Overview

(As of January 25, 2025) I'm a researcher working on using machine learning for code to develop state-of-the-art AI-assisted developer tools to make writing, fixing, and using software easier and more enjoyable.

Education

Academic Qualifications

2016-2021 **PhD in Computer Science**
Massachusetts Institute of Technology, Cambridge, MA.

2013-2016 **Masters in Computer Science**
New York University: Courant Institute of Mathematical Sciences, NY, NY.
GPA: 3.89, MS Research/Thesis Fellowship Award Fall 2015, funding work on A2Q (an order-aware optimizing query compiler for AQuery)

2007-2011 **Bachelor of Arts in Economics and Minor in German Studies**
University of Pennsylvania, Philadelphia, PA.
GPA: 3.93, Phi Beta Kappa, Summa Cum Laude, Dean's List (08, 09, 10)

Industry Work Experience

07/2024 – current **Staff Software Engineer** *DevAI Team*, Google, Atlanta, GA.

06/2022 – 05/2024 **Senior Researcher** *PROSE Team*, Microsoft, Remote.

06/2021 – 06/2022 **Researcher** *PROSE Team*, Microsoft, Remote.
- Working on program synthesis technologies for a variety of developer, data scientist, and end-user applications. A lot of my work focuses on developing and applying large language models to programming tasks, such as program repair and natural language to code synthesis. As part of my job, I also manage and mentor junior researchers through the PROSE research fellowship program.

Summer 2020 **Intern** *Facebook AI Research*, Facebook, Remote.
- Worked with the SysML team at FAIR on a novel tensor compiler, writing C++ for JIT compilation, benchmarking against Halide/TVM

Fall 2018 **Part-Time Research Visitor** *Big Code Team*, Facebook, Remote.
- Applied deep learning to identify and highlight core code functionality in early ML4Code models.

Summer 2018 **Intern** *Software Engineering*, Facebook, Boston.
- Applied deep learning to code search and contributed to some of the earliest ML4Code models in this space.

Summer 2015 **Intern** *Data Science*, Cloudera, San Francisco.

2011 – 2014 **Full-Time Securitized Credit Research Associate** *Non-Agency Mortgages and US Housing*, Morgan Stanley, New York.

215-900-5308 • *josepablocam@gmail.com*
www.josecambronero.com • *www.github.com/josepablocam*

Summer 2010 **Richard B. Fisher Scholar** *Fixed Income Generalist Sales and Fixed Income Credit Strategy*, Morgan Stanley, New York.

Summer 2009 **Douglas Paul Scholar** *Investment Banking and Alternative Investments*, Morgan Stanley, New York.

Academic Work Experience

Fall and Spring 2021 **Advanced Undergraduate Research Class** *TA*, MIT.

2015 – 2016 **Graduate Course in Compiler Construction** *Grader*, NYU.

Fall 2014 **Graduate Course in Programming Languages** *Teaching Assistant*, NYU.

Language skills

- **Programming Languages:** Proficient in: Python, Javascript/Typescript, R, C#.
- **Natural Languages:** Native fluency in English and Spanish. Working proficiency in German.

Service

- **Program Committee ICSE 2024**
- **Program Committee Table Representation Learning Workshop (at NeurIPS) 2023**
- **Program Committee Table Representation Learning Workshop (at NeurIPS) 2022**
- **Artifact Evaluation Committee OOPSLA 2020**
- **Artifact Evaluation Committee CAV 2020**
- **Artifact Evaluation Committee PPoPP 2018**

Mentoring/Advising

- Jennifer McCleary (MIT) MEng Thesis: pancreatic cancer risk modeling (Fall 2019 - January 2020)
- Alex Berg (MIT) Undergraduate research: pancreatic cancer risk modeling (Summer 2020)
- Thomas Xiong (MIT) MEng Thesis: pancreatic cancer risk modeling (Fall 2020 - Spring 2021)
- Lori Zhang (MIT) Undergraduate research: pancreatic cancer risk modeling (Summer 2020 - Spring 2021)
- Harshit Joshi (Microsoft): PROSE Research fellow, automated program repair (Fall 2021 to July 2023 – joining Stanford PhD program 2023)
- Mukul Singh (Microsoft): PROSE Research fellow, NL-to-Code (Spring 2022 to date)
- Abishai Ebenezer (Microsoft): PROSE Research fellow, automated program repair (Fall 2022 to July 2023)
- Jialu Zhang (Yale/Microsoft): Summer intern in the PROSE team, working on automated program repair (Summer 2022). Part of thesis committee.

Publications

[1] M. Singh, J. Cambronero, S. Gulwani, V. Le, C. Negreanu, and G. Verbruggen. Datavinci: Learning syntactic and semantic string repairs. *to appear SIGMOD 2025*, 2025.

[2] P. Rondon, R. Wei, J. Cambronero, J. Cito, A. Sun, S. Sanyam, M. Tufano, and S. Chandra. Evaluating agent-based program repair at google. *arXiv preprint arXiv:2501.07531 (to appear ICSE SEIP 2025)*, 2025.

[3] U. Singh, J. Cambronero, S. Gulwani, A. Kanade, A. Khatry, V. Le, M. Singh, and G. Verbruggen. An empirical study of validating synthetic data for formula generation. *arXiv preprint arXiv:2407.10657 (to appear NAACL Findings 2025)*, 2024.

[4] H. Dong, J. Zhao, Y. Tian, J. Xiong, M. Zhou, Y. Lin, J. Cambronero, Y. He, S. Han, and D. Zhang. Encoding spreadsheets for large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors,

215-900-5308 • *josepablocam@gmail.com*
www.josecambronero.com • *www.github.com/josepablocam*

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20728–20748, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.

[5] S. Barke, C. Poelitz, C. S. Negreanu, B. Zorn, J. Cambronero, A. D. Gordon, V. Le, E. Nouri, N. Polikarpova, A. Sarkar, et al. Solving data-centric tasks using large language models. *NAACL 2024*, 2024.

[6] A. Singha, J. Cambronero, S. Gulwani, V. Le, and C. Parnin. Tabular representation, noisy operators, and impacts on table structure understanding tasks in llms. *arXiv preprint arXiv:2310.10358 (Table Representation Learning at NeurIPS 2023)*, 2023.

[7] M. Singh, J. C. Sánchez, S. Gulwani, V. Le, C. Negreanu, M. Raza, and G. Verbruggen. Cornet: Learning table formatting rules by example. *Proc. VLDB Endow.*, 16(10):2632–2644, jun 2023.

[8] M. Singh, J. Cambronero, S. Gulwani, V. Le, and G. Verbruggen. Emfore: Online learning of email folder classification rules. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 2280–2290, New York, NY, USA, 2023. Association for Computing Machinery.

[9] M. Singh, J. Cambronero, S. Gulwani, V. Le, C. Negreanu, and G. Verbruggen. CodeFusion: A pre-trained diffusion model for code generation. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11697–11708, Singapore, Dec. 2023. Association for Computational Linguistics.

[10] M. Singh, J. Cambronero, S. Gulwani, V. Le, C. Negreanu, E. Nouri, M. Raza, and G. Verbruggen. Format5: Abstention and examples for conditional table formatting with natural language. *VLDB 2024*, 2023.

[11] T. Phung, V.-A. Pădurean, J. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. Singla, and G. Soares. Generative ai for programming education: Benchmarking chatgpt, gpt-4, and human tutors. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 2*, ICER '23, page 41–42, New York, NY, USA, 2023. Association for Computing Machinery.

[12] T. Phung, J. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. Singla, and G. Soares. Generating high-precision feedback for programming syntax errors using large language models. *EDM 2023*, 2023.

[13] H. Joshi, J. C. Sanchez, S. Gulwani, V. Le, I. Radiček, and G. Verbruggen. Repair is nearly generation: Multilingual program repair with llms. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[14] H. Joshi, A. Ebenezer, J. Cambronero, S. Gulwani, A. Kanade, V. Le, I. Radiček, and G. Verbruggen. Flame: A small language model for spreadsheet formulas. *arXiv preprint arXiv:2301.13779 (AAAI 2024)*, 2023.

[15] A. D. Gordon, C. Negreanu, J. Cambronero, R. Chakravarthy, I. Drosos, H. Fang, B. Mitra, H. Richardson, A. Sarkar, S. Simmons, et al. Co-audit: tools to help humans double-check ai-generated content. *arXiv preprint arXiv:2310.01297 (PLATEAU 2024)*, 2023.

[16] J. Cambronero, S. Gulwani, V. Le, D. Perelman, A. Radhakrishna, C. Simon, and A. Tiwari. Flashfill++: Scaling programming by example by cutting to the chase. *Proceedings of the ACM on Programming Languages*, 7(POPL):952–981, 2023.

[17] J. Zhang, J. Cambronero, S. Gulwani, V. Le, R. Piskac, G. Soares, and G. Verbruggen. Repairing bugs in python assignments using large language models. *OOPSLA 2024*, 2022.

[18] B. Wasti, J. P. Cambronero, B. Steiner, H. Leather, and A. Zlateski. Loopstack: a lightweight tensor algebra compiler stack. *arXiv preprint arXiv:2205.00618*, 2022.

[19] R. Bavishi, H. Joshi, J. Cambronero, A. Fariha, S. Gulwani, V. Le, I. Radiček, and A. Tiwari. Neurosymbolic repair for low-code formula languages. *Proc. ACM Program. Lang.*, 6(OOPSLA2), oct 2022.

[20] F. Zogaj, J. P. Cambronero, M. C. Rinard, and J. Cito. Doing more with less: characterizing dataset downsampling for automl. *Proceedings of the VLDB Endowment*, 14(11):2059–2072, 2021.

[21] M. Samak, J. P. Cambronero, and M. C. Rinard. Searching for replacement classes. *arXiv preprint arXiv:2110.05638*, 2021.

[22] T. H. Dang, J. P. Cambronero, and M. C. Rinard. Inferring drop-in binary parsers from program executions. *arXiv preprint arXiv:2104.09669*, 2021.

[23] L. Appelbaum, A. Berg, J. P. Cambronero, T. H. Y. Dang, C. C. Jin, L. Zhang, S. Kundrot, M. Palchuk, L. A. Evans, I. D. Kaplan, et al. Development of a pancreatic cancer prediction model using a multinational medical records database., 2021.

[24] J. P. Cambronero, J. Cito, and M. C. Rinard. Ams: generating automl search spaces from weak specifications. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 763–774, 2020.

[25] L. Appelbaum, J. P. Cambronero, J. P. Stevens, S. Horng, K. Pollick, G. Silva, S. Haneuse, G. Piatkowski, N. Benhaga, S. Duey, et al. Development and validation of a pancreatic cancer risk model for the general population using electronic health records: An observational study. *European Journal of Cancer*, 143:19–30, 2020.

[26] L. Appelbaum, J. P. Cambronero, K. Pollick, G. Silva, J. P. Stevens, H. J. Mamon, I. D. Kaplan, and M. Rinard. Development and validation of a pancreatic cancer prediction model from electronic health records using machine learning., 2020.

[27] J. P. Cambronero, J. Shen, J. Cito, E. Glassman, and M. Rinard. Characterizing developer use of automatically generated patches. In *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 181–185. IEEE, 2019.

[28] J. P. Cambronero and M. C. Rinard. Al: autogenerating supervised learning programs. *Proceedings of the ACM on Programming Languages*, 3(OOPSLA):1–28, 2019.

[29] J. P. Cambronero, T. H. Dang, N. Vasilakis, J. Shen, J. Wu, and M. C. Rinard. Active learning for software engineering. In *Proceedings of the 2019 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, pages 62–78, 2019.

[30] J. Cambronero, H. Li, S. Kim, K. Sen, and S. Chandra. When deep learning met code search. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 964–974, 2019.

[31] J. Cambronero, P. Stanley-Marbell, and M. Rinard. Incremental color quantization for color-vision-deficient observers using mobile gaming data. *arXiv preprint arXiv:1803.08420*, 2018.

[32] J. Cambronero, J. K. Feser, M. J. Smith, and S. Madden. Query optimization for dynamic imputation. *Proceedings of the VLDB Endowment*, 10(11):1310–1321, 2017.

215-900-5308 • *josepablocam@gmail.com*
www.josecambronero.com • *www.github.com/josepablocam*