MIMIR

Open Positions 2026:
Data Engineer, ML Scientist, Multi-Agent Dev

Funded by
SPRIN-D
Bundesagentur für Sprunginnovationen

Building the best AI router, optimized for quality, compliance, and cost.

Routing, the problem of selecting which model or agent to use for a given task, becomes critical infrastructure as AI systems become more capable, complex and expensive. The best router maximizes output quality while minimizing cost, and ensures compliance with customer specific requirements like the EU AI Act and individual enterprise risks. MIMIR is building that router. Our work spans the full technical stack: novel architectures for learned routing and outcome prediction, evaluation frameworks for measuring agent behaviour and routing quality, and the multi-agent infrastructure that makes intelligent orchestration possible at scale. We're a deep-tech team working on hard ML problems with real product leverage.

HOW THESE ROLES WORK

Each role below lists example projects that are real work happening at MIMIR right now. An employee joining for a 6-month engagement (SPRIN-D initial funding runtime, longer horizons likely) would typically focus on one to two of these.
Roles can be structured as a thesis project (Bachelor's or Master's), a PhD research collaboration, or a paid part-time / full-time position. Hybrid arrangements combining academic credit and compensation are possible, and we are happy to publish in most areas.

Data Engineer

Build the data pipelines and benchmark datasets that power AI agent evaluation at scale.

Thesis / Collab / PT / FT | **~6 months** | **Tübingen / Remote**

ABOUT THE ROLE

Evaluating AI agents and multi-agent systems requires large amounts of high-quality, well-structured data. As a Data Engineer at MIMIR, you'll design and implement the pipelines, datasets, and data infrastructure that power our evaluation suite. You'll work at the intersection of ML data engineering and agentic AI research, turning diverse sources into reliable, reproducible benchmarks for agent behaviour, routing quality, and system-level properties that are genuinely hard to measure. This role has direct relevance to regulatory

compliance: your work ensures that routing decisions meet EU AI Act requirements where applicable.

EXAMPLE PROJECTS

Below are representative projects. You'd work on one or two of these over an initial 6-month engagement, depending on your interests and the team's current priorities.

- **Agent Evaluation Benchmark Construction.** Designing and building novel benchmark datasets for evaluating AI agent behaviour, which need to cover tool use, multi-step planning, delegation between agents, and failure modes in multi-agent pipelines. We fill the gaps where no adequate evaluation data currently exists.
- **Evaluation Landscape Analysis.** Systematically mapping the landscape of existing AI evaluation benchmarks, in particular assessing which ones reliably measure agent capabilities, routing quality, and compliance-relevant properties, and producing a structured gap analysis.
- **Pipeline Development.** Building and maintaining robust data pipelines that ingest, clean, version, and serve benchmark datasets to the evaluation infrastructure, thereby ensuring reproducibility and traceability across model versions and routing configurations.
- **Benchmark Quality & Validation.** Assessing the reliability and validity of both existing and newly developed benchmarks. Testing for metric stability, sensitivity to real performance differences, and robustness to prompt variations and distributional shift.
- **Systems Evaluation & Pipeline Integration.** Running benchmarks against real agentic systems and feeding results into the routing pipeline. Tracking dataset performance over time, refining data composition based on live evaluation, and closing the loop between evaluation and deployment.
- **Data Generation & Curation.** Subsampling and curating existing public datasets, generating synthetic agent trajectories via LLMs, and applying programmatic augmentation or expert annotation to create targeted evaluation splits for specific agent capabilities.

WHAT WE'RE LOOKING FOR

- Strong Python skills, particularly for data processing (pandas, HuggingFace datasets, Spark, or similar)
- Interest in AI agents, multi-agent systems, or AI evaluation
- Systematic, detail-oriented approach to data quality
- Bonus: experience with LLM-based data generation, synthetic data, or annotation workflows
- Bonus: familiarity with agentic AI frameworks or evaluation methodology
- Bonus: Experience building or maintaining data pipelines: ETL, versioning, and reproducibility practices
- Bonus: interest in AI governance, regulatory compliance (EU AI Act), or enterprise risk assessment

WHAT YOU'LL GAIN

- Hands-on experience with frontier agentic AI systems and the open problems in evaluating them
- Publication opportunities in AI evaluation and multi-agent systems research
- Deep expertise in data engineering for ML, applied to one of the field's most challenging problems

- Collaboration with ML researchers, systems engineers, business lawyers and policymakers
- Direct impact on a production routing system optimizing for quality, compliance, and cost
- Fair compensation for paid positions; rates depend on experience and engagement model

---

ML Scientist

Design novel architectures for learned routing and model outcome prediction.

Thesis / Collab / PT / FT **~6 months** **Tübingen / Remote**

ABOUT THE ROLE

Building the best AI router requires solving hard ML problems that sit at the frontier of architecture research. How do you train a model to predict which other model will perform best on a given task, when the models you're routing between are themselves constantly improving? How do you handle retrieval from a rapidly growing database of past routing decisions? How do you optimize a routing layer that needs to learn attention over an unordered set of candidate models? As an ML Scientist at MIMIR, you'll work on these questions. This is deep architecture work with direct product impact.

EXAMPLE PROJECTS

Below are representative projects. You'd work on one or two of these over an initial 6-month engagement, depending on your interests and the team's current priorities.

- **Learned Top-K Selection Layer.** Designing and implementing learned attention mechanisms for Top-K model selection. Training keys and values that enable a routing layer to efficiently select the K most promising models or agents for a given input, with differentiable scoring, while only seeing limited data during training.
- **Attention Over Unordered Fragments.** Developing attention architectures that operate over unordered, only locally causal fragments, enabling the router to reason over heterogeneous sets of model outputs or partial agent trajectories without imposing a global ordering.
- **Fast Inner Product Retrieval.** Building fast retrieval systems based on inner product search over a frequently appended, designable database of routing decisions and outcomes, optimizing for both speed and accuracy as the database scales.
- **Inference with Partial Information.** Designing inference methods that handle partially incomplete retrieval information. Make routing decisions where not all candidate model scores are available, or where retrieval from the outcome database returns approximate results.
- **Multi-Objective Training with Imbalanced Data.** Training multi-objective architectures where different prediction heads (e.g., quality, cost, compliance) have very differently balanced amounts of training data, handling class imbalance and objective weighting in a principled way.
- **Several other research questions for which we explore optimal solutions include:**
    - Inference with partially incomplete retrieval information

- Training multiple objectives (as heads) with very differently balanced amounts of training data (correctness, fairness scores, compliance scores)
- Context sharing between AI agents
- Task-specific AI agent and tool cost estimation
- Data governance in multi-agent systems

WHAT WE'RE LOOKING FOR

- Strong background in deep learning, particularly attention mechanisms, retrieval, or model architectures
- Experience implementing and training neural network architectures from scratch (PyTorch or JAX)
- Interest in router design, model selection, or meta-learning
- Comfort reading and implementing ideas from recent ML papers
- Bonus: experience with ranking models, recommendation systems, or learned optimizers
- Bonus: familiarity with efficient attention variants (linear attention, sparse attention, etc.)
- Bonus: background in information retrieval, approximate nearest neighbor search, or vector databases

WHAT YOU'LL GAIN

- Deep hands-on experience with novel architecture design at the frontier of learned routing
- Publication opportunities in top-tier ML venues (NeurIPS, ICML, ICLR)
- Collaboration with researchers and engineers building a production routing system
- Expertise in a genuinely open problem space: router mechanisms and architectures are still very much unsolved
- Fair compensation for paid positions; rates depend on experience and engagement model

---

Developer: Router Output & Multi-Agent Platform

Build the routing infrastructure, multi-agent systems, and interfaces that bring evaluation to users.

Thesis / Collab / PT / FT **~6 months** **Tübingen / Remote**

ABOUT THE ROLE

The technical heart of MIMIR is a system that routes evaluation tasks across models and agents, orchestrates multi-agent pipelines, aggregates results, and surfaces them to users in useful forms. As a Developer focused on platform and infrastructure, you'll build those layers: the routing platform that orchestrates evaluation runs across heterogeneous AI systems, the multi-agent frameworks that enable coordination between models, the APIs that expose results to external users, and the interfaces that make evaluation data legible and actionable. Specialize in any one of the technical areas below or combine them.

EXAMPLE PROJECTS

Below are representative projects. You'd work on one or two of these over an initial 6-month engagement, depending on your interests and the team's current priorities.

- **Agentic Routing Platform.** Designing and building a model routing platform that intelligently dispatches evaluation tasks across different models and agent backends. Besides centrally selecting the right model for a given task, this includes handling request queuing, load balancing, and result aggregation across a heterogeneous set of AI systems.
- **Multi-Agent System Development & Evaluation.** Building and evaluating multi-agent systems, designing coordination protocols, delegation strategies, and failure recovery mechanisms that enable multiple AI agents to work together effectively on complex evaluation tasks.
- **Evaluation API.** Building a public-facing API that enables external developers and enterprises to submit AI systems for evaluation, query benchmark results, and retrieve structured output, with clean versioning, authentication, and documentation.
- **Server & Infrastructure.** Setting up and hardening the server infrastructure underpinning MIMIR's evaluation pipeline. Implement components of containerization, CI/CD, monitoring, and reliability engineering for a system that runs real evaluations against frontier models.
- **Website & Evaluation Dashboard.** Building a tech demo, user-facing website and interactive dashboard for MIMIR, enabling researchers, developers, and enterprise users to explore evaluation results, compare model performance across benchmarks, and understand routing decisions in real time.
- **Output Handling & Result Formatting.** Developing structured output pipelines that transform raw evaluation results from multi-agent runs into clean, machine-parseable formats (JSON, structured reports) and building the tooling to track, version, and diff evaluation results over time.

WHAT WE'RE LOOKING FOR

- Strong backend or full-stack development skills (Python, TypeScript/JavaScript, or similar)
- Experience with API design, server setup, or web development; specialisation in any of these is fine
- Comfort with infrastructure tooling: Docker, cloud providers, CI/CD pipelines
- Interest in AI systems, LLMs, or developer tooling for ML
- Bonus: experience with LLM APIs, agent frameworks, or model orchestration
- Bonus: experience with data visualisation, dashboards, or real-time result streaming
- Bonus: familiarity with multi-agent coordination protocols or distributed systems

WHAT YOU'LL GAIN

- Direct ownership of the platform that researchers and enterprises use to evaluate real AI systems
- Experience building production infrastructure for multi-agent systems and learned routing
- Collaboration with ML researchers, systems engineers, and domain experts
- Publication and open-source contribution opportunities
- Fair compensation for paid positions; rates depend on experience and engagement model

HOW TO APPLY

We hire for curiosity, rigour, and genuine interest in the problem. If one of these roles sounds like your kind of work, even if you don't tick every box, we'd like to hear from you. Please send a short note explaining which role and which example project(s) interest you most, along with a CV or link to relevant work.

Contact: contact@mimir.fit | **Website:** mimir.fit