

David Guzmán

✉ david.guzman@mila.quebec ☎ +1 (437) 973 8963 🌐 davidguzmanr.github.io in enrique-david-guzman-ramirez

Education

Ph.D. in Computer Science , <i>McGill University & Mila - Quebec AI Institute</i> .	<i>August 2025 – Present</i>
M.Sc. in Applied Computing , <i>University of Toronto</i> , GPA: 3.93/4.	<i>September 2022 – June 2024</i>
B.Sc. in Data Science , <i>National Autonomous University of Mexico</i> , GPA: 9.60/10.	<i>August 2019 - June 2021</i>
B.Sc. in Physics , <i>National Autonomous University of Mexico</i> , GPA: 9.54/10.	<i>August 2015 - June 2020</i>

Experience

Data Scientist <i>Pēq</i>	<i>Remote</i> <i>September 2025 - Present</i>
-------------------------------------	--

- Designed and shipped production Python and R packages powering Pēq's geo-based A/B test incremental lift measurement for CPG brands, automating balanced test/control market selection and incremental ROAS (iROAS) estimation across 30+ retailers.
- Developed Bayesian hierarchical lift models in Stan (MCMC and Variational Inference) that scale from ~20 geographic markets to thousands of stores in minutes on GCP, producing non-negative lift estimates with credible intervals for client reporting.
- Built a weekly sales forecasting library and automated the end-to-end measurement pipeline with GCP Cloud Run, eliminating manual market-selection work and integrating LLM-generated insights via the Claude API.

Data Engineer <i>J.D. Power</i>	<i>Mexico City, Remote</i> <i>March 2024 - August 2025</i>
---	---

- Designed and deployed scalable ETL pipelines with PySpark and BigQuery SQL, processing multi-source data across multiple client accounts and reducing latency by 30%.
- Created internal Python libraries to standardize and automate pipeline deployment via GitLab, decreasing manual coding effort by 20% and increasing reusability across projects.
- Developed and integrated robust data quality and validation checks, improving data accuracy and reducing data-related issues by 30%.

Machine Learning Research Intern <i>National Research Council Canada</i>	<i>Toronto, Canada</i> <i>May 2023 - December 2023</i>
--	---

- Collaborated in the *Speech Generation for Indigenous Language Education* project to develop state-of-the-art speech synthesis systems and techniques with Deep Learning for Indigenous languages in Canada.
- Contributed to the development of *EveryVoice TTS Toolkit*, a tool for creating text-to-speech systems focused on low-resource languages, achieving robust performance with up to 90% less data.

Data Scientist <i>Grupo Salinas</i>	<i>Mexico City</i> <i>November 2021 - July 2022</i>
---	--

- Developed a collaborative filtering recommender system for the homepage of the E-commerce area with PyTorch, which increased precision at 10 by 50%.
- Developed a content-based recommender system and association rules for the product detail pages of the E-commerce area with scikit-learn and MLxtend, automating the original two week manual process.
- Deployed ML APIs on AWS SageMaker for real-time personalization and monitored performance using AWS CloudWatch.

Skills

- **Programming Languages:** Python, R, SQL, Julia, Fortran.
- **Machine and Deep Learning:** Scikit-learn, XGBoost, PyTorch, Keras, TensorFlow, PySpark MLlib.
- **NLP:** NLTK, spaCy, Hugging Face, LangChain.
- **Data Analysis and Manipulation:** Numpy, SciPy, Pandas, Matplotlib, Seaborn, Bokeh, Plotly.
- **Cloud services:** AWS, GCP.
- **Databases:** PostgreSQL, Redshift, MongoDB, Redis.
- **Miscellaneous:** Linux, Git, LaTeX, Slurm, REST APIs, Scrum.
- **Certifications:** [AWS Certified Machine Learning - Specialty](#) .
- **Languages:** Spanish (Native), English (Proficient), French (Intermediate).

Publications

- **David Guzmán**, Luel Hagos Beyene, Jesujoba Oluwadara Alabi, Yejin Jeon, Dietrich Klakow, David Ifeoluwa Adelani. “*OpenBibleTTS: Large-Scale Speech Resources and TTS Models for Low-Resource Languages*.” arXiv preprint, 2026. Available at: <https://arxiv.org/abs/2606.09553> .
- Minh Duc Bui, **David Guzmán**, Abteen Ebrahimi, et al. “*Findings of the AmericasNLP 2026 Shared Task on Cultural Image Captioning for Indigenous Languages*.” Proceeding of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP). 2026. Available at ACL Anthology: <https://aclanthology.org/2026.americasnlp-6.27/> .
- Kosei Uemura, **David Guzmán**, Quang Phuoc Nguyen, Jesujoba Oluwadara Alabi, En-Shiun Annie Lee, David Ifeoluwa Adelani. “*MERLIN: Multi-Stage Curriculum Alignment for Multilingual Encoder-LLM Integration in Cross-Lingual Reasoning*.” Presented at the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2026. Available at ACL Anthology: <https://aclanthology.org/2026.eacl-long.277/> .
- Steve Bakos, Félix Gaschi, **David Guzmán**, Riddhi More, Kelly Chutong Li, En-Shiun Annie Lee “*AlignFreeze: Navigating the Impact of Realignment on the Layers of Multilingual Models Across Diverse Languages*.” Presented at the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2025. Available at ACL Anthology: <https://aclanthology.org/2025.naacl-short.48> .
- Aidan Pine, Erica Cooper, **David Guzmán**, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha’ Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, Junichi Yamagishi. “*Speech Generation for Indigenous Language Education*.” 2025. Computer Speech & Language, Volume 90. Available at: <https://doi.org/10.1016/j.csl.2024.101723> .
- Tong Su, Xin Peng, Sarubi Thillainathan, **David Guzmán**, Surangika Ranathunga, En-Shiun Annie Lee. “*Unlocking Parameter-Efficient Fine-Tuning for Low-Resource Language Translation*.” Presented at the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2024. Available at ACL Anthology: <https://aclanthology.org/2024.findings-naacl.263> .