

Tomás Pablo Korenblit

EDUCATION	Universidad Nacional de San Martín BSc Data Science	2022 – 2027 (expected)
PUBLICATIONS	Not All Instructions Are Forgotten Equal T. P. Korenblit <i>JAIIO 2026 (ASAIID)</i> – Aggregate compliance hides per-instruction failures. Some instructions get dropped deep in a conversation while the mean still looks fine, so a guardrail set at the start can quietly stop holding. Argues for watching instructions one at a time, not in aggregate – Restating an instruction only helps for a handful of them. Most are already followed by default, so blanket repetition mostly wastes tokens – Scored with deterministic checkers instead of an LLM judge and the effects were estimated with a hierarchical Bayesian model	2026 paper repo
	PowerBench: LLM Refusal in Power-Grabbing Requests W. Brau, T. Gimenez Molina, G. A. Heredia, T. P. Korenblit, G. Labastie, N. Martorell <i>Apart Research Global South AI Safety Hackathon</i> . In preparation – Adding self-benefit to a harmful request lowers refusal. Models refuse pure disempowerment (harm to others) more than power-grabbing (harm plus self-gain) – Request refusal does not correlate with model capability – Multilingual, single-turn first-person requests, scored by an LLM judge	2026 paper
EXPERIENCE	Research – Studying LLM threat models. Extending the PowerBench work – Facilitating BlueDot’s Technical AI Safety course, and taking their Projects course Ascendancy <i>Data Scientist</i> – 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) Invisible Technologies <i>Software Engineer</i> (Jun 2025) <i>QC Data Analyst</i> (Jun 2023) <i>AI Data Trainer</i> (Dec 2022) – Designed a RAG platform to create almost all RLHF annotation interfaces – Wrote validation tools and dashboards (Python, Looker) used company-wide. Reviewed output quality for the LLM training teams Klouser <i>Data Scientist</i> / <i>Data Engineer</i> – Made a lead generation system and trained a classifier on client profiles to qualify them (Python, Pandas, PostgreSQL, Scikit-learn)	Apr 2026 – Present Jan 2026 – Apr 2026 Dec 2022 – Dec 2025 Dec 2021 – Oct 2022