY

I study how structure and parsimony arise in representations of language. Using tools from manifold learning (representational level) and information theory (behavioral level), I'm interested in the interaction between representational geometry and functional demands of language. I'm especially interested in how low-dimensional geometry allows biological/artificial brains to handle the vast complexity of language.

PUBLICATIONS

1. **Emily Cheng**, Chris Wang, Greta Tuckute, Marco Baroni, Andrei Barbu. Mapping linguistic complexity in the brain. *In prep.*
2. **Emily Cheng**, Carmen Amo Alonso, Marco Baroni. Linearly Controlled Language Generation with Performative Guarantees. *In prep.*
   Workshop version in MINT@NeurIPS 2024.
3. **Emily Cheng*** and Richard Antonello*. Evidence from fMRI supports a two-phase abstraction process in language models. *In prep.*
   Workshop version in UniReps@NeurIPS 2024. **Awarded best abstract.** (Oral)
4. **Emily Cheng** and Francesca Franzon. Principles of semantic and functional efficiency in grammatical patterning. Preprint, 2024. *Under final review at PNAS.*
5. Jin Hwa Lee*, Thomas Jiralerspong*, Jade Yu, Yoshua Bengio, **Emily Cheng**. Geometric signatures of compositionality across a language model's lifetime. *In Proceedings of ACL 2025.* (Oral)
   Workshop version in NeurReps@NeurIPS 2024.
6. **Emily Cheng**, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, Marco Baroni. Emergence of a High-Dimensional Abstraction Phase in Language Transformers. *In Proceedings of ICLR 2025.* (Poster)
7. **Emily Cheng**, Corentin Kervadec, Marco Baroni. Bridging Information-Theoretic and Geometric Compression in Language Models. *In Proceedings of EMNLP 2023.* (Oral)
8. **Emily Cheng**, Mathieu Rita, Thierry Poibeau. On the Correspondence between Compositionality and Imitation in Emergent Neural Communication. *In Findings of ACL 2023.*
9. **Emily Cheng**, Yen-Ling Kuo, Josefina Correa, Ignacio Cases, Boris Katz, and Andrei Barbu. Quantifying the Emergence of Symbolic Communication. *In Proceedings of CogSci 2022.*
10. **Emily Cheng**, Yen-Ling Kuo, Ignacio Cases, Boris Katz, and Andrei Barbu. Towards Modeling the Emergence of Symbolic Communication. *In Proceedings of the ICRA-2021 Social Intelligence Workshop.*

TALKS & PRESENTATIONS

Linearly Controlled Language Generation with Performative Guarantees
Poster at MINT Workshop at NeurIPS December 2024 *Vancouver*

Geometric Signatures of Compositionality Across a Language Model's Lifetime
Poster at NeurReps Workshop at NeurIPS December 2024 *Vancouver*

Evidence from fMRI supports a two-phase abstraction process in language models.
Talk at RycoLab Seminar @ ETH Zurich November 2024 *Zürich*
Talk at Grammar and Cognition Group @ Universitat Pompeu Fabra October 2024 *Barcelona*
Oral presentation at UniReps@NeurIPS December 2024 *Vancouver*

High-Dimensional Abstraction Phase in Language Transformers
Talk at Area Science Park June 2024 *Trieste, Italy*
Talk at Infolab@MIT CSAIL group meeting June 2024 *Cambridge, MA*
Talk at Johns Hopkins Center for Speech and Language Processing February 2025 *Baltimore*

Bridging Information-Theoretic & Geometric Compression in LMs
Oral presentation at Deep Learning Barcelona Symposium December 2024 *Barcelona*
Talk at Zaslavsky lab (NYU Psychology) August 2024 *NYC*
Oral presentation at EMNLP December 2023 *Singapore*
Talk at Swiss AI Lab (IDSIA) November 2023 *Lugano*
Talk at Evolution in Language (EviL) Seminar September 2023 *Online*

Interplay of functional & semantic aspects in shaping inflectional morphology: a case study on Romance languages November 2023 *Zürich*
Poster at Crosslinguistic Perspectives in Linguistics Conference (X-PPL)
Talk at Evolution in Language (EviL) Seminar December 2024 *Online*
Poster at Mediterranean Morphology Meeting June 2025 *Zadar, Croatia*

EXPERIENCE

Apple Machine Learning Research Summer 2025
MLR Intern *Cambridge, UK*

· Calibrating prompt-based steering in generative models.

Universitat Pompeu Fabra Fall 2022-
Doctoral thesis *Barcelona*

· Explore relationship between linguistic structure, information-theoretic compression, and intrinsic dimensionality in neural language models.

Ecole Normale Supérieure, CNRS Spring-Fall 2022
Visiting researcher *Paris*

· Supervised by Thierry Poibeau. Supported by Paris AI Research Institute.
· Explore relationship between compositionality of emergent languages and ease of imitation learning.

MIT CSAIL Infolab Fall 2020 - Fall 2021
Master's Research *Cambridge, MA*

· Thesis: *Understanding Symbolic Communication*
· Supervised by Boris Katz and Andrei Barbu.
· Characterized the transition from sub-symbolic to symbolic communication between human players, and later machine players via a communication game.

AWARDS & GRANTS

NeurIPS UniReps Workshop Best Abstract 2024
Free registration to NeurIPS 2024 *Vancouver, CA*

UPF Department of Linguistics "Estada" Grant, €3000 2024
Competitive grant for research stay with MIT CSAIL *Barcelona, ES*

UPF Department of Linguistics Travel Grant, €2000 2023
Competitive grant for travel to EMNLP conference *Barcelona, ES*

Brains, Minds, Machines Travel Grant 2023
Fully paid BMM Summer Course, ≈ 10% acceptance *Woods Hole, MA*

Fulbright France Open Research Grant Semifinalist 2021
Meta-learning in low-resource multilingual generalization *Paris, France*

In collaboration with LATTICE at CNRS and École Normale Supérieure.

TEACHING & SUPERVISION

Brains Minds and Machines Summer Course Summer 2024
Teaching Assistant *Woods Hole, MA*

Gave tutorials and supervised course projects.

Madagascar ML Summer School Winter 2022
Guest Lecture

Introduction to Statistical Learning Theory.

6.031 Software Construction Fall 2020
Graduate Teaching Assistant *Cambridge, MA*

Held lab hours, graded assignments for students in intermediate Java software class.

MIT-France Global Teaching Labs January 2020
Instructor *Grenoble, France*

Taught middle, high, and preparatory school students for STEM outreach program.

MIT Math Learning Center Fall 2018 - Spring 2019
Teaching Assistant *Cambridge, MA*

Office hours for Differential Equations, Linear Algebra, Probability, Physics and Calculus.

MIT Math Department Fall 2017, Fall 2018
Grader *Cambridge, MA*

Probability and Random Variables (18.600) and Statistics (18.650).

ACTIVITIES & OUTREACH

Conference and workshop organization
CoNLL 2025 Publication Chair; COLT Symposium 2025;

Reviewer
NeurIPS, ICLR 2025; LanGame@NeurIPS workshop, UniReps@NeurIPS workshop, NeurIPS, ARR, ICLR 2024; EMNLP 2023; NeurIPS 2022; NeurIPS, ICLR 2021

INDUSTRY EXPERIENCE

Palantir Technologies Summer 2020
Software Engineering Intern *Remote*

Developed insurance risk models using PySpark and built map visualization backend in Java, arcGIS

Two Sigma Investments Summer 2019
Quant Research Intern: News Team *New York, NY*

Built alpha models in Python and Groovy to forecast equity and options returns with news data.

Virtu Financial January 2019
Quant Research Intern *New York, NY*

Developed cross-asset market impact models in Python for cash equities execution.

Goldman Sachs Summer 2018
Securities Quant Research Intern: Equities Flow Vol, FICC SMM Execution Services *New York, NY*

- Developed alpha models in Python to forecast realized volatility for single stock options
- Built an order fill model for trading simulation in Java

COURSEWORK

Brains Minds and Machines Summer Course Summer 2023
Project: Intrinsic Dimensionality of Brain Responses to Language *Woods Hole, MA*

Institute of Language, Communication, and the Brain Summer School Summer 2022
Marseille

Official Coursework

6.867 Machine Learning (G)	6.860 Statistical Learning Theory (G)	6.337 Numerical Methods (G)
6.435 Bayesian Inference (G)	6.031 Software Construction	18.615 Stochastic Processes (G)
6.864 NLP (G)	6.046 Design & Analysis of Algorithms	6.436 Probability Theory (G)
6.884 Sensorimotor Learning (G)	24.933 Semantics & Pragmatics (G)	9.190 Comp. Linguistics (G)

SKILLS

Computer Languages	Python, Java, C/C++, Julia
Software & Tools	Pandas/Numpy/Scipy, PyTorch, Git, Linux

LANGUAGES

English (native), Mandarin (fluent), French (C1), Spanish (C1)