LLM Data & Evaluation Engineer – Alexandra Institute, Danish Foundation Models

Do you want to shape what makes Denmark's open, national language models truly good – as part of Danish Foundation Models? Do you love digging into data, building evaluations that are hard to game, and designing RL environments that teach models real skills? Then you might be our new colleague!

About the job

Few places in Europe let you decide what a large language model learns – and publish everything openly. Danish Foundation Models (DFM) is one of them. DFM is a government-funded consortium of the Alexandra Institute, Aarhus University, the University of Copenhagen and the University of Southern Denmark, building open-source language models and AI infrastructure for Denmark. At the Alexandra Institute's AI Lab, our team develops and releases DFM's open models. With our recent Munin 2.0 release – a 120B (12B active) mixture-of-experts model post-trained on the LUMI and Gefion supercomputers – we saw how much leverage careful post-training data and reinforcement learning give when working with a finite compute budget.

We are now looking for an LLM Data & Evaluation Engineer to shape what our upcoming models learn and how we evaluate them. You will work on post-training data mixtures, synthetic data, RL environments and reward design, and modern evaluation suites – both to improve our own models and to help build open, trustworthy evaluation of Danish and European language models for the wider community.

You will join a small, flat team working closely with our training engineers, where you own real technical decisions and see your data, environments and evaluations directly guide our next training runs.

As a GTS institute (a Research and Technology Organisation), we have a special role in Danish society. We act as independent advisors with a strong focus on open standards, open source and independent digital solutions – often with a research foundation and always with a responsible mindset. In the DFM project, you will collaborate with researchers at our partner universities as well as with companies and public organisations putting open Danish language models to use.

(More excited about distributed training, clusters and RL systems? Take a look at our companion opening: **LLM Post-Training & Systems Engineer**.)

Your future role

You will own what goes into Munin models and how we, and the wider community, know they got better – sitting right alongside the training engineers, so your data mixes, environments and evaluation insights directly move the models.

- **Evaluation & setting the release bar:** Build and own our evaluation stack on modern open standards like Inspect AI and Harbor – realistic end-to-end agentic evaluations, covering Danish, reasoning, tool use, instruction following, and safety. Hunt down contamination, saturation and blind spots, and define what "good enough to release" means.
- **RL environments & reward design:** Apply the same task, environment and verifier craft to RL training – designing interactive environments, automated verifiers and rubric-based rewards that provide reliable training signal.
- **Model diagnostics & error analysis:** Look closely at the traces. Run controlled ablations, inspect model trajectories and failure modes, and catch regressions or reward hacking early – giving the training engineers clear, evidence-based findings to act on.
- **What goes into the models:** Own the quality of our post-training data – designing SFT

and RL data mixtures, synthetic data generation, filtering and distillation that build new capabilities across Danish and English – in close collaboration with the training engineers, and with open, well-documented provenance (building alongside efforts like Danish Dynaword).

- **Open releases & research**: Publish open datasets, eval suites, technical write-ups, and, where relevant, research papers.
- **Bring open models into the real world:** Work with companies and public organisations on their use cases – advising on how to evaluate, adapt and put open Danish language models to use, and bringing what we learn back into our data and evaluations.

The competences we're looking for

We imagine that you:

- Have hands-on experience building and iterating on LLM data pipelines, synthetic data generation, RL tasks/environments or rigorous evaluation suites that improved or diagnosed real models.
- Are fluent in Python and modern open-source LLM tooling, such as running large-scale batch generation and evaluations with vLLM or SGLang, working with the Hugging Face ecosystem, and using or extending modern evaluation frameworks.
- Have sharp empirical instincts: you know how to design clean ablations, spot data contamination, metric gaming or reward hacking, and separate genuine capability improvements from benchmark noise.
- Are driven by strong scientific curiosity and the discipline to inspect raw data and model traces yourself, rather than treating data as a black box.
- Have experience with – or interest in – designing verifiable RL environments, agentic benchmarks, LLM-as-a-judge calibration or human evaluation protocols.
- Are familiar with open data licensing and copyright, and with domain-specific Danish/European text (e.g. public sector, legal, healthcare, technical).
- Enjoy talking with companies and organisations about their use cases and translating their needs into meaningful evaluations.
- Ideally have a public track record: open-source contributions, released open datasets, benchmarks or evaluation tools, technical blog posts or papers (a plus, but not a requirement).
- Communicate fluently in English and can explain complex concepts to both technical and non-technical audiences. Danish (or another Scandinavian language) is a big plus, as much of our data and evaluation work concerns Danish language and culture – but it is not a requirement.
- Already have the right to work in Denmark (see below).

We are open to different levels of experience and educational backgrounds - from strong recent graduates and PhDs to experienced engineers and researchers from industry or academia. What matters most is real hands-on depth and the drive to learn fast. You don't need to tick every box.

Your personality, passion and approach to work are just as important to us. You are self-driven and take ownership, but you also enjoy being part of a team, collaborating closely, exploring and trying out new technologies – and sharing knowledge with your colleagues.

Your future workplace

You will join the Alexandra Institute's AI Lab, a team of about 25 people split between Aarhus (Katrinebjerg) and Copenhagen (the IT University of Copenhagen). You will be based in Copenhagen, where the DFM team sits, with the option of working from home a day or two a week.

Expect a warm, informal atmosphere and colleagues who are exceptionally good at what they do. We work flexible hours, start every Friday with breakfast together, and offer a lunch scheme and health insurance. You'll have a lot of freedom and be trusted to use it

well, with plenty of room to grow, both professionally and personally.

Sounds interesting?

Then we would love to hear from you. Send us your CV and a short note (half a page is plenty) no later than **6 December**. In the note, tell us about a dataset, RL environment, evaluation framework or model failure analysis you have owned: what you did, what was hard, and what you learned. We don't need a formal cover letter; we would rather hear about your work. Links to GitHub, Hugging Face, blog posts or papers are very welcome.

First-round interviews will take place December 10th to 14th.

Second-round interviews will take place December 16th to 18th.

Please note: Due to our timeline, we can only consider candidates who already hold the right to work in Denmark (EU/EEA citizens or existing permit holders).

If you have questions about the position, you are welcome to contact Head of AI Lab Mie Hvas at mie.hvas@alexandra.dk / +45 20 63 99 58.

At the Alexandra Institute, we believe that innovation thrives when our employees have many different backgrounds – professionally as well as in terms of gender, religion, nationality and ethnicity. We therefore encourage all qualified candidates to apply, regardless of background.