



Master's Thesis & Internship:

Advanced Interpretability for Generative AI

Location: Munich (Hybrid 3 days in-office) / Distributed (US Collaboration)

Duration: 6+ Months (Thesis) | 3-6 Months (Internship)

Start Date: Immediate

Language: English (German optional)

Who we are: Sureel.ai

At **Sureel.ai**, we believe that for AI to truly benefit society, it must be transparent and accountable. We aren't just solving a legal or IP problem; we are building the fundamental infrastructure for **AI Attribution and Lineage**. Our mission is to ensure that as AI scales, we don't lose the thread of human legacy. We want to make the world a better place by providing the tools that allow innovation to move faster, not slower, because users, creators, and regulators can finally trust what is happening inside the "black box." We are a venture-backed, technically-driven team split between Munich and the US.

The Vision

AI capability is no longer the bottleneck; knowing what models are actually doing is. We are moving into an era where we must prove *why* a model made a decision, especially under real-world constraints like noisy data, privacy requirements, and adversarial behavior. This position is for students who want to work at the edge where research meets production. We are building systems that interrogate neural networks in real-time. The environment is technically demanding, but we keep it grounded: high standards, frequent feedback, and enough room to explore ambitious ideas that turn into an excellent thesis.



Research Focus & Creative Autonomy

Your work will contribute directly to the core science of how we "interrogate" neural networks. We focus on several key pillars:

- **Training Data Attribution (TDA):** Developing methods to trace model behavior back to specific training fragments to understand influence and preserve data lineage.
- **Mechanistic Interpretability:** Reverse-engineering internal neural "circuits" to discover how facts, styles, and logic are represented in weights.
- **Neural Tracing & Provenance:** Connecting internal activations into inspectable chains of influence to build evidence-based "why" answers.
- **Robustness & Evaluation:** Stress-testing the faithfulness of XAI methods to ensure they remain stable under production-grade scrutiny.

While we provide these tracks as a foundation, we want you to bring your own ideas. If you have a novel hypothesis about model transparency or a "wild" idea to help preserve human legacy in the age of AI, we will provide the compute power and mentorship to help you prove it.

The Publication-First Goal

We don't view a Master's thesis as a mere graduation requirement; we view it as a potential breakthrough. Our minimum goal for every student is a co-authored contribution to a peer-reviewed publication at top-tier venues (e.g., ICML, NeurIPS, ICLR, or specialized XAI workshops). You'll be mentored to produce research that the global AI community can build upon.

Team, Office, and The "Vibe"

You'll be part of a distributed team, collaborating closely with our US-based engineers while being anchored in our Munich office. We believe in the power of "in-person" bandwidth; the best breakthroughs often happen when staring at the same whiteboard. Our hybrid model requires you to be in the Munich office at least 3 days a week (but we are flexible). It's a mix of deep-work sprints and high-bandwidth collaboration. We aim for excellence without the ego. We take the science seriously, but we make sure the process is genuinely fun.



Requirements

Must Have:

- **Enrollment:** Currently enrolled at **TU Munich** (Informatics, Mathematics, Data Engineering, Electrical Engineering, or similar).
- **Programming:** High proficiency in **Python** and the **PyTorch** ecosystem.
- **Theory:** A solid grasp of Deep Learning fundamentals (Transformers, training dynamics, loss landscapes, algebra, etc).
- **Scientific Rigor:** The ability to read complex papers critically and a mindset that values controlled experiments over "lucky" results.

Nice to Have:

- Prior exposure to interpretability methods (attribution, probing, SAEs).
- Experience with provenance systems, Knowledge Graphs, or data pipelines.
- A desire to bridge the gap between "Black-Box" AI and human-readable evidence.

Who You Are

- **Deeply Curious:** You aren't satisfied with models that just "work." You want to take them apart to understand the underlying mechanics of *why*.
- **An Original Thinker:** You're not afraid to pitch a unique idea or challenge the status quo. We value students who think outside the box rather than just following a leaderboard.
- **A Collaborative Soul:** You enjoy the energy of a shared office. You're ready to dive into a whiteboard session in Munich and then sync with the US team to see how your research scales.
- **Resilient:** Research is messy and experiments fail. You have the grit to stay curious through the "it's not working" phase until you find the breakthrough.
- **High Standards, Low Ego:** You're here to do the best work of your degree so far and enjoy the process of getting there with a team that has your back.

How to Apply

Send your **CV (PDF)** and a **cover letter** to tamay@surreel.ai describing:

1. Tell us about yourself, who you are, what you are interested in
2. Which track interests you most (or your own research idea).
3. A project or paper you found genuinely exciting and why.