Swiss Institute of Bioinformatics

Using Large Language Models for Biodata Exploration: From Theory to Practice

Ana Claudia Sima, Vincent Emonet,
Tarcisio Mendes, Panayiotis Smeros

kru@sib.swis

Schedule

- The theory : LLM basics
- Coffee break @ 10:30
- From Theory (closer to) Practice : LLM question answering over biodata •

Lunch @ 12:00

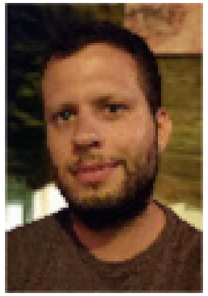
- Hands-on : RAG-based LLM application for biodata exploration

• Coffee break @ 15:00

Who we are – Knowledge Representation Unit @ SIB

• Tarcisio Mendes, Ana Claudia Sima – KRU co-leads (Lausanne / Zurich) • Vincent Emonet – Research Software Engineer, lead developer of ExpasyGPT

• Panayiotis Smeros, Research Scientist, broad experience in ML / LLMs for information retrieval, classification




What to expect from today

- An overview of LLMs, basics on how they are trained and potential uses

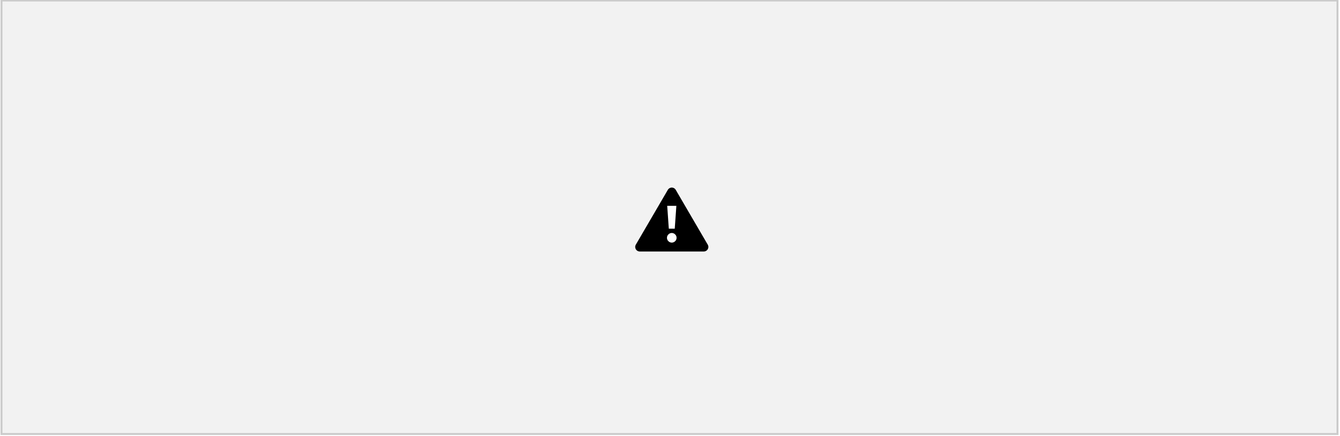
- A deep dive into using LLMs for question answering with context -
  From unstructured and structured sources (inputs)
  - Towards unstructured and structured answers (outputs)

- RAG-based application for interacting with biodata
  - Build your own mini "ExpasyGPT" !

Part I. The Theory

What is a Large Language Model?

A (deep learning) model trained to predict the next word in a sentence



How is it trained?

- Self-supervised learning on HUGE amounts of text
- Learning to predict 1 word at a time

Source: https://amitness.com

Large??

"Large" = number of parameters

- 8B, 70B, 200B…1 Trillion? (GPT 4o)

"Large" = amount of text seen

- GPT4: trained on 45TB (13 Trillion tokens!) of data (text, code)

Source: GKD: Generalized Knowledge Distillation for Auto-regressive Sequence Models

Closed-source LLMs: The GPT family

GPT- 4o (“research preview”)

Multimodal

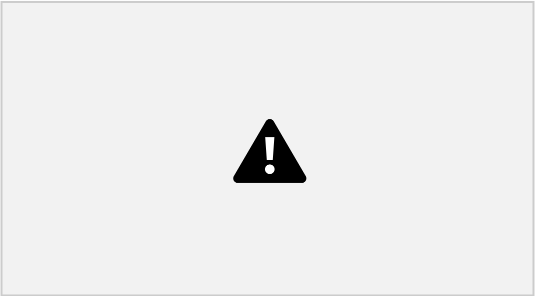
GPT- o1

Reasoning

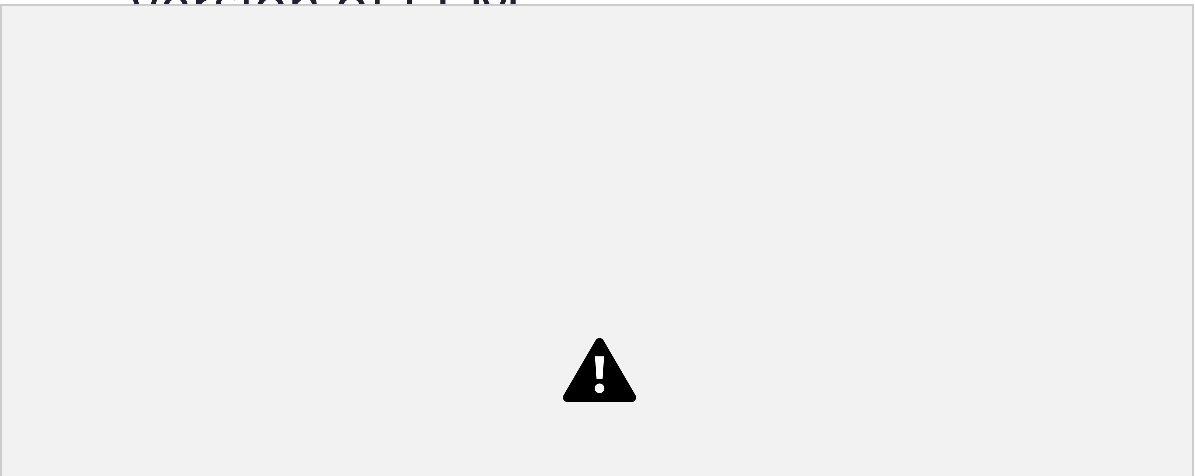
GPT- 4.1

GPT- 4.5

Source: https://generativeai.pub/



- Originally based on GPT3.5
- A "chatty" version of LLM



Note: ChatGPT is *not an LLM per-se, but a service with many "bells and whistles" leveraging an LLM*

(but also including web search, integration with external tools, etc)

The “art” of prompting

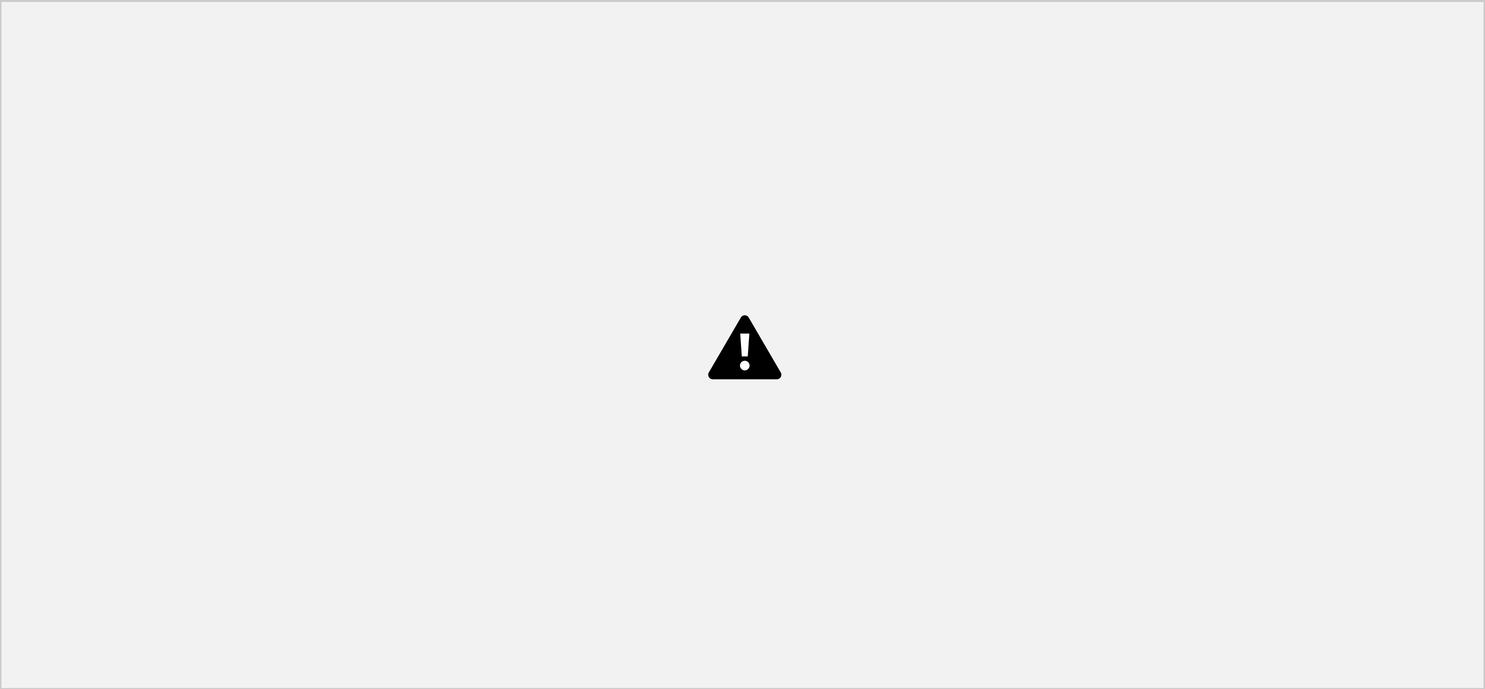
- Prompting = a guiding question / instruction given to the model to shape the generated response

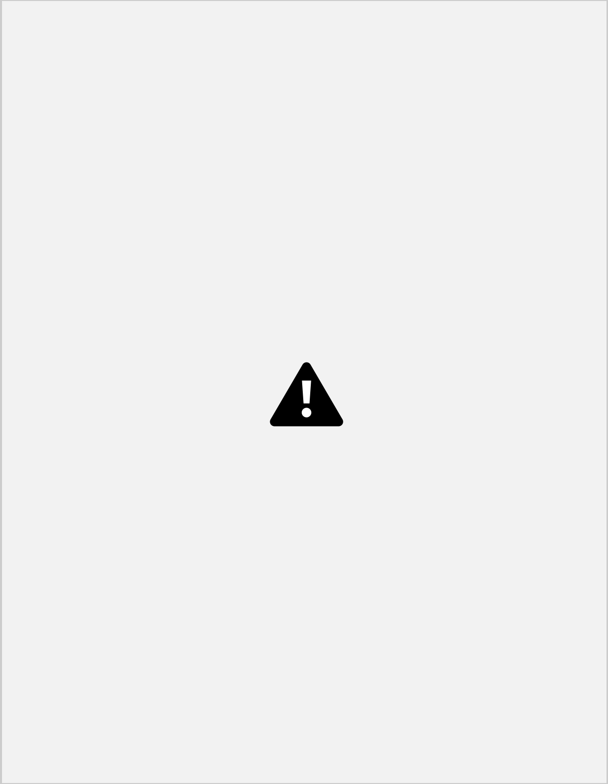
McDuff, D., Schaekermann, M., Tu, T. et al. Towards accurate differential diagnosis with large language models. Nature (2025). https://doi.org/10.1038/s41586-025-08869-4

The “art” of prompting

- Prompting = a guiding question / instruction given to the model to shape the generated response

- Brittle!





Vinay, Rasita, et al. "Emotional Prompting Amplifies Disinformation

Generation in AI Large Language Models." *Frontiers in Artificial Intelligence* 8: 1543603.

What can you use an LLM for?

- Summarisation
- Programming (Co-Pilot)
- Question Answering
- Reports from images
- ...

- Recognise special pieces of information in a text ◦ "Named Entity Recognition"
- Sentiment analysis
- Fraud detection
- ...

Classification tasks

Generative tasks

Key LLM training steps

2) Fine-tune

4) Validate / Evaluate

Tokenization

- Breaking down sentences into *tokens*

- The totality of tokens = the *vocabulary* of an LLM
    - Determined by the specific choice of *tokenizer*

- Tokens do *not* necessarily always correspond to *words*
  - *Why?*

What is a Large Language Model? (Generalization)

- A (deep learning) model trained to **predict** the next **word** in a **sentence**
  - Token = words / amino-acids / genes / ....
- What is a Language?
  - A **vocabulary** +

- **Sequences of tokens** that represent **information** in that language

Pre-training: 2) embedding

- LLMs, like any machine learning system, work with *numbers*
- *Embeddings = a way to transform words into numbers, while preserving their semantics*

Source: https://dkharazi.github.io/notes/ml/nlp/embedding

Pre-training: Attention!

Source: A Multiscale Visualization of Attention in the Transformer Model, ACL 2019 Systems Demonstrations

“Self-Supervised” / “Unsupervised” – but training data *matters*

- The illusion of self-supervision – actually a lot of effort for curating good data
  - Relies on work of thousands of human “data annotators”
  - Ethical issues: breaches of copyrighted material
  - Removing bias from training dataset
  - etc

Result: Foundation Model

Source: https://viso.ai/deep-learning/foundation-models/

Non-exhaustive list of pitfalls and limitations…

- Hallucinations

- When it's wrong, it's confidently wrong!

- Cutoff date
  - An LLM has no knowledge past its training date

ANY SOLUTIONS?

  - Example: GPT 4.1 (April '25) has knowledge up to June '24
- COST!
  - 100 Million $$$ to develop!
  - Environmental considerations…
- PRIVACY!
  - Please don't put sensitive information in ChatGPT…
- Lack of interpretability
- Lack of provenance / attribution
- Bias
- ….

Open-source LLMs: Llama, Mistral… •

Usually come in various size, the smaller the cheaper to run, but less “smart” �
~7B: inference can run on your laptop in acceptable time if you have a GPU
� **~70B**: requires a larger server for decent inference time (1 A100/H100) � **~400B**: largest, requires custom supercomputers (~4 H100s or more)

- Mixture-Of-Experts approaches such as Mixtral 8x7B or 8x22B combine multiple smaller models for better results with same hardware requirements

- HuggingFace: the most comprehensive resource to find open-source models
  - e.g. https://huggingface.co/meta-llama/Meta-Llama-3-8B

- Initially based on GPT3.5
- A "chatty" version of LLM
  - Based on Instruction DataSet (small, e.g. 50K etc) – Alpaca Style Prompt Template
- "Secret sauce": Reinforcement Learning through Human Feedback (**RLHF**)

We trained an initial model using supervised fine-tuning: ***human AI trainers provided conversations in which they played both sides—the user and an AI assistant****. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.*

Source: https://openai.com

Reinforcement Learning through Human Feedback

30

RLHF – an ongoing journey…

Crowdsourcing expertise?

Retrieval Augmented Generation (RAG)

- Complement the base model with *external knowledge base*

- This can be:
  - PDFs
  - Structured Data
    - CSVs
    - Relational DBs
    - ...
  - Images
  - ....

Source: https://blog.gopenai.com

Fine-tuning a foundation model

- Requires a dataset of examples in the style of “instructions”, input and output

- Note: usually **expensive** to do

Part II. From Theory (Closer To) Practice

How can you use LLMs to explore biodata?

- Directly ask questions (“zero-shot”)

- Formulate code (e.g. API calls) to answer questions

- Answering questions with context
    - Structured data (CSV files)
    - Unstructured data (PDF files)

Ask LLM questions over biodata directly

“Cheating!”

Disable and try again!

Asking chatGPT question directly ("zero-shot")

Warning: in general, asking LLMs specialized questions directly will likely lead to hallucinated answers!

Recommended: Ask Model To Ask an API

Recommended: Ask Model To Ask an API

Analysing local files?

(LLMs with context)

Example: the SIB resources CSV

Download file from here

Use ChatGPT for QA over CSV

Use ChatGPT for QA over CSV

Use ChatGPT for QA over CSV

Always

Verify!

Using LLM for QA over PDF?

Using ChatGPT for QA over complex PDF

Use an example ESMO clinical guideline, ask for a description of some of the figures or an analysis of a decision tree in one of the figures

E.g. *[Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up]*

Question Answering over Structured Data

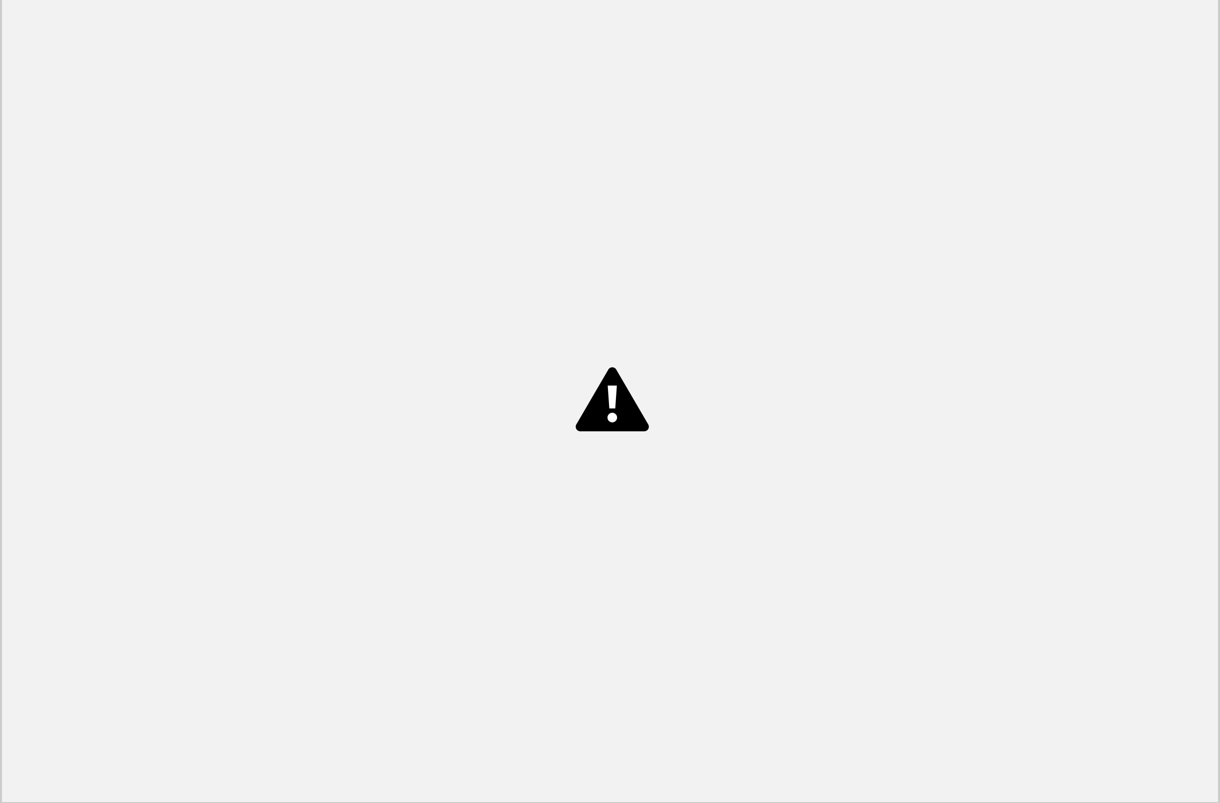
Expasy

Knowledge Graphs curated by the SIB

What is a Knowledge Graph?

- Graph data model
  - Easy to **interlink** and extend
  - Data usually stored as triples
  - Triple = <subject, predicate, object>

  - E.g. "*<gene X> isExpressedIn <anatomic entity Y>*"



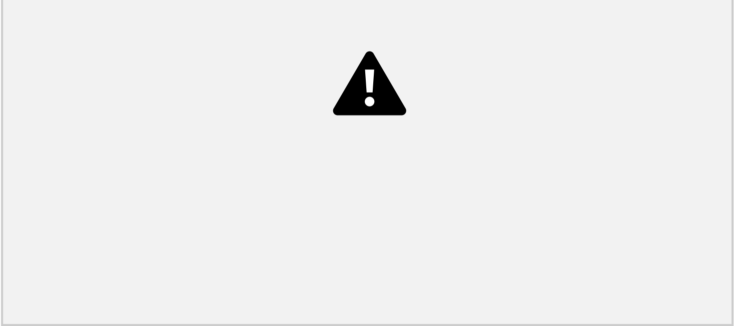
- Can be queried using SPARQL:
  - Select * where {

```
        ?gene isExpressedIn ?anatEntity.
}
```

- Play an important role in **Semantic Interoperability**

Why are these useful?

- Explicit semantics through: - Standard identifiers
- Shared vocabularies
- Ontologies



Data Integration and FAIRification

- Ask complex questions that span across disciplines and across datasets

What are the human genes involved in lung cancer with an ortholog expressed in the mouse?

Research question: LLMs + KGs = ?

How do we extract information from RDF data?

Retrieval Augmented Generation (RAG)

Pan, Shirui, et al. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." ***arXiv preprint arXiv:2306.08302*** **(2023).**

```
SELECT ?gene ?orthologous_protein2 WHERE {
 SELECT * {
 SERVICE <http://sparql.uniprot.org/sparql> {
        SELECT ?protein1 WHERE {
        ?protein1 a up:Protein;
               up:organism/up:scientificName 'Homo sapiens' ;
               up:annotation ?annotation .
               ?annotation rdfs:comment ?annotation_text.
               ?annotation a up:Disease_Annotation .
                FILTER CONTAINS (?annotation_text, ”lung cancer")
        }
 }

 SERVICE <https://sparql.omabrowser.org/sparql/> {
        SELECT ?orthologous_protein2 ?protein1 ?gene WHERE {
               ?protein_OMA a orth:Protein .
               ?orthologous_protein2 a orth:Protein .
               ?cluster a orth:OrthologsCluster .
               ?cluster orth:hasHomologousMember ?node1 .
               ?cluster orth:hasHomologousMember ?node2 .
               ?node2 orth:hasHomologousMember* [.......]
                FILTER(?node1 != ?node2)
        }
 }

 SERVICE <https://bgee.org/sparql/> {
        ?gene genex:isExpressedIn ?anatEntity .
?anatEntity rdfs:label ‘lung' .
?gene orth:organism ?org .
?org obo:RO_0002162 taxon:10090 .}
```

What are the human genes involved in lung cancer with an ortholog expressed in the mouse?

LLMs for Biodata Exploration in Practice:
ExpasyGPT

⚠ Problem

Writing SPARQL queries is hard and time-consuming. LLMs are great at it, but they need context

For complex questions finding the right context is not trivial.

We need a minimal structured way (metadata) to describe SPARQL endpoints.

System architecture

Some Intermediate Conclusions…

- The LLM space is constantly evolving…
  - Tasks that seemed hard 6 months ago are now solved
  - Powerful interfaces to both structured and unstructured data…
  - …but need to be adapted to your use case

- Careful when using (especially) closed-source LLMs
  - Don’t use them for sensitive data!

- Infrastructure is a problem
  - Self-hosted (requires GPUs) or hosting platforms (paid)
  - Small models are (clearly) less performant
  - But big models are expensive to host

References

- EMBL-EBI webinars on LLMs: https://www.ebi.ac.uk/training/events/large-language-models-and-their-applications-bioinformatics/
- Developing an LLM: Building, Training, Fine-tuning (relatively technical video): https://www.youtube.com/watch?v=kPGTx4wcm_w - Sebastian Raschka, author of "Build a Large Language Model (from scratch)"
- Detailed notebook for running LLMs for various text analyses (includes analysis of next word probability, embeddings generated etc): https://colab.research.google.com/github/michael-franke/npNLG/blob/main/neural_pragmatic_nlg/07-LLMs/07b-pretrained-LLMs.ipynb
- Interactive overview of existing Language Models (from BERT to GPT3): https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/?trk=article-ssr-frontend-pulse_little-text-block
- Will we run out of data? Limits of LLM scaling based on human-generated data: https://arxiv.org/pdf/2211.04325, interesting preprint analysing the limits of scaling beyond current model sizes due to the lack of sufficient additional training data in the near future
- Other LLMs to interact with: Gemini, claude, …
- Language models for biological research: a primer, Nature Methods, '24
- Extended LLM course with multiple hands-on notebook examples: https://github.com/mlabonne/llm-course
- https://github.com/ncbi-nlp/LLM-Medicine-Primer
- McDuff, D., Schaekermann, M., Tu, T. *et al.* Towards accurate differential diagnosis with large language

models. *Nature* (2025). https://doi.org/10.1038/s41586-025-08869-4

- Vinay, Rasita, et al. "Em~~otional~~ ~~Prompting Amplifies Disinformation G~~eneration in AI Large Language Models." *Frontiers in Artificial Intelligence* 8: 1543603.

Thank you!

DATA SCIENTISTS FOR LIFE

Ana-claudia.sima@sib.swiss

Part III. Hands-on: Setup

Install: uv, qdrant, get (Mistral) API key