

Generative DNNs at Unibas: imagination, expectations, and infrastructure

Dr. Geoffrey Fucile
sciCORE and the Center for Data Analytics
31.10.2024



Swiss Institute of
Bioinformatics



Swiss Tropical and Public Health Institute
Schweizerisches Tropen- und Public Health-Institut
Institut Tropical et de Santé Publique Suisse

sciCORE and the Center for Data Analytics (CeDA)

sciCORE – Center for Scientific Computing


- Core facility for entire Unibas ecosystem
- Compute infrastructure for research, including sensitive data platform (sciCORE+)
- Data management services and scientific programmers
- Scientific support services (training and consulting)

CeDA

- Core facility for entire Unibas ecosystem
- Engages primarily in research collaborations and consultations
- Additional activities in outreach and training
- Expertise in quantitative data analysis
- Academic anchor in Department of Math and Computer Science (Prof Ivan Dokmanic)

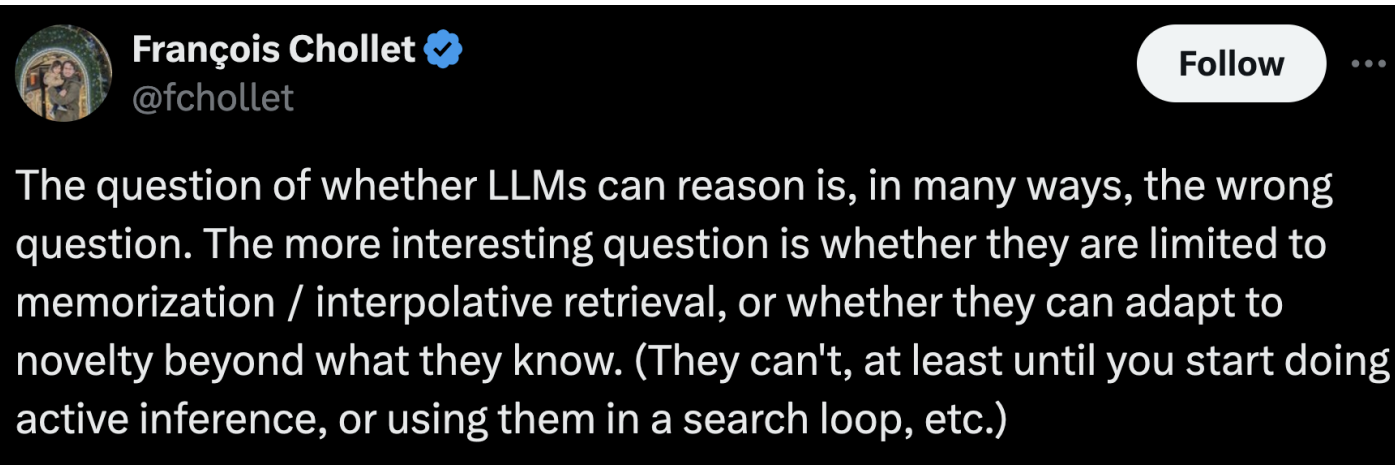
A brief scientific digression...



Speech and Natural Language Processing
Paper | October 2024

 Machine Learning Research

GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

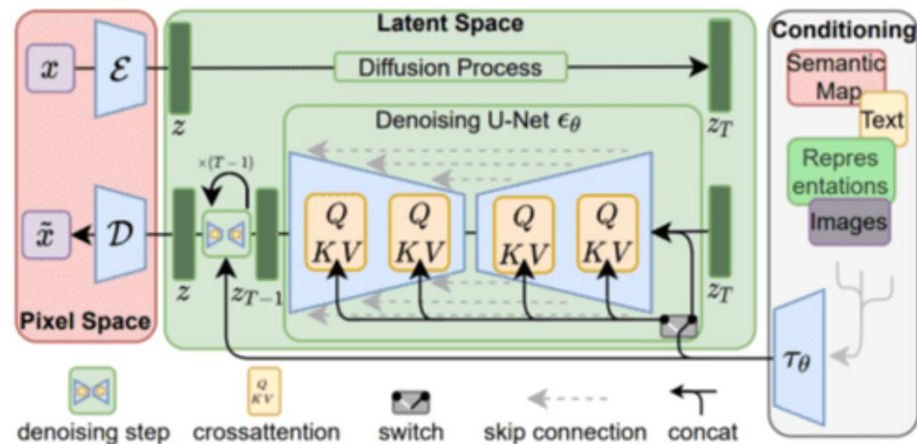
Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio,
Mehrddad Farajtabar



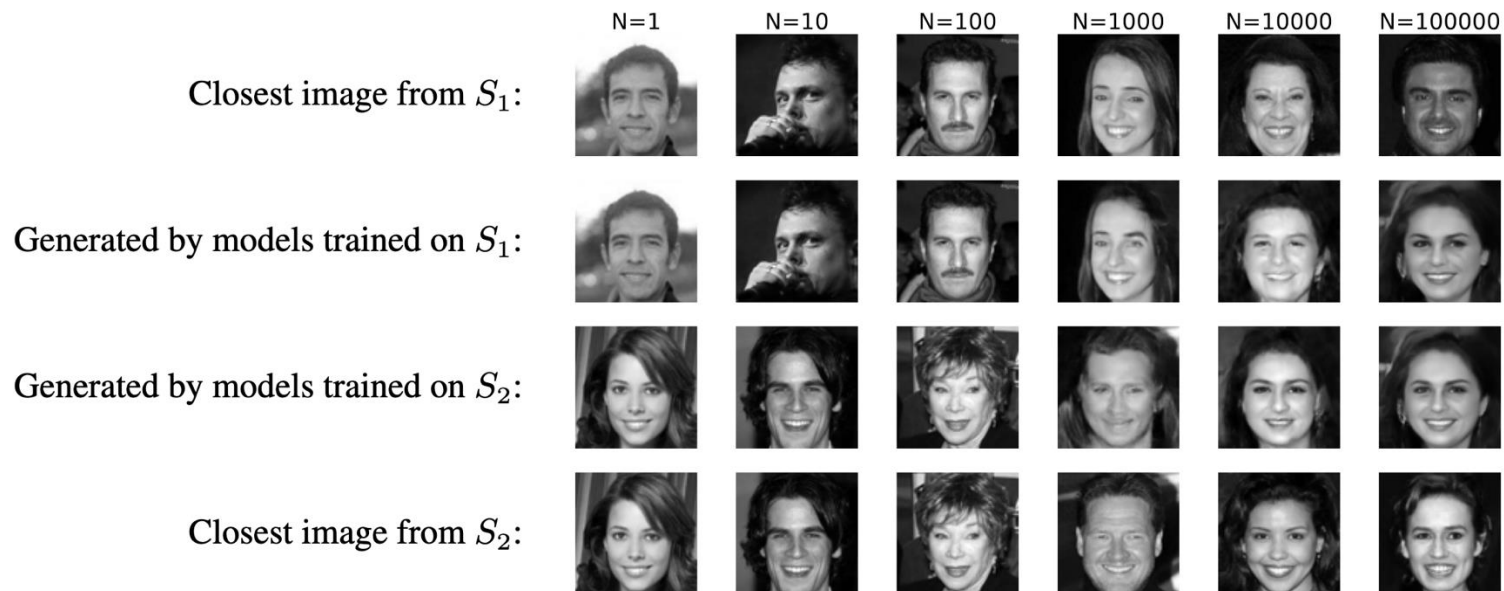
 **François Chollet** 
@fchollet Follow ⋮

The question of whether LLMs can reason is, in many ways, the wrong question. The more interesting question is whether they are limited to memorization / interpolative retrieval, or whether they can adapt to novelty beyond what they know. (They can't, at least until you start doing active inference, or using them in a search loop, etc.)

A brief scientific digression...

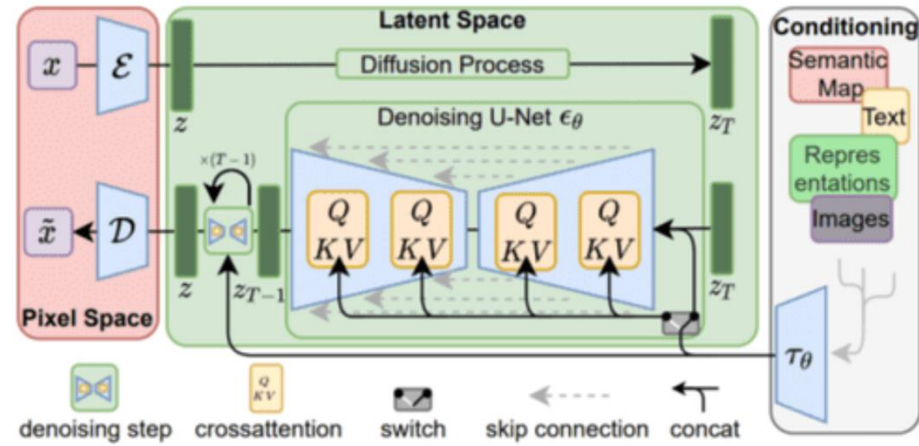


<https://huggingface.co/blog/Esmail-AGumaan/diffusion-models>

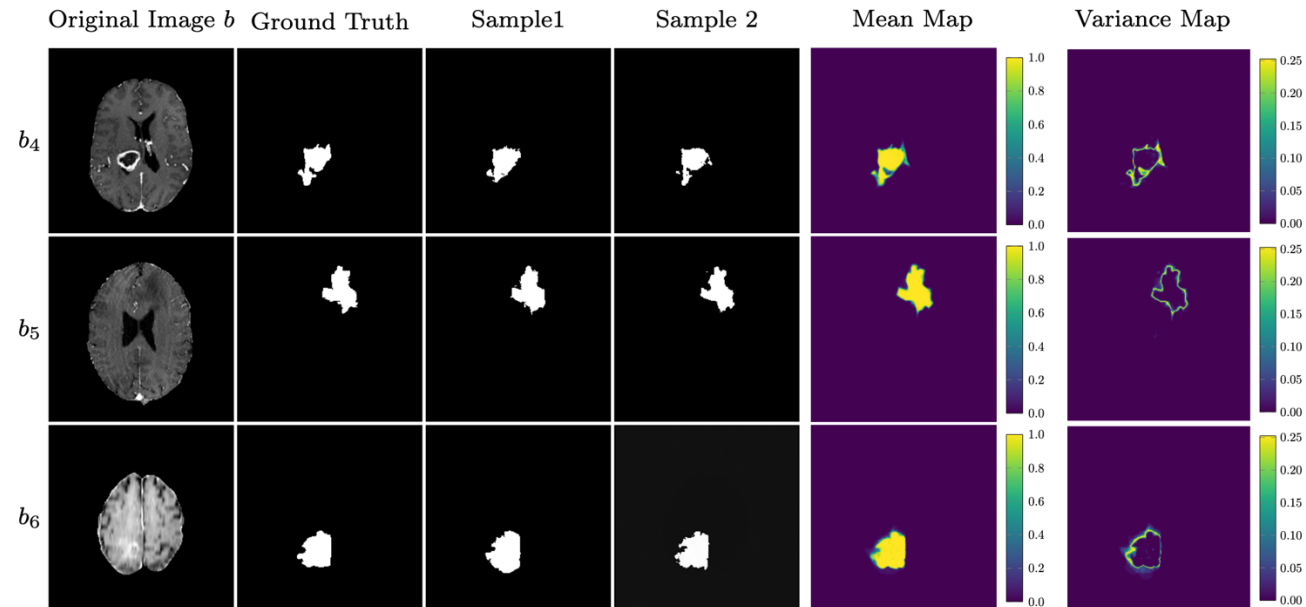


Kadkhodaie, Mallat et al 2024 <https://arxiv.org/pdf/2310.02557>

A brief scientific digression...



<https://huggingface.co/blog/Esmail-AGumaan/diffusion-models>



Wolleb, Cattin et al. Proceedings of Machine Learning Research 2022

Model collapse

”Synthetic data, although theoretically indistinguishable from real data, is almost always biased, inaccurate, not well representative of the real data, harmful, or presented out-of-context. Using such data as training data leads to issues with the quality and reliability of the trained model.”

https://en.wikipedia.org/wiki/Model_collapse

Article | [Open access](#) | Published: 24 July 2024

AI models collapse when trained on recursively generated data

[Iliia Shumailov](#) , [Zakhar Shumaylov](#) , [Yiren Zhao](#), [Nicolas Papernot](#), [Ross Anderson](#) & [Yarin Gal](#) 

Nature **631**, 755–759 (2024) | [Cite this article](#)

Strong Model Collapse

E Dohmatob, Y Feng, J Kempe
arXiv preprint arXiv:2410.04840, 2024

The University of Basel's AI initiative

As part of a holistic AI initiative, the University of Basel is addressing the challenges posed to our society by the advances in AI development through a comprehensive range of offers and targeted measures. The university's goal is to make a decisive contribution to ensuring that AI is used for the benefit of all. The university's AI initiative is committed to the motto "No One Left Behind".

<https://www.unibas.ch/de/Universitaet/KI-Initiative.html>

Some guiding principles:

- Efforts should be consolidated and coordinated in the spirit of efficiency
- Open source and development of on-prem expertise when possible
- Enable use of the best methods wherever needed
- Funding TBD

Use cases

- **Research**
 - AlphaFold, small molecule “synthesis”
 - Literature synthesis
 - Simulation of psychosomatic interactions
 - Models of risk perception
 - Genetic counseling
 - Cultural evolution from ancient Greek papyri fragments
 - Foundation models for molecular/cellular biology, seismology
 - Theories of ML
 - Code pilots
 - ...
- **Administration and teaching**
 - General chat-bot
 - Documentation navigation
 - Assistance with recruitment procedures
 - Image/Video/Sound \leftrightarrow Text for courses, communications/outreach
 - Access to specific models in exam environment (bring-your-own-device)

Use cases

- **Research**

- This is the easy case (although much still to optimize)
- Cluster has ~15 A100 nodes (40 and 80 GB, 4 cards per node)
- Requests submitted for H200 nodes and AMD products (MI300X)
- Solutions for storage, interactive computing (OpenOnDemand), etc
- Solutions for sensitive data (sciCORE+)

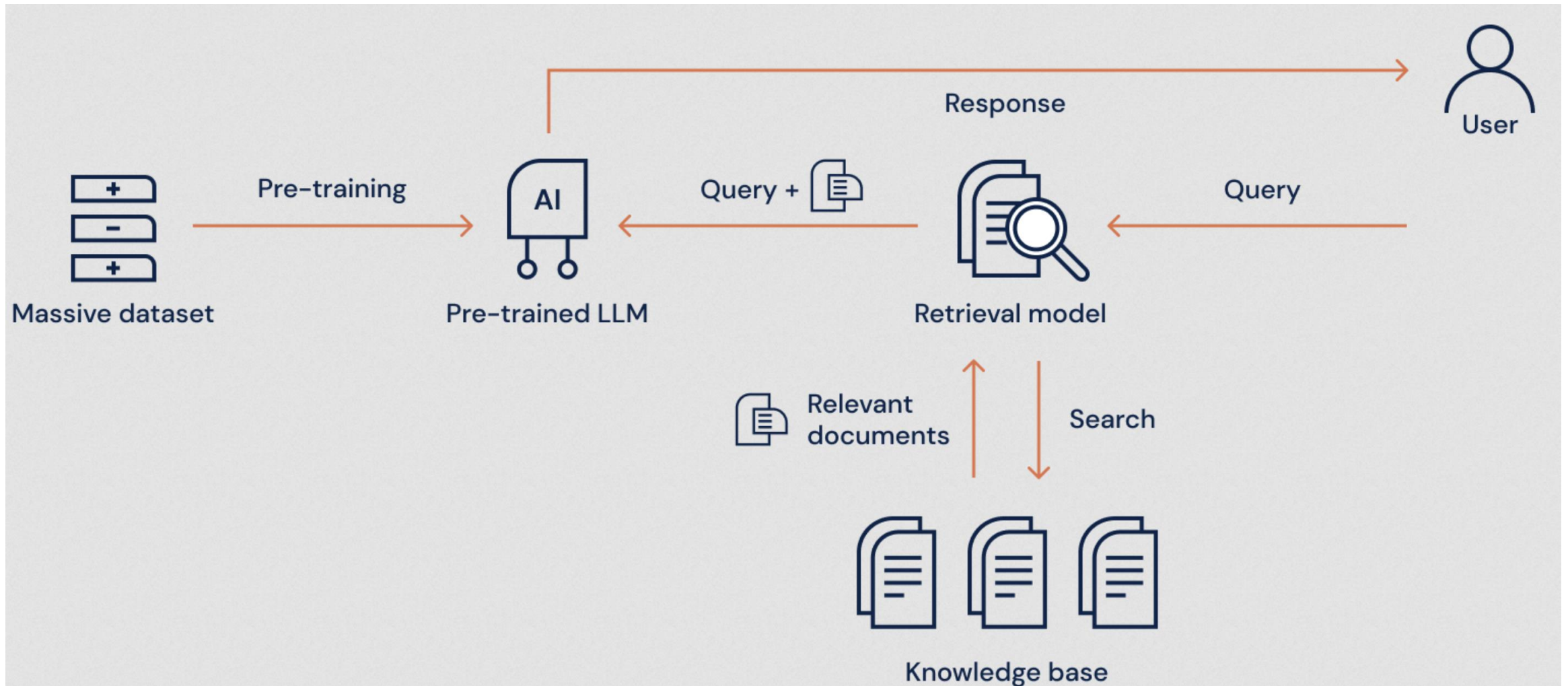
- **Administration and teaching**

- Much more complex
- Stability/availability and fail safes
- Front-end and API development
- Efficiency orchestration
- Quality control/assessment
- So far, no one has concrete requirements

Playground for development in admin/teaching space

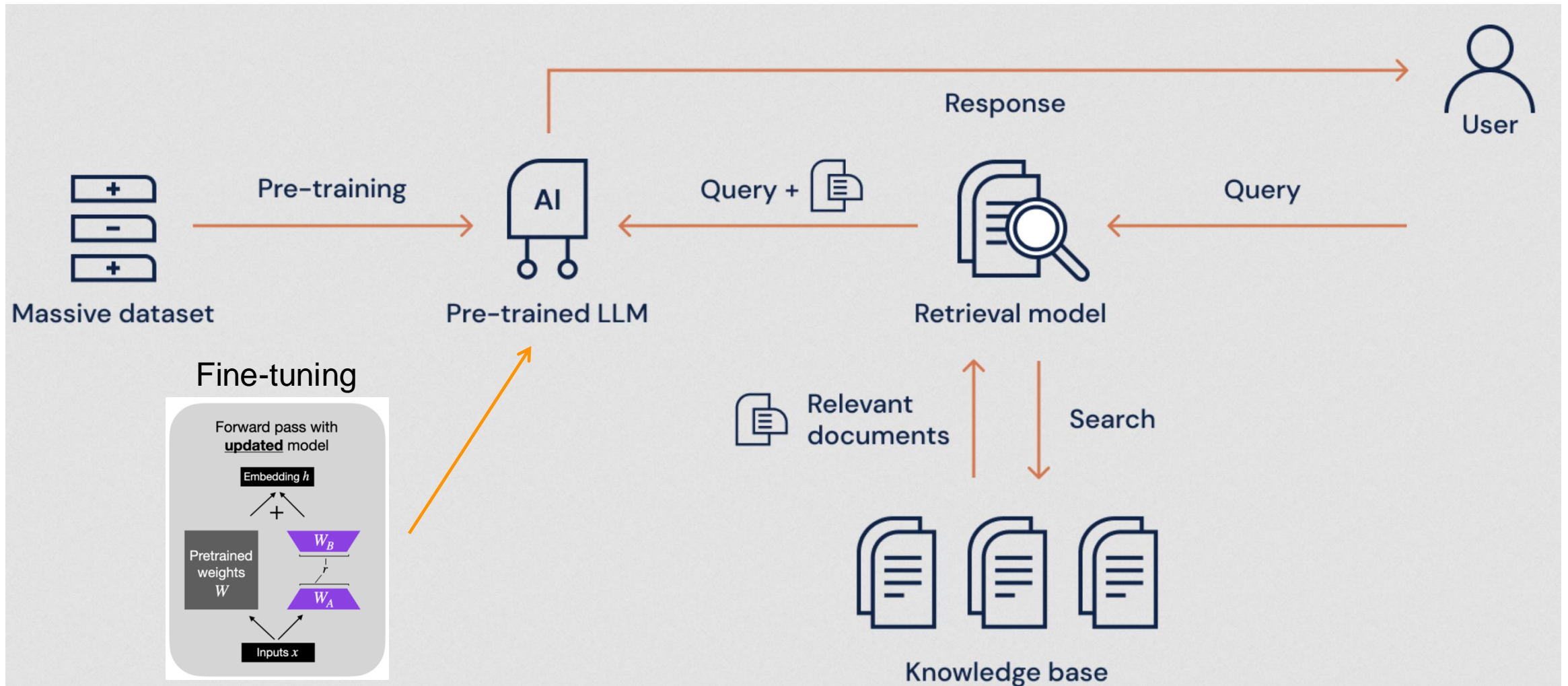
- Joint proposal between CeDA and IT-Services (enterprise IT)
- Create a space for test deployments towards defining concrete requirements
- Compute resources
 - H200 node (4X)
 - Storage and local scratch
- Consultant time
- In-kind contributions from CeDA and IT-S for config, organization, consultation
- Users come with competence to deploy

Retrieval Augmented Generation (RAG)



<https://www.6clicks.com/hubfs/Retrieval-Augmented%20Generation%20Process.png>

Retrieval Augmented Generation (RAG)



<https://lightning.ai/pages/community/lora-insights/>

<https://www.6clicks.com/hubfs/Retrieval-Augmented%20Generation%20Process.png>

Retrieval Augmented Generation (RAG)

Query:

How do I get a sciCORE account?

Response:

1. If you have a UniBasel e-mail address:

Go to <https://support.scicore.unibas.ch/> and login with your Uni Basel email credentials, then click on "Account creation" and fill out the form.

2. If you do not have a UniBasel e-mail:

Fill the form "[Antrag für eine Zugangsberechtigung zu universitären Informatikmitteln](#)" and contact the Genehmiger (approving officer) for your institute.

The following documents were most relevant to your query:

<https://wiki.biozentrum.unibas.ch/display/scicore/1.+sciCORE+accounts+in+Ubuntu+Cluster>

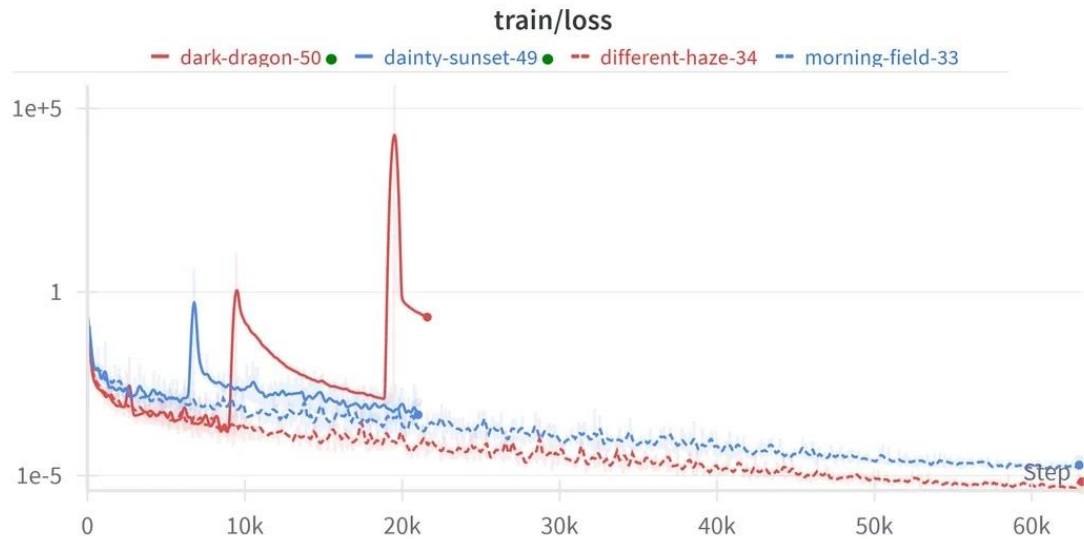
<https://wiki.biozentrum.unibas.ch/display/scicore/1.+sciCORE+accounts>

RAG example – Nvidia Llama 3.1 Nemotron 70B in memory

```
+-----+
| NVIDIA-SMI 555.42.02                Driver Version: 555.42.02          CUDA Version: 12.5          |
+-----+-----+-----+-----+-----+-----+
| GPU  Name                Persistence-M | Bus-Id                Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |      Memory-Usage     | GPU-Util  Compute M. |
|=====+=====+=====+=====+=====+=====+
|   0  NVIDIA A100-SXM4-80GB      On      | 00000000:2F:00.0 Off |         0          |
| N/A   48C    P0              93W / 500W | 70783MiB / 81920MiB |    59%    Default |
|                                           |                       |           Disabled |
+-----+-----+-----+-----+-----+-----+
|   1  NVIDIA A100-SXM4-80GB      On      | 00000000:B0:00.0 Off |         0          |
| N/A   47C    P0              271W / 500W | 68904MiB / 81920MiB |    51%    Default |
|                                           |                       |           Disabled |
+-----+-----+-----+-----+-----+-----+

Processes:
| GPU  GI   CI           PID  Type  Process name                GPU Memory |
|      ID  ID                |                | Usage   |
+-----+-----+-----+-----+-----+-----+
|   0  N/A  N/A           6610   G  /usr/lib/xorg/Xorg           107MiB |
|   0  N/A  N/A          2007808  C  python                       70656MiB |
|   1  N/A  N/A           6610   G  /usr/lib/xorg/Xorg            4MiB |
|   1  N/A  N/A          2007808  C  python                       68880MiB |
+-----+-----+-----+-----+-----+-----+
```

Pixtral 12B



Prompt:

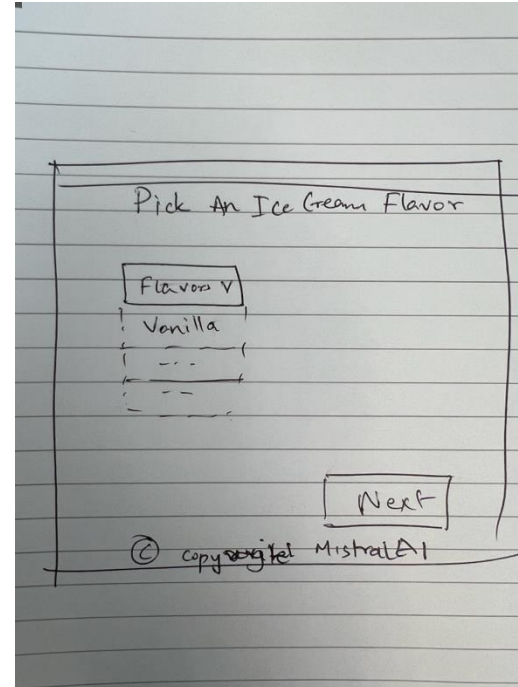
When did things start to go wrong for dark dragon?

Response:

The image is a line graph titled 'train/loss' that shows the training loss over steps for four different models: dark-dragon-50, dainty-sunset-49, different-haze-34, and morning-field-33. The x-axis represents the steps, ranging from 0 to 60k, and the y-axis represents the loss on a logarithmic scale from 1e-5 to 1e+5.

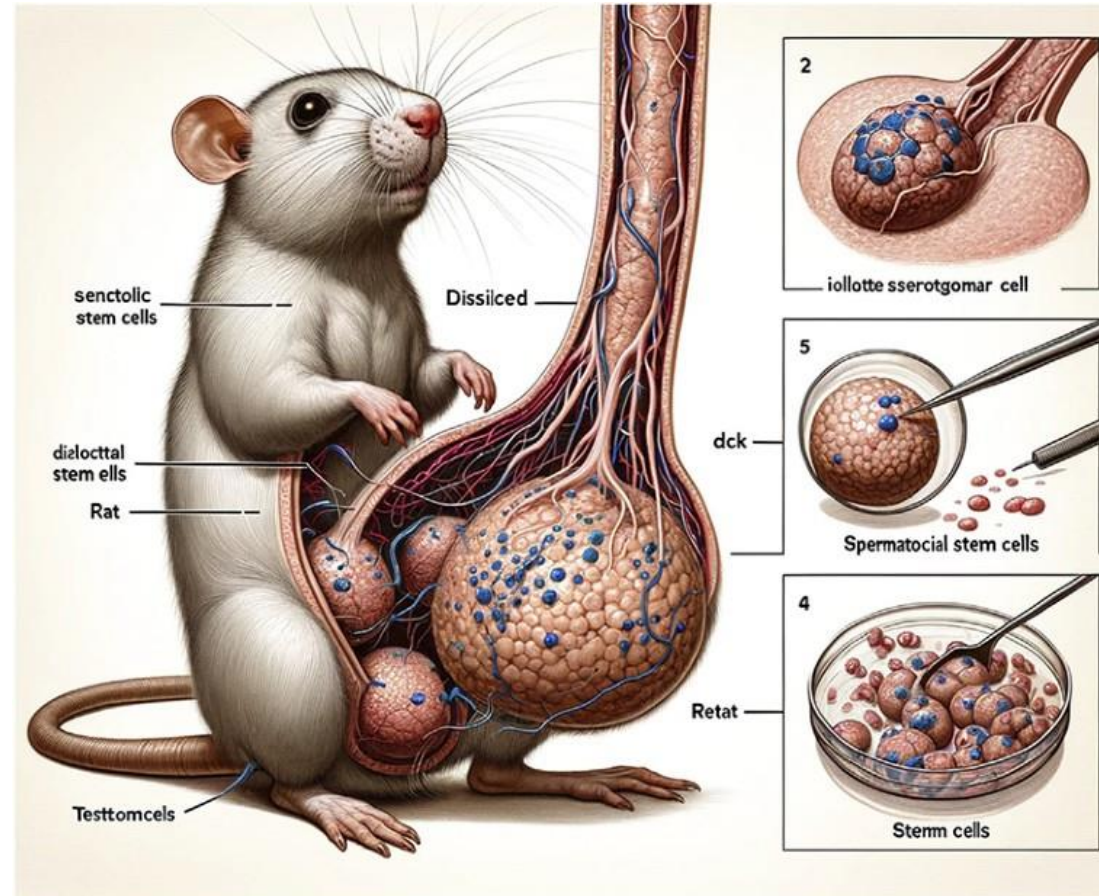
To determine when things started to go wrong for the dark-dragon-50 model, we need to look at the red line representing this model. Initially, the loss decreases, indicating that the model is learning effectively. However, around the 10k step mark, the loss starts to spike significantly. This spike indicates that the model's performance is deteriorating. The loss continues to spike and fluctuate, reaching very high values, which suggests that the model is not learning effectively and is likely overfitting or encountering other issues.

Therefore, the critical point where things started to go wrong for the dark-dragon-50 model is around the 10k step mark.



```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <title>Pick an Ice Cream Flavor </title>
  <style>
  body {
    font-family: Arial, sans-serif;
    display: flex;
    justify-content: center;
    align-items: center;
    height: 100vh;
    background-color: #f0f0f0;
  }
  .container {
    background: white;
    padding: 20px;
    border-radius: 8px;
    box-shadow: 0 10px rgba(0, 0, 0, 0.1);
  }
  h1 {
    text-align: center;
  }
  select {
    width: 100%;
    padding: 10px;
    margin-top: 10px;
    margin-bottom: 20px;
  }
  button {
    width: 100%;
    padding: 10px;
    background-color: #4CAF50;
    color: white;
    border: none;
    border-radius: 4px;
    cursor: pointer;
  }
  button:hover {
    background-color: #45a049;
  }
  </style>
</head>
<body>
  <div class="container">
    <h1>Pick An Ice Cream Flavor </h1>
    <select id="flavorSelect">
      <option value="vanilla">Vanilla </option>
      <option value="chocolate">Chocolate </option>
      <option value="strawberry">Strawberry </option>
      <option value="cookie_dough">Cookie Dough </option>
      <option value="mint_chocolate">Mint Chocolate </option>
    </select>
    <button onclick="next()">Next </button>
  </div>
  <script>
  function next() {
    const selectedFlavor = document.getElementById("flavorSelect").value;
    alert("You selected: " + selectedFlavor);
  }
  </script>
</body>
</html>
```

Stable Diffusion, DeepFloyd etc



<https://www.nature.com/articles/d41586-024-00659-8>

X. Guo *et al.* *Front. Cell Dev. Biol.*

Zoo of LLM “Orchestration” Software



LlamaIndex

Anything LLM

Open WebUI



MemGPT



Langfuse



LangChain



LM Studio



LiteLLM



gpt4all

LLM



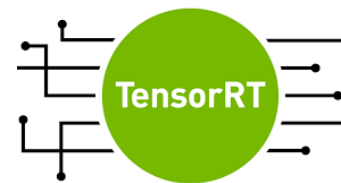
Chat UI



Ollama



LLaMA.cpp



TensorRT



airia

Enterprise AI Simplified



Kelly Sommers

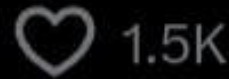
@kellabyte

Follow



Premature abstraction is WAY more prevalent and worse than premature optimization and really handcuffs code bases for years and years.

5:24 PM · Jul 18, 2024 · **116.2K** Views



To start ...

- Identify models that cover broadest range of use cases
- Build simple APIs that use protocols compatible with common tools (VSCode, Confluence, etc)
- Simplicity to adapt to rapidly changing landscape of models

Unanswered questions

- Abuse monitoring
 - Usage profiling
 - Cost models
 - Environmental impact assessment
-



University
of Basel

Thank you
for your attention.