Catherine Arnett

catherine.arnett@gmail.com catherinearnett.github.io

Professional Experience

EleutherAI, *remote*
NLP Researcher 2025 - 2026

PleIAs, *Paris, France (remote)*
Research Scientist 2024-2025
Research Intern Summer 2024

Education

PhD, Linguistics, UC San Diego 2019-2025
With a specialization in Computational Social Science
Dissertation: "A Linguistic Approach to Crosslingual and Multilingual NLP"
Committee: Farrell Ackerman (chair), Benjamin Bergen, Leon Bergen, Victor Ferreira

MA, Chinese and Linguistics (Hons), University of Edinburgh 2014-2018
Including one year exchange at Zhejiang University, Hangzhou, P.R. China

Publications

High-Impact Conference Proceedings

1. **Catherine Arnett**, Tyler A. Chang, Stella Biderman, Benjamin K. Bergen (2025). Explaining and Mitigating Crosslingual Tokenizer Inequities. The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS). San Diego, CA, USA.

2. **Catherine Arnett**, Tyler A. Chang, James A. Michaelov, and Benjamin K. Bergen (2025). On the Acquisition of Shared Grammatical Representations in Bilingual Language Models. The 63rd Annual Meeting of the Association for Computational Linguistics (ACL). Vienna, Austria.

3. **Catherine Arnett** and Benjamin K. Bergen (2025). Why do language models perform worse for morphologically complex languages? The 31st International Conference on Computational Linguistics (COLING). Abu Dhabi, UAE and online. **Best Paper Award**

4. Pavel Chizhov*, **Catherine Arnett***, Elizaveta Korotkova, Ivan P. Yamshchikov (2024). BPE Gets Picky: Efficient Vocabulary Refinement During Tokenizer Training. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, FL, USA. *equal contribution.

5. Tyler A. Chang, **Catherine Arnett**, Zhuowen Tu, Benjamin K. Bergen (2024). When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, FL, USA. **Outstanding Paper Award**

6. James Michaelov, **Catherine Arnett**, Benjamin K. Bergen (2024). Revenge of the Fallen? Recurrent Models Match Transformers at Predicting Human Language Comprehension Metrics. The First Conference on Language Modeling (COLM). Philadelphia, PA, USA.

7. James A. Michaelov*, **Catherine Arnett***, Tyler A. Chang, Benjamin K. Bergen (2023). Structural priming demonstrates abstract grammatical representations in multilingual language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore. *equal contribution

8. **Catherine Arnett** (2019). Pathways of Change in Romance Motion Events: A Corpus-based Comparison. *Proceedings of the Thirtieth Western Conference on Linguistics (WeCOL).* Vol 23. Fresno, CA, USA.

Workshop Proceedings

1. **Catherine Arnett**, Marisa Hudspeth, and Brendan O'Connor (2025). Evaluating Morphological Alignment of Tokenizers in 70 Languages. Tokenizer Workshop at ICML. Vancouver, Canada.

2. Sander Land and **Catherine Arnett** (2025). BPE Stays on SCRIPT: Structured Encoding for Robust Multilingual Pretokenization. Tokenizer Workshop at ICML. Vancouver, Canada. **Best Paper Award.**

3. **Catherine Arnett***, Pamela D. Rivière*, Tyler A. Chang, and Sean Trott (2024). Different Tokenization Schemes Lead to Comparable Performance in Spanish Number Agreement. Special Interest Group on Computational Morphology and Phonology (SIGMORPHON) at NAACL. Mexico City, Mexico. *equal contribution

4. **Catherine Arnett***, Tyler A. Chang*, Benjamin K. Bergen (2024). A Bit of a Problem: Measurement Disparities in Dataset Sizes Across Languages. 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages (SIGUL), at LREC-COLING. Torino, Italy. *equal contribution

Refereed Journal Articles

1. Jesse Quinn, Matthew Goldrick, **Catherine Arnett**, Victor S. Ferreira, Tamar H. Gollan (2024). Syntax Drives Default Language Selection in Bilingual Connected Speech Production. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Conference Presentations

1. Sander Land and **Catherine Arnett** (2025). BPE Stays on SCRIPT: Structured Encoding for Robust Multilingual Pretokenization. EurIPS.

2. James A. Michaelov and **Catherine Arnett** (2025). Benchmark Disaggregation Improves Interpretability of Training Dynamics. Interpreting Cognition in Deep Learning Models (CogInterp) Workshop at NeurIPS. San Diego, CA, USA.

3. Jennifer Meng Lu, **Catherine Arnett**, Ruochen Zhang. (2025) Mechanisms of In-Context Syntactic Generalization in Language Models. Blackbox NLP @ EMNLP. Suzhou, China.

4. Jennifer Meng Lu, **Catherine Arnett**, Ruochen Zhang. (2025) Mechanisms of In-Context Syntactic Generalization in Language Models. New England Mechanistic Interpretability Workshop (NEMI). Northeastern Univeristy.

5. **Catherine Arnett** (2025). Toxic Commons: Curating Open-Source Pre-Training Data. *The First Conference of the International Association for Safe & Ethical AI*. Official event of the AI Action Summit. Paris, France.

6. **Catherine Arnett**, Tyler A. Chang, James A. Michaelov, and Benjamin K. Bergen (2023). Crosslingual Structural Priming and the Pre-Training Dynamics of Bilingual Language Models. *The 3rd Workshop on Multilingual Representation Learning co-located with EMNLP 2023*. Singapore.

7. **Catherine Arnett** & Maho Takahashi (2022). Creating a Baseline to Evaluate Correlations Between Language and Environment. *Machine Learning and the Evolution of Language*. Kanagawa/online. [abstract] [poster pdf]

8. **Catherine Arnett** & Eva Wittenberg (2020). Multiple Meanings of Doubling Up: Mandarin Verbal Reduplication. *The 26th Architectures And Mechanisms for Language Processing Conference* (AMLaP). Potsdam/online. [poster pdf]

9. **Catherine Arnett** & Eva Wittenberg (2020). Conceptual Effects of Verbal Reduplication in Mandarin Chinese. *North American Conference on Chinese Linguistics 32*. Storrs, CT/online.

10. **Catherine Arnett** & Eva Wittenberg (2019). Conceptual Effects of Verbal Reduplication in Mandarin Chinese. California Meeting on Psycholinguistics. Santa Cruz, California. [poster pdf]

11. **Catherine Arnett** (2018). Diachronic study of the typology of motion verbs in the Romance languages. Undergraduate Linguistics Association of Britain Conference, University of Edinburgh.

Manuscripts

1. Antonio Lopardo, Avyukth Harish, **Catherine Arnett**, Akshat Gupta (under review). Weight Tying Biases Token Embeddings Towards the Unembedding Space.
2. Pedro Ortiz Suarez*, Laurie Burchell*, **Catherine Arnett***, and 94 other (under review). CommonLID: Re-evaluating State-of-the-Art Language Identification Performance on Web Data. *Equal contribution.
3. James A. Michaelov, **Catherine Arnett**, Pamela D Riviere, Tyler A. Chang, Samuel M. Taylor, Cameron Robert Jones, Sean Trott, Roger P. Levy, Ben Bergen, Micah Altman (under review). Improving the Reliability of Scientific Research involving Pretrained Language Models: Open Weights and Beyond.
4. Pavel Chizhov, **Catherine Arnett**, Mattia Nee, Pierre-Carl Langlais, Ivan P. Yamshchikov. (under review). Common Nonsense Reasoning: Towards Comprehensive Evaluation of Multiple-Choice Question Benchmarks.
5. Tyler A. Chang*, **Catherine Arnett***, and 300+ others (preprint). Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures. *Equal contribution.
6. Pierre-Carl Langlais, Pavel Chizhov, **Catherine Arnett**, Carlos Rosas Hinostroza, Mattia Nee, Eliot Krzystof Jones, Irène Girard, David Mach, Anastasia Stasenko, and Ivan P. Yamshchikov (under review). Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training.
7. Tyler A. Chang, **Catherine Arnett**, Zhuowen Tu, Benjamin K. Bergen (under review). Goldfish: Monolingual Language Models for 350 Languages.

Invited Talks

Linguistics in the Age of LLMs [slides]
Guest Lecture, *Introduction to Linguistics*, Harvard University — July 2025

Glitches in the Embedding Matrix [slides][video]
EleutherAI — June 2025

When is Multilinguality a Curse? Language Models for 350 Languages [slides]
Cambridge NLIP Seminar Series, Cambridge University — June 2025

Best Practices for Open Multilingual LLM Evaluation
PyTorch Day, GOSIM AI. Paris, France. — May 2025

Why do language models perform worse for morphologically complex languages?
NLP Seminar, University of Massachusetts Amherst — September 2025
tinlab, Boston University — March 2025

Characterizing Shared Multilingual Representations with Structural Priming
Brown NLP, Brown University — March 2025

Media and Outreach

NLP Blog Posts
There is no such thing as a tokenizer-free lunch
An Analysis of Multilingual Models on Hugging Face
Best Practices for Open Multilingual LLM Evaluation
Small is Beautiful
wHy DoNt YoU jUsT uSe ThE lLaMa ToKeNiZeR??

Author and Subject-Matter Expert,
Cognitive Foundations, an open Textbook
Update and rewrite portions of the *Language* chapter using bookdown in R.

Teaching Experience

Graduate Teaching Consultant for International Instructors, UC San Diego 2022-2024
Instruct incoming international graduate students in teaching skills and subject-specific English

Instructor of Record, UC San Diego Summer 2023
LIGN 170: Psycholinguistics
Course Website

Teaching Assistant, UC San Diego 2019-2022
LIGN 101: Introduction to Language, Fall 2020-Spring 2022
LIDS 19: Independent Language Study, Winter 2020-Spring 2020
LIGN 170: Psycholinguistics, Fall 2019

Selected Scholarships, Fellowships, and Grants

Yankelovich Graduate Research Funding, UC San Diego, $3,500	2024
Linguistics Department Anti-Racist Pedagogy Micro-Fellowship, UC San Diego Linguistics, $500	2023
Summer Graduate Teaching Scholar, UC San Diego, $1,000	2023
LaVerne Noyes Foundation Endowed Fellowship, UC San Diego, $6,000	2022
Academic Senate Grant, UC San Diego, $9,000	2022
Friends of the International Center Fellowship, UC San Diego, $2000	2020
Eric Liddell China Saltire Scholarship, University of Edinburgh, £5,000	2016
St. Andrews Society of North Carolina Scholarship, $37,000	2014-2018

Awards

Best Paper Award, Tokenization Workshop @ ICML	2025
Best Paper Award, COLING 2025	2025
Outstanding Paper Award, EMNLP 2024	2024
Teaching Assistant Excellence Award, UCSD Linguistics	2023-2024
Edinburgh Award, Peer Learning and Support	2018

Skills and Languages

Languages

English - Native
Mandarin - Proficient
Spanish - Proficient

Coding and Software

R/RStudio tidyverse, lme4, bookdown
Python pandas, matplotlib, seaborn, numpy, BeautifulSoup, transformers
basic SQL, basic HTML

Academic Service

Organizing

Shared Task Organizer, Multilingual Representation Learning Workshop @ EMNLP 2025
Shared Task Organizer, Workshop on Data Quality in Multilingual Pre-Training Data @ COLM 2025
Local organizer, Eval Eval 2025 (unofficial workshop at NeurIPS 2025)

Reviewing

Ad Hoc Journal Reviewer: Language Resources and Evaluation (2025), Journal of Cognitive Science (2025)

Conference Area Chair: LREC (2026), ACL ARR (Feb 2025, May 2025)

Conference Reviewer: LREC (2026), NeurIPS Position Paper Track (2025), NeurIPS (2025), COLM (2025), COLING (2025), Undergraduate Linguistics Association of Britain (2020)

Workshop Program Chair: Multilingual Representation Learning Workshop @ EMNLP 2025, Workshop on Multiligual Data Quality Signals @ COLM 2025

Workshop Reviewer: CogInterp @ NeurIPS (2025), WMDQS @ COLM (2025), MELT @ COLM (2025), Student Research Workshop @ ACL (2025), Tokenizer Workshop @ ICML (2025), Multilingual Representation Learning @ EMNLP (2024, 2025)

Ethics Reviewing: ACL (2025), EMNLP (2024)

Award Committees: Consultant, TMLR Outstanding Paper Committee (2025)

Advising

External PhD Committee Members

Marisa Hudspeth, University of Massachussetts Amherst 2025-

Undergraduate Honors Theses Committee

Yuhan Fu, Cognitive Science B.S. Honors Thesis, UC San Diego 2024
"Is There Evidence for Embodied Simulation During Language Translation?"

Professional Memberships

Association for Computational Linguistics 2023-