David Guzman Piedrahita

davidguzman1120@gmail.com

davidguzman1120 | davidguzmanp

Zurich, Switzerland

Education

- **University of Zurich** *September 2022 - Ongoing*
  Master's in Computer Science Zurich, Switzerland
  - Major in Artificial Intelligence with focus on **Natural Language Processing** and **Large Language Models**.
  - Current GPA: 5.7/6.

- **University Of Bergamo** *September 2019 - July 2022*
  Bachelor's in Computer Science Engineering Bergamo, Italy
  - Final grade: 107/110
  - Ranked top 1% of faculty.
  - Ranked 2nd among engineering students, 1st among Bachelor students in my year and program.

Experience

- **ETH Zurich & Max Planck Institute for Intelligent Systems [🌐]** *September 2025 - Present*
  Research Assistant Zurich, Switzerland & Tübingen, Germany
  - Led a **position paper** on the **sociopolitical risks** of increasingly agentic AI (**ICML 2026 spotlight**), working alongside a senior author and advisor network including **Yoshua Bengio**, **Stuart Russell**, **Audrey Tang**, and **Bernhard Schölkopf**.
  - Developed **multi-agent LLM systems** to study the emergence and evolution of **sanctioning norms** in cooperative social dilemmas under the supervision of Professor Zhijing Jin and Bernhard Schölkopf.
  - Wrote and co-designed a funding proposal under faculty supervision; project funded by the **AI Safety Fund Award** ($499,838, 2025–2026). PI: Prof. Zhijing Jin.
  - Mentored and onboarded **three AI safety research interns** during a South Africa research program with CAIRF / AI Safety South Africa collaborators.

- **University of Zurich [🌐]** *March 2024 - August 2024*
  Research Assistant Zurich, Switzerland
  - **Fine-tuned large scale LLMs** on Cerebras chips under the supervision of Professor Rico Sennrich.
  - **Improved domain-specific performance by 68%**, as measured by domain metrics. The model can be used as an **LLM chatbot with highly specific knowledge about Switzerland**.

- **ETH Zurich [🌐]** *September 2023 - February 2024*
  Research Assistant Zurich, Switzerland
  - Conducted large-scale web scraping initiatives, storing millions of PDFs in **PostgreSQL** databases.
  - **Reduced HTML extraction time by 43%** by prioritizing direct extraction from URLs over browser simulation.
  - Coordinated with cross-functional teams to maintain **clean coding practices**, thorough **documentation**, and version control via Git.

Publications and Conferences

C=Conference, J=Journal, P=Patent, S=In Submission, T=Thesis

[C.1] **Guzman Piedrahita, D.**, Banerjee, D., Li, C., Zhang, T. J., Blin, K., Simko, S., Pandey, P. S., Strauss, I., Mihalcea, R., Schölkopf, B., & Jin, Z. (2026). **Position: Safe Models Do Not Guarantee Safe Societies: The Case for Sociopolitical Risk**. In *International Conference on Machine Learning (ICML) 2026, Position Paper Track*. July 10-11, 2026, Seoul, South Korea. *Spotlight presentation*.

- Led a position paper arguing that **model-level safety** does not by itself address **sociopolitical risk** from increasingly agentic AI systems.
- **Extended preprint**, ***AI Poses Risks to Democratic and Social Systems***, with a broader senior author and advisor network including **Yoshua Bengio**, **Stuart Russell**, **Audrey Tang**, **Bernhard Schölkopf** and **Richard Mallah**.

[C.2, T.2] **Guzman Piedrahita, D.**, Yang, Y., Sachan, M., Ramponi, G., Schölkopf, B., & Jin, Z. (2025). **Corrupted by Reasoning: Reasoning Language Models Become Free-Riders in Public Goods Games**. In *Conference on Language Modeling (COLM) 2025*. October 7-10, 2025, Montreal, Canada. Review scores in the 95th percentile. *Best Oral Paper — REALM Workshop, ACL 2025*. *(Master's Thesis, University of Zurich, ETH Zurich & Max Planck Institute for Intelligent Systems)*

- Investigated the behavior of **reasoning-augmented language models** in **public goods games**, finding that reasoning can lead to **free-riding behavior**.
- Developed **multi-environment benchmarks** using **game-theoretic scenarios** to systematically simulate agent interactions and the influence of punishments and rewards.

[C.3] **Guzman Piedrahita, D.**, Strauss, I., Schölkopf, B., Mihalcea, R., & Jin, Z. (2026). **Democratic or Authoritarian? Probing a New Dimension of Political Biases in Large Language Models**. In *European Chapter of the Association for Computational Linguistics (EACL) 2026*. *Oral presentation*.

- Proposed a novel methodology to assess **LLM alignment** with the **democracy-authoritarianism spectrum**.
- Found that LLMs generally favor **democratic values** and leaders, but exhibit increased favorability toward **authoritarian figures** when prompted in Mandarin.

[C.4] Samway, K., Kleiman-Weiner, M., **Guzman Piedrahita, D.**, Mihalcea, R., Schölkopf, B., & Jin, Z. (2025). **Are Language Models Consequentialist or Deontological Moral Reasoners?**. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2025*. November 5-9, 2025, Suzhou, China.

- Introduced a **taxonomy of moral rationales** to systematically classify **reasoning traces** according to **consequentialism** and **deontology**.
- Revealed that LLM **chains-of-thought** tend to favor **deontological principles**, while **post-hoc explanations** shift toward **consequentialist rationales**.

[S.1] Backmann, S., **Guzman Piedrahita, D.**, Tewolde, E., Mihalcea, R., Schölkopf, B., & Jin, Z. (2025). **When Ethics and Payoffs Diverge: LLM Agents in Morally Charged Social Dilemmas**. Manuscript submitted for publication in *arXiv*.

- Introduced **Moral Behavior in Social Dilemma Simulation (MoralSim)** to evaluate how LLMs behave in the **prisonerś dilemma** and **public goods game** with morally charged contexts.
- Showed substantial variation across models in both their general tendency to **act morally** and the **consistency of their behavior** across game types.

[S.2] Fazla, A., Krauter, L., **Guzman Piedrahita, D.**, & Michail, A. (2025). **Robustness of Misinformation Classification Systems to Adversarial Examples Through BeamAttack**. Manuscript submitted for publication in *arXiv*.

- Introduced **BeamAttack**, a novel algorithm for generating **adversarial examples** in natural language processing through the application of **beam search**.
- Utilized a **masked language model** to predict contextually plausible alternatives to the words to be replaced, enhancing the **coherence of generated adversarial examples**.

[T.1] **Bachelorś Thesis: LSTM-based Time Series Forecasting for Air Quality**. Completed at *University of Bergamo*

- Evaluated **LSTM neural networks** for air-quality **time-series forecasting** in Lombardy, Italy.
- Employed **surrogate models**, specifically **LIME**, to assess **feature importance** and enhance **interpretability**.

Awards and Scholarships

- **ICML 2026 Spotlight Paper** *2026*
  International Conference on Machine Learning (ICML)
  ◦ Accepted as a **Position Paper Track spotlight** for *Position: Safe Models Do Not Guarantee Safe Societies: The Case for Sociopolitical Risk*.
- **AI Safety Fund Award** *May 2025*
  AI Safety Fund
  ◦ Co-authored the winning proposal with Prof. Zhijing Jin, securing $499,838 to study sanctioning norms in multi-agent LLM systems.
- **Zurich AI Safety Day + Anthropic API Research Credit** *2025*
  Zurich AI Safety Day
  ◦ Invited to poster presentation on LLM bias evaluation.
  ◦ Awarded $500 API research credit at Zurich AI Safety Day to support AI safety experiments.
- **Oral Paper Award** *July 2025*
  ACL 2025 NLP - REALM Workshop

◦ Awarded for the paper "Corrupted by Reasoning: Reasoning Language Models Become Free-Riders in Public Goods Games".

- **Full Scholarship** *September 2020 - July 2022*
  University of Bergamo [🌐]
  ◦ Full scholarship awarded based on academic merit after the first year.
  ◦ Extended from the second to the third year based on sustained academic excellence.

Presentations and Academic Service

- **EACL 2026 Oral Presentation** *2026*
  European Chapter of the Association for Computational Linguistics (EACL)
  ◦ Presented ***Democratic or Authoritarian? Probing a New Dimension of Political Biases in Large Language Models*** as an **archival oral presentation**.
- **IASEAI 2026 Oral Presentation** *2026*
  IASEAI
  ◦ Presented ***Democratic or Authoritarian? Probing a New Dimension of Political Biases in Large Language Models*** as a **non-archival oral presentation** at IASEAI 2026 in Paris, France.
- **ICML 2026 AI4GOOD Workshop Co-Organizer** *July 2026*
  International Conference on Machine Learning (ICML)
  ◦ Co-organizing the **Trustworthy AI for Good (AI4GOOD) Workshop** on trustworthy AI, sociopolitical risk, AI for social good, and cooperative AI.
  ◦ Organizing team includes **Rada Mihalcea**, **Milind Tambe**, **David Lie**, **Zhijing Jin**, **Terry Jingchen Zhang**, and **Changling Li**; confirmed speakers/panelists include **Yoshua Bengio**, **Joel Leibo**, **Oana Ignat**, **Maksym Andriushchenko**, and **Lewis Hammond**.

Languages

Languages:

- English (Fluent, TOEFL C1)
- Italian (Fluent, PLIDA C1)
- Spanish (Mother tongue)