

EDUCATION	Universidad Nacional de San Martín BSc Data Science (GPA 8.64 / 10)	2022 – 2027 (expected)
PUBLICATIONS	<p>PowerBench: Measuring Language Model Bias in Power-shifting Requests 2026 N. Martorell, W. Brau, G. A. Heredia, T. P. Korenblit, G. Labastíe, T. Gimenez Molina <i>arXiv preprint arXiv:2610.02303</i> Preprint</p> <ul style="list-style-type: none"> – Evaluates 24 models across eight languages and human/AI-agent requesters – Finds nationality biases and greater refusal when AI agents seek power <p>Counterfactual Resampling to Analyse Model Behaviour 2026 T. P. Korenblit <i>BlueDot Impact Technical AI Safety sprint, facilitated by BAISH</i> writeup repo</p> <ul style="list-style-type: none"> – Extends Betley et al.'s Value Leakage introducing per-conversation measurements by resampling truncations of the CoT under a flipped prompt – CoTs that deny being influenced do not concentrate on the influenced traces, against the hypothesis that the model selectively covers its tracks <p>Not All Instructions Are Forgotten Equal 2026 T. P. Korenblit <i>Memorias de las 55 JAIIO (ASAIID)</i> paper repo</p> <ul style="list-style-type: none"> – Measures retention of in-context instructions individually in multi-turn coding sessions – Per-instruction effects range from -2.4 to +5.4 log-odds and only three of twelve instructions benefit from being restated 	
EXPERIENCE	<p>AISAR Oct 2026 – Present <i>Research Fellow</i> (part-time)</p> <ul style="list-style-type: none"> – Research on white-box AI control. <p>Independent Research Apr 2026 – Present</p> <ul style="list-style-type: none"> – Research on AI Safety. Interests: scalable oversight, interpretability. – Facilitator for BlueDot Impact's Technical AI Safety course – Contributor to Alineados, BAISH's Spanish-language AI safety newsletter. <p>Ascendancy <i>Data Scientist</i> Jan 2026 – Apr 2026</p> <ul style="list-style-type: none"> – Built a graph database of investors, founders, and companies across the Americas, MENA, and Europe, with warm-intro pathfinding and community detection – Wrote an assistant to turn English into Cypher, so non-technical users can query the graph directly (Python, Neo4j, PostgreSQL, FastAPI, GCP) <p>Invisible Technologies Dec 2022 – Dec 2025 <i>Software Engineer</i> (Jun 2025) <i>QC Data Analyst</i> (Jun 2023) <i>AI Data Trainer</i> (Dec 2022)</p> <ul style="list-style-type: none"> – Designed a RAG platform to create almost all RLHF annotation interfaces – Wrote validation scripts and dashboards (Python, Looker) used company-wide. Reviewed output quality for the RLHF data teams <p>Klouser <i>Data Scientist / Data Engineer</i> Dec 2021 – Oct 2022</p> <ul style="list-style-type: none"> – Made a lead generation system and trained a classifier for client profiles (Python, Pandas, PostgreSQL, Scikit-learn) 	