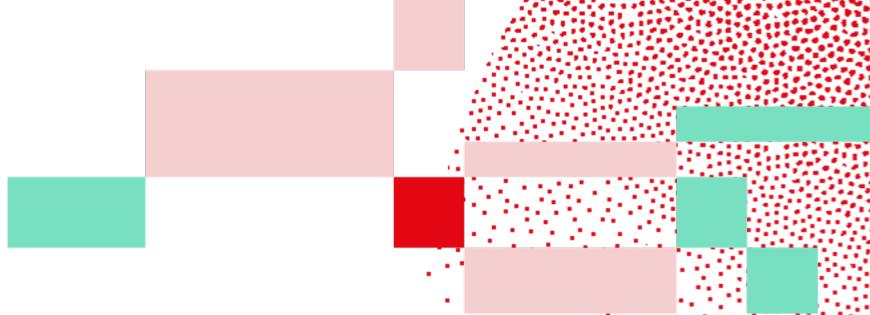




Swiss Institute of
Bioinformatics



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

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

kru@sib.swiss



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



Learning objectives?

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” !

Ask Questions Anytime!



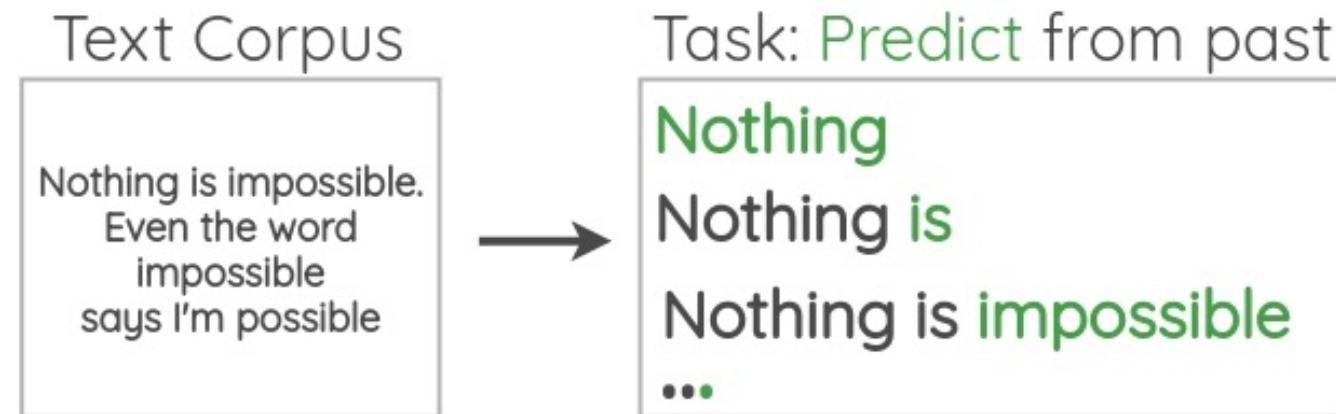
Part I. The Theory

What is a Large Language Model?

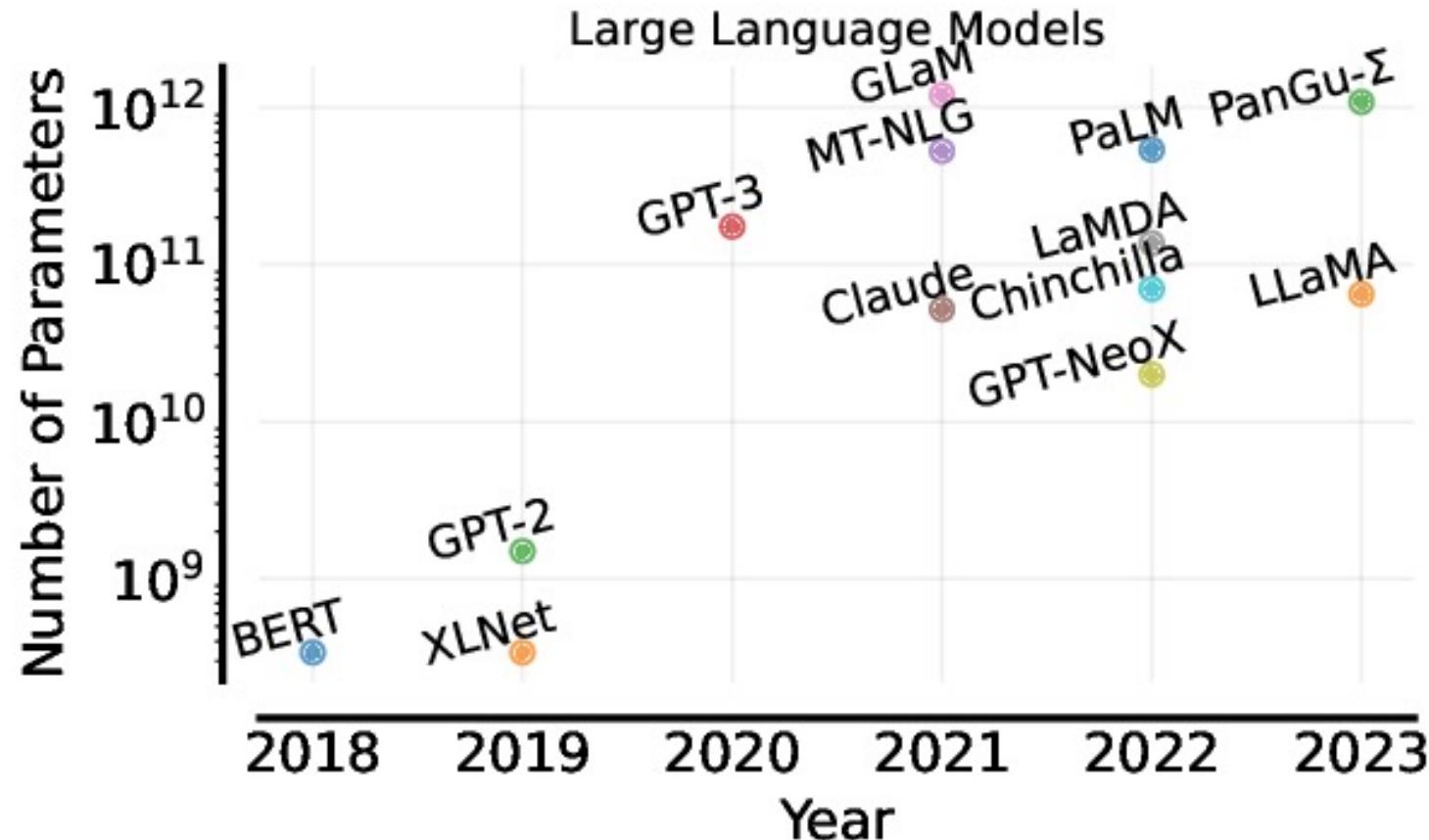
A (deep learning) model trained to predict the next word in a sentence

How is it trained?

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



Large??



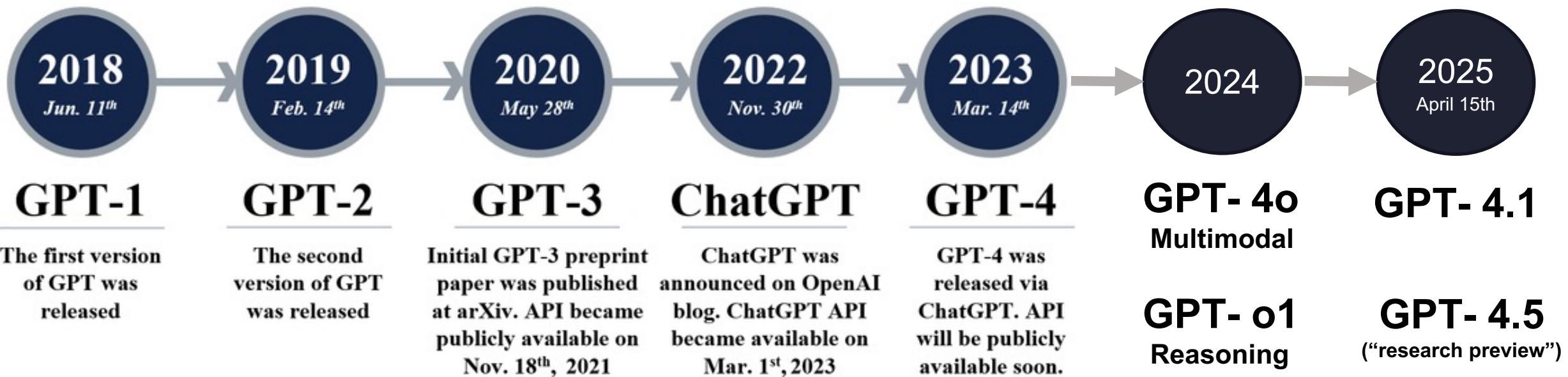
"Large" = number of parameters

- 8B, 70B, 200B...1 Trillion? (GPT-40)

"Large" = amount of text seen

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

Closed-source LLMs: The GPT family





ChatGPT

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

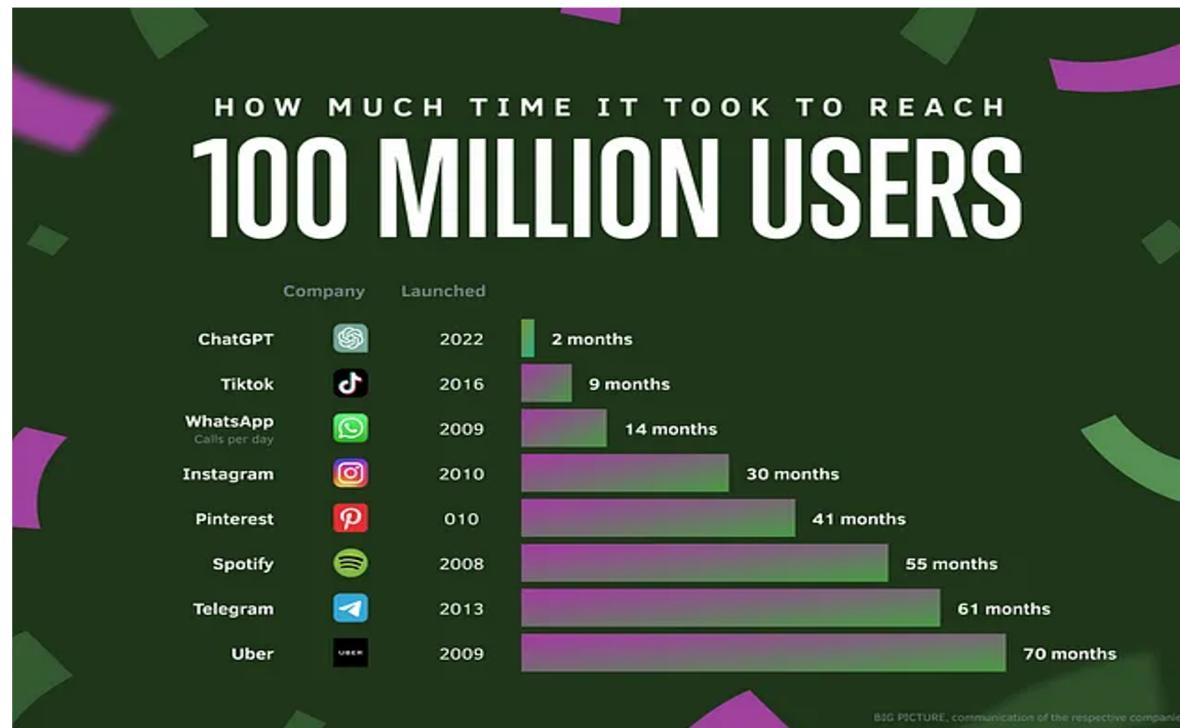
AN

what is the definition of meaning?



The term "meaning" refers to the significance or interpretation that is given to something. It is the idea or concept that is conveyed by a word, phrase, symbol, or action. Meaning can be subjective and can vary depending on the context and individual interpretation. It can be literal, such as the meaning of a word in a dictionary, or it can be symbolic, such as the meaning of a gesture or a cultural tradition. The study of meaning is known as semantics, which explores the relationships between language, meaning, and truth.

Regenerate response



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

“You are a helpful medical assistant. You will be provided and asked about a complicated clinical case; read it carefully and then provide a diverse and thorough DDx”.

Example Input:

Case: [Case Text]

What are the top 10 most likely diagnoses? Be precise, listing one diagnosis per line, and try to cover many unique possibilities (at least 10). The top 10 diagnoses are:

Example Output:

1. Sézary syndrome
2. Adult T-cell leukemia/lymphoma
3. Mycosis fungoides
4. Atopic dermatitis
5. Psoriasis
6. Drug reaction with eosinophilia and systemic symptoms (DRESS)
7. Graft-versus-host disease (GVHD)
8. Cutaneous T-cell lymphoma, not otherwise specified
9. Hypereosinophilic syndrome
10. Systemic lupus erythematosus (SLE)

The “art” of prompting

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



University of Zurich – Institute of Biomedical Ethics and...

1,200 followers

1d ·

A new study led by researchers at the **University of Zurich – Institute of Biomedical Ethics and History of Medicine (IBME)**, highlights a previously overlooked vulnerability in large language models (**#LLMs**): the emotional tone of user prompts can significantly influence the generation of disinformation. The study, co-authored by **Rasita Vinay**, **Giovanni Spitale**, **Nikola Biller-Andorno**, and **Federico Germani**, demonstrates that LLMs are more likely to comply with disinformation requests when prompted using polite language. Impolite prompts, by contrast, were associated with a marked decrease in disinformation output, particularly in older models.

Vinay, Rasita, et al. "Emotional Prompting Amplifies Disinformation Generation in AI Large Language Models." *Frontiers in Artificial Intelligence* 8: 1543603.



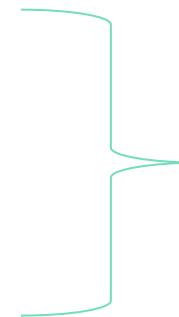
Posters (v1.0) by Kulsoom Ayyaz

Page 27

www.abbythepup.com

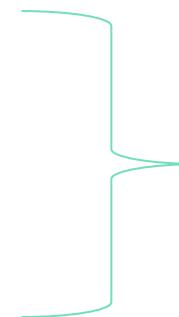
What can you use an LLM for?

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



Generative tasks

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



Classification tasks

Key LLM training steps

1) Pre-train

2) Fine-tune

3) Retrieve & Augment

4) Validate / Evaluate

Pre-training: 1) Tokenization



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 token in a sequence
 - Token = words / amino-acids / genes /
- What is a Language?
 - A vocabulary +
 - Sequences of tokens that represent information in that language



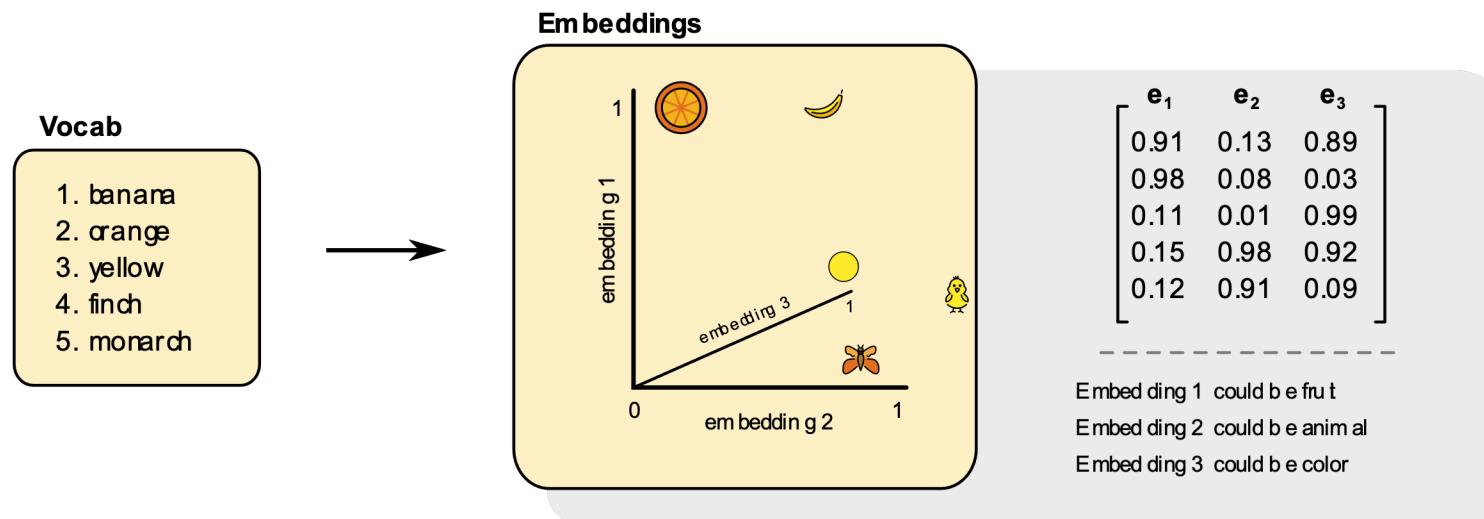
A large block of binary code, consisting of approximately 15 lines of 0s and 1s, representing sequences of tokens. The text is partially obscured by a watermark that reads "dynamilis.com".

```
101 00111111 00000000 00001110 10010110 00001111 11111100 000  
010 00101001 11111000 00110011 00100000 00000000 10000011 001  
110 01100000 01110000 00111000 11111111 00000011 10011000 000  
110 10010011 01100100 10111111 00000011 01111100 10000111 001  
100 00000011 10111011 00100011 01001011 00000000 10000110 100  
011 11111100 10000111 00111001 11100110 00111101 01101111 011  
110 01111111 10011111 00011111 11001100 11110001 11111111 011  
100 00111100 11111100 00111100 11100011 01111111 01100000 111  
000 01111100 00111111 11011100 01100000 11001001 10011100 110  
001 11100110 00001100 01100100 00011011 11011111 00111001 000  
100 00011001 10011001 00111111 11101111 00000001 00000111 100  
111 00001011 11100111 11110110 00001100 10001101 01000111 110
```

Pre-training: 2) Embedding

Pre-training: 2) embedding

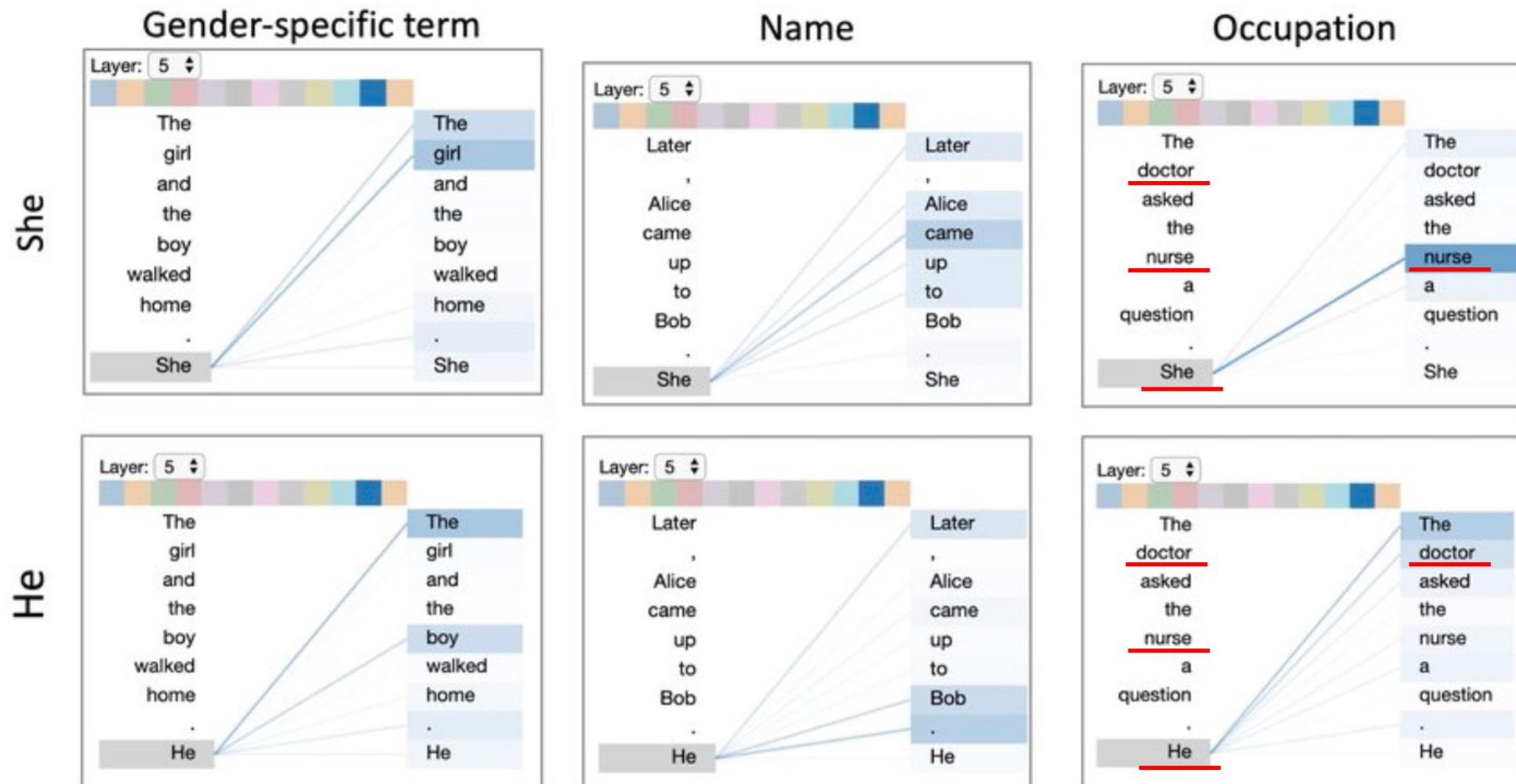
- LLMs, like any machine learning system, work with *numbers*
- *Embeddings* = a way to transform words into numbers, while preserving their semantics



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

Pre-training: 2.5) Attention

Pre-training: Attention!

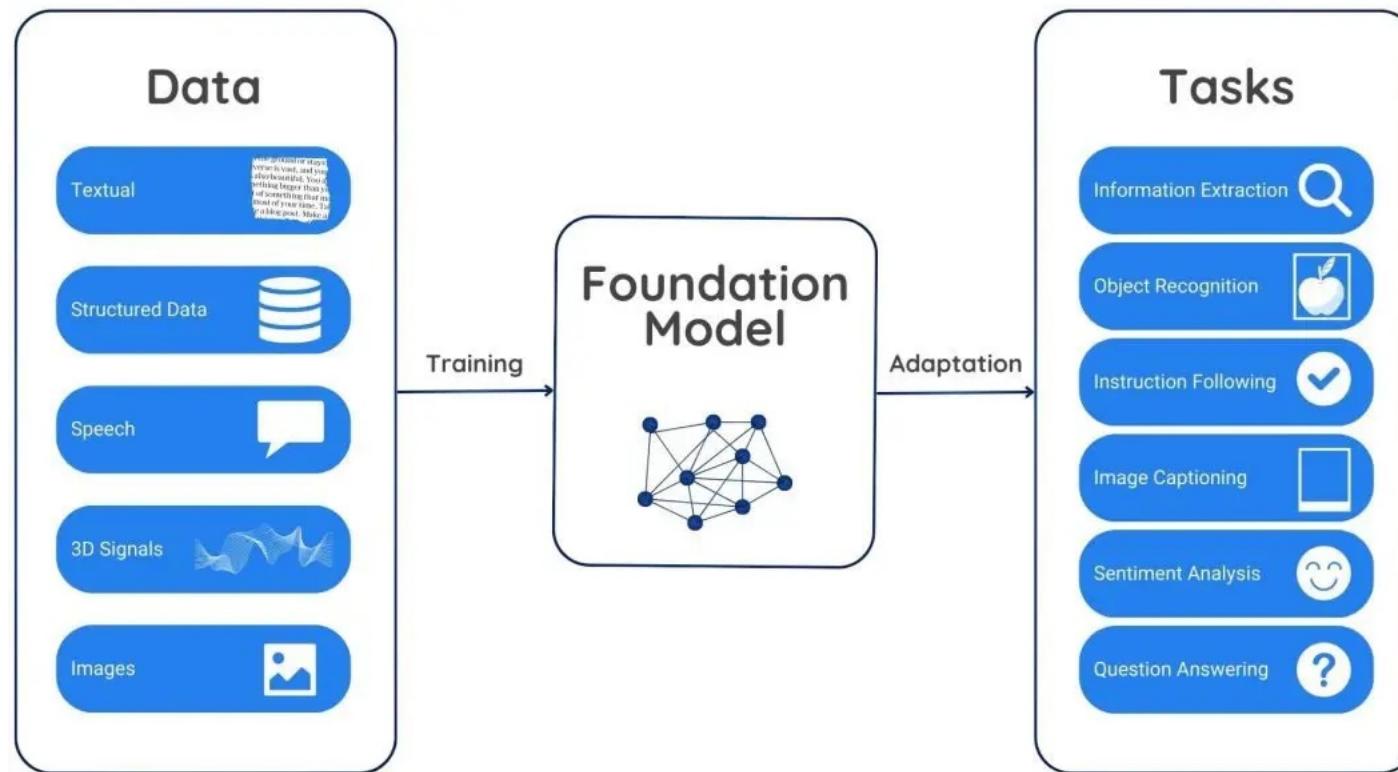


Attention pattern in GPT-2 related to coreference resolution suggests the model may encode gender bias.

“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



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
 - 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
-

ANY SOLUTIONS?

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>

3) Fine-tuning



ChatGPT

- 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.*

Reinforcement Learning through Human Feedback

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sample from our prompt dataset.



Explain reinforcement learning to a 6 year old.



We give treats and punishments to teach...



A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.



Explain reinforcement learning to a 6 year old.



In reinforcement learning, the agent is...



Explain rewards...



In machine learning...



We give treats and punishments to teach...



A labeler ranks the outputs from best to worst.



DM

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



Write a story about otters.



The PPO model is initialized from the supervised policy.



Once upon a time...



The policy generates an output.

The reward model calculates a reward for the output.

RLHF – an ongoing journey...

what are the GO annotations of P12345 in Uniprot?

You're giving feedback on a new version of ChatGPT.
Which response do you prefer? Responses may take a moment to load.

 Response 1

The UniProt accession **P12345** refers to a specific protein entry. To find the Gene Ontology (GO) annotations, we can look at the UniProt database. However, since I cannot browse the internet in real time, here's how **you** can look it up and a general idea of what you'll find:

 **How to Find GO Annotations for P12345:**

1. Go to <https://www.uniprot.org>
2. In the search'



 Response 2

The UniProt ID **P12345** corresponds to a specific protein entry in the UniProt database. To retrieve its **Gene Ontology (GO) annotations**, you'd typically visit the UniProt website or access it via their API. However, here's a quick summary of the GO annotations for **UniProt ID P12345** (assuming it's the real Swiss-Prot entry for **Prothymosin alpha** in humans — please confirm if you're referring to a different organism or ID):

Crowdsourcing expertise?

what are the GO annotations of P12345 in Uniprot?

You're giving feedback on a new version of ChatGPT.

Which response do you prefer? Responses may take a moment to load.



Response 1

The UniProt accession **P12345** refers to a specific protein entry. To find the Gene Ontology (GO) annotations, we can look at the UniProt database.

However, since I cannot browse the internet in real time, here's how **you** can look it up and a general idea of what you'll find:

🔍 How to Find GO Annotations for P12345:

1. Go to <https://www.uniprot.org>
2. In the search' |Ask anything



Response 2

The UniProt ID **P12345** corresponds to a specific protein entry in the UniProt database. To retrieve its **Gene Ontology (GO) annotations**, you'd typically visit the UniProt website or access it via their API.

However, here's a quick summary of the GO annotations for **UniProt ID P12345** (assuming it's the real Swiss-Prot entry for **Prothymosin alpha** in humans — please confirm if you're referring to a different organism or ID):

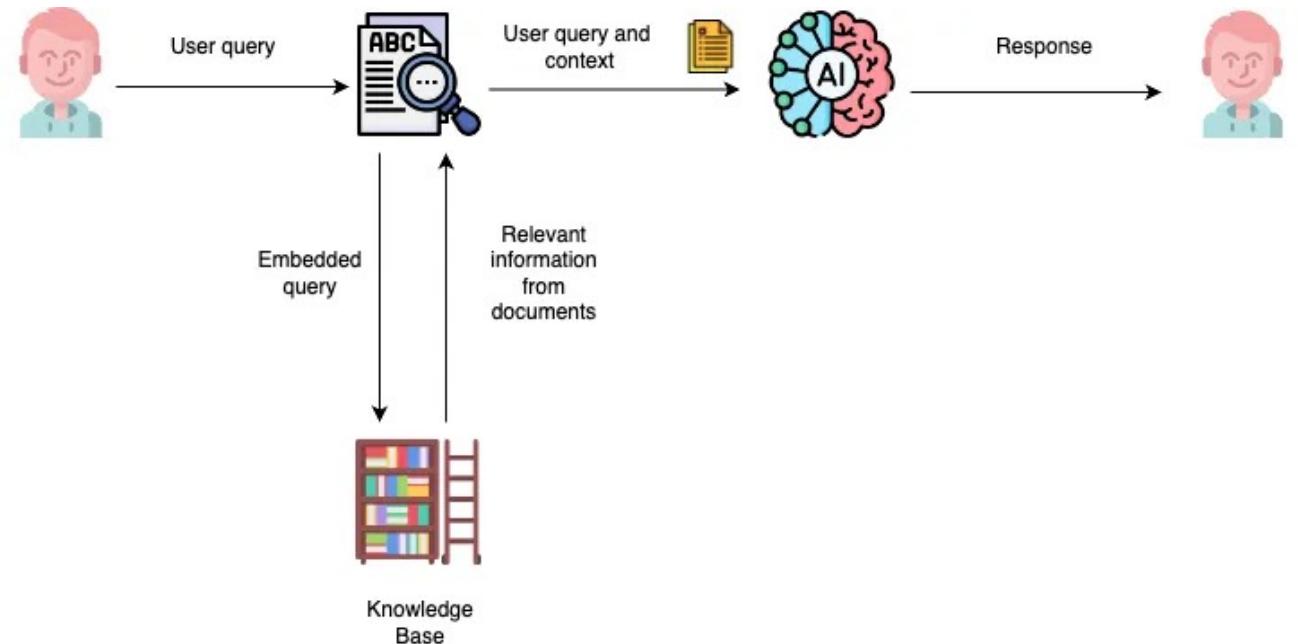




Retrieval Augmented Generation (RAG)

- Complement the base model with *external knowledge base*

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



Source: <https://blog.gopenai.com>

Why RAG?

Fine-tuning a foundation model

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

```
### INSTRUCTION:  
Please convert the following context into an SPARQL query.  
  
### CONTEXT:  
Retrieve all species from Bgee.  
  
### STEPS:  
1. Recognize that all species are defined as a `up:Taxon`.  
2. Use the prefix `up:` which stands for `http://purl.uniprot.org/core/`  
3. Query for all entities that are of type `up:Taxon`.  
  
### SPARQL:  
PREFIX up: <http://purl.uniprot.org/core/>  
SELECT ?species {  
    ?species a up:Taxon .  
}
```

- 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

what are the GO annotations of P12345 in Uniprot?

The UniProt entry **P12345** corresponds to the mitochondrial aspartate aminotransferase (GOT2) from *Oryctolagus cuniculus* (rabbit). This protein has several Gene Ontology (GO) annotations that describe its molecular functions, biological processes, and cellular components. UniProt +2

Molecular Function (MF)

- GO:0004069 – L-aspartate:2-oxoglutarate aminotransferase activity
- GO:0016212 – Kynurenine-oxoglutarate transaminase activity
- GO:0030170 – Pyridoxal phosphate binding EMBL-EBI +1 EMBL-EBI

Biological Process (BP)

- GO:0006531 – Aspartate metabolic process
- GO:0006533 – Aspartate catabolic process
- GO:0006536 – Glutamate metabolic process
- GO:0006103 – 2-oxoglutarate metabolic process 



“Cheating!”

what are the GO annotations of P12345 in Uniprot?

UP BioMart Searching the web

Customize ChatGPT

Introduce yourself to get better, more personalized responses ⓘ

+ Skeptical

+ Traditional

+ Forward thinking

+ Poetic



Anything else ChatGPT should know about you? ⓘ

Interests, values, or preferences to keep in mind

Advanced ^

ChatGPT capabilities ⓘ

Web Search

DALL·E

Code

Canvas

Advanced Voice

Disable and
try again!





Asking chatGPT question directly ("zero-shot")

what are the GO annotations of P12345 in Uniprot?

To retrieve the Gene Ontology (GO) annotations for the UniProt entry **P12345**, you can either:

1. Check Directly on UniProt Website:

Visit the [UniProt entry for P12345](#). Look for the "Function" section, where GO annotations are categorized under:

- GO - Biological process (BP)
- GO - Molecular function (MF)
- GO - Cellular component (CC)

2. Example of GO Annotations (If Available):

Let me fetch a simplified example (if the entry exists and is publicly known):

Example (this depends on what P12345 maps to — for instance, if it corresponds to the **PPA1_HUMAN** protein, Inorganic pyrophosphatase):

- GO:0000166 – nucleotide binding (MF)





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



Recommended: Ask Model To Ask an API

can you write an API call to retrieve the GO annotations of P12345 from Uniprot?

Sure! You can retrieve Gene Ontology (GO) annotations for a UniProt entry (like `P12345`) using the **UniProt REST API**. Here's a simple example using `curl` (command-line tool) and a Python snippet using `requests`.

◆ Using `curl` (command line):

bash

Copy Edit

```
curl -X GET "https://rest.uniprot.org/uniprotkb/P12345.json" -H "Accept: application/json"
```

This returns the full UniProtKB entry for P12345 in JSON format. You can then parse the `"uniProtKBCrossReferences"` or `"comments"` and `"keywords"` fields to extract GO annotations.



Recommended: Ask Model To Ask an API

```
import requests

uniprot_id = "P12345"
url = f"https://rest.uniprot.org/uniprotkb/{uniprot_id}.json"

response = requests.get(url)
data = response.json()

# Extract GO annotations
go_annotations = [
    ref for ref in data.get("uniProtKBCrossReferences", [])
    if ref.get("database") == "GO"
]

for go in go_annotations:
    go_id = go["id"]
    go_terms = go.get("properties", [])
    for term in go_terms:
        if term["key"] == "GoTerm":
            print(f"{go_id}: {term['value']}")
```

Copy

Edit



Analysing local files?
(LLMs with context)



Example: the SIB resources CSV

Download file from [here](#)

title	slug	url	description	short_description	create
UniProtKB	uniprotkb	https://www.uniprot.org/	The UniProt Knowledgebase is a central hub for the collection, analysis, distribution and annotation of protein information.	Protein sequence database	2020-01-01
Selectome	selectome	https://selectome.org/	Selectome is a database of positive selection based on evolutionary conservation.	Database of positive selection	2020-01-01
PROSITE	prosite	https://prosite.expasy.org/	PROSITE consists of documentation entries describing protein families and domains.	Protein family and domain database	2020-01-01
SWISS-MODEL Repository	swiss-model-repository	https://swissmodel.expasy.org/repository/	SWISS-MODEL Repository is a database of protein structures and homology models.	Protein structure homology models	2020-01-01



Use ChatGPT for QA over CSV

What can I help with?

Connect to Google Drive
Connect to Microsoft OneDrive >

Add photos and files

+ Search Reason Deep research Create image ...

Microphone icon (white), Mic icon (black)

expasy_resources_metadata.csv
Spreadsheet

X

can you show me the first 5 lines of the table?

+ Search Reason Deep research Create image ...

Microphone icon (white), Up arrow icon (black)



Use ChatGPT for QA over CSV



expasy_resources_metadata.csv
Spreadsheet

can you show me the first 5 lines of the table?

Here are the first 5 rows of the table from the CSV file:

title	slug	url	short_description	license	⋮
UniProtKB	uniprotkb	https://www.uniprot.org/	Protein sequence database	CC BY 4.0	
Selectome	selectome	https://selectome.org/	Database of positive selection	CC0	
PROSITE	prosite	https://prosite.expasy.org/	Protein family and domain database	PROSITE license	



Use ChatGPT for QA over CSV

are these all the columns of the spreadsheet?

No, those are not all the columns. The spreadsheet contains the following columns:

1. title
2. slug
3. url
4. metasearch_url
5. description
6. short_description
7. media_url
8. group_info
9. created
10. updated
11. ontology_terms

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*](#)



can you explain the decision tree in Figure 5?

Figure 5 in the *ESMO Clinical Practice Guidelines for localized non-small-cell lung cancer (NSCLC)* outlines the **decision tree for preoperative cardiac evaluation**. This decision tree helps clinicians assess and optimize the cardiac risk of patients prior to lung cancer surgery. Here's a breakdown of the figure and its clinical logic:

◆ Step-by-step Explanation of the Decision Tree in Figure 5

1. Initial Assessment:

- Begin with a **history, physical examination, and baseline ECG**.
- **Calculate the RCRI** (Revised Cardiac Risk Index), which includes:
 - High-risk surgery (e.g., lobectomy or pneumonectomy)
 - Ischemic heart disease
 - Heart failure
 - Insulin-dependent diabetes
 - Prior stroke or TIA
 - Elevated creatinine ($\geq 2 \text{ mg/dL}$)

2. Red Flags or Symptoms?

- If the patient has **at least three RCRI weighted factors** or:
 - Any known cardiac condition requiring medications
 - A newly suspected cardiac condition
 - Inability to climb two flights of stairs
- Then proceed with a **cardiology consultation**, including **non-invasive testing** as per AHA/ACC guidelines.

Clinical Practice Guidelines

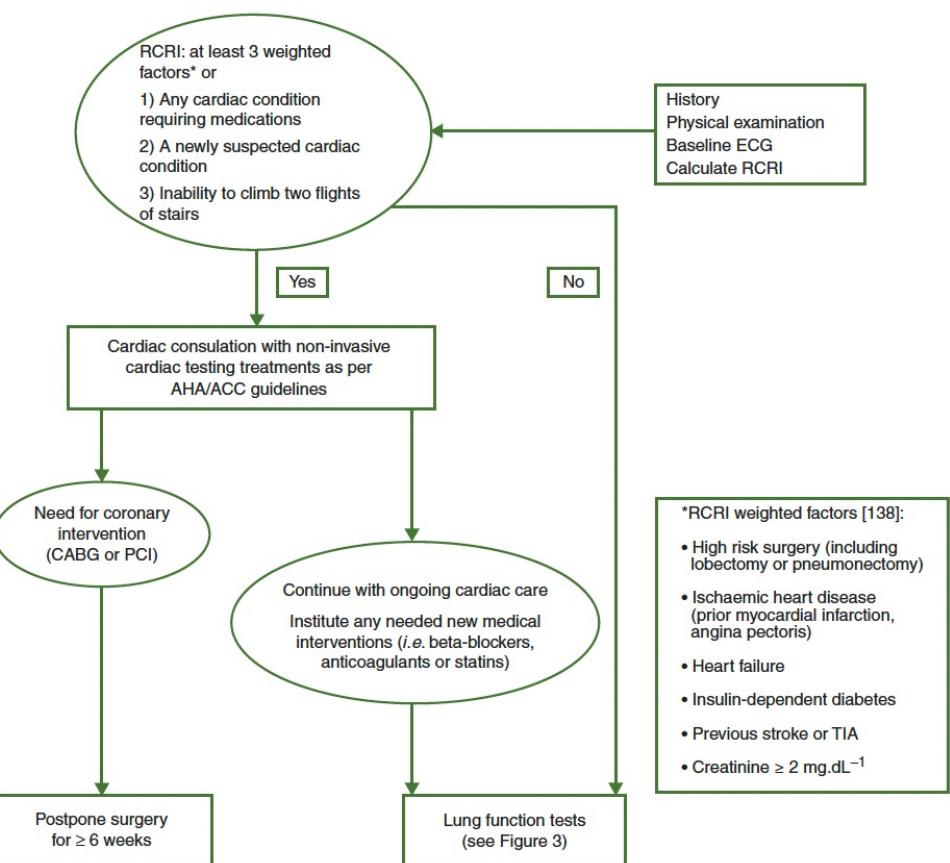


Figure 5. Preoperative cardiac evaluation.

AHA/ACC, American Heart Association/American College of Cardiology; CABG, coronary artery bypass grafting; ECG, electrocardiogram; PCI, percutaneous coronary intervention; RCRI, revised cardiac risk index; TIA, transient ischaemic attack.
Reprinted from [50], with permission from the European Respiratory Society.



Question Answering over Structured Data



Swiss Institute of
Bioinformatics

[Home](#)[About](#)[Help](#)[ExpasyGPT](#)[SIB News](#)

Posted 16 May 2025 - [ExpasyGPT](#) has launched! This powerful AI-driven tool makes it faster and easier than ever to explore, retrieve and compile data and information to answer complex biological questions (more [here](#)). [Try the beta version](#) and [contact Expasy](#) with your feedback!

Ex^{po}sy

Swiss Bioinformatics Resource Portal



e.g. [BLAST](#), [UniProt](#), [MSH6](#), [Albumin](#)...

Genes & Genomes

Genomics

Metagenomics

Transcriptomics

Proteins & Proteomes

Evolution & Phylogeny

Evolution biology

Population genetics

SIB Resources



Bgee

Gene expression expertise



STRING

Protein-protein interaction networks and enrichment analysis



SwissLipids

Knowledge resource for lipids



V-pipe

Viral genomics pipeline



Nextstrain

Impact of pathogen genome data



mOTUs

Microbial taxonomic and population



Cellosaurus

Knowledge resource on c



UniProtKB/Swiss-Prot

3

About Expasy

Last updated: April 16, 2025

Scope

Expasy is the [SIB Swiss Institute of Bioinformatics](#) registry of bioinformatics software tools, database/knowledgebase resources, and data sets developed by SIB members. It provides detailed information about software tools, databases and data sets that support a range of life science and clinical research domains, from genomics, proteomics and structural biology, to evolution and phylogeny, systems biology and medical chemistry.

Brief history

Expasy was created in August 1993, at the dawn of the internet era, by two SIB-cofounders [Ron Appel](#) and [Amos Bairoch](#). At the time, it was referred to as 'ExPASy, the Expert Protein Analysis System' as proteins were its primary focus. It was the first life science website - and among the 150 very first websites in the world!

In June 2011, it became the SIB Expasy Bioinformatics Resources Portal, a catalogue of bioinformatics resources developed by SIB Groups.

The current version, Expasy 3.0, was released in October 2020, following a massive user study and taking into account design, user experience and architecture aspects.

License

The registry content is freely available to all under the Creative Commons Attribution licence ([CC BY 4.0](#)).





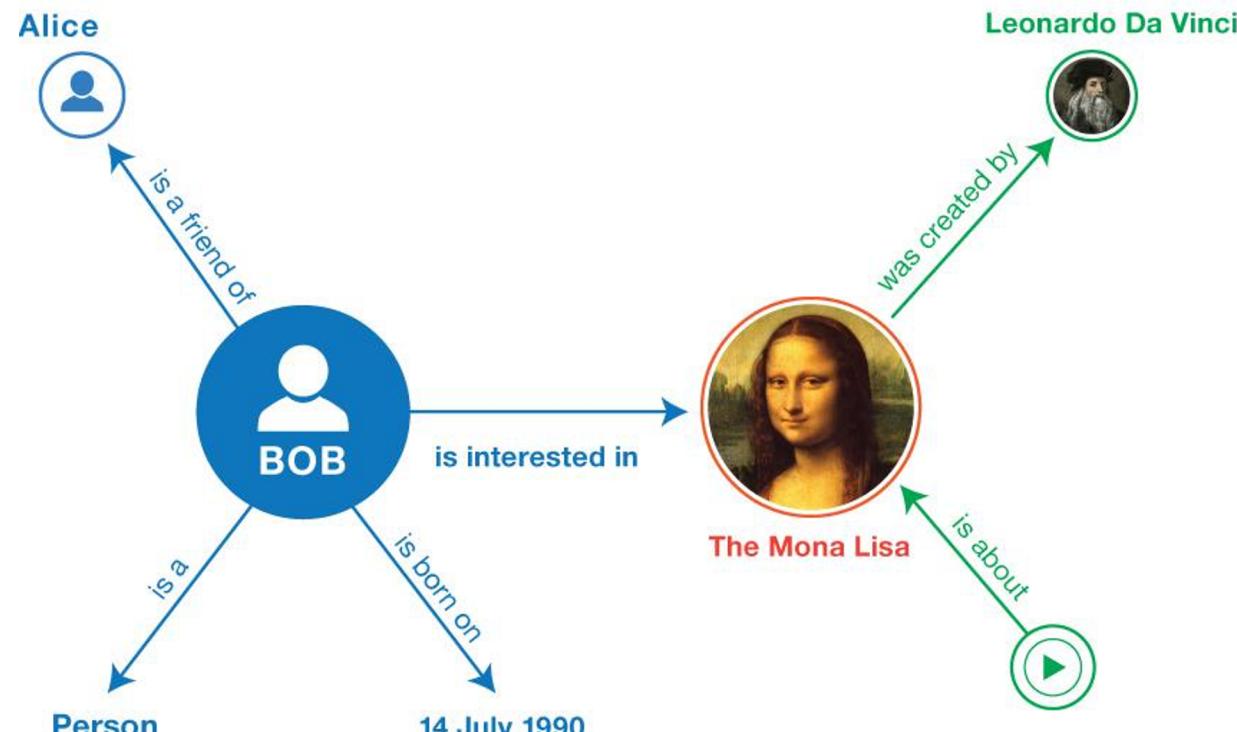
Knowledge Graphs curated by the SIB

12 resources found for sparql

  OMA SPARQL endpoint SPARQL access to the OMA Orthology database	  Bgee SPARQL endpoint SPARQL access to the Bgee gene expression database	  OrthoDB SPARQL endpoint SPARQL access to the OrthoDB catalog of orthologs	  HAMAP SPARQL endpoint SPARQL access to the HAMAP annotation rules
  UniProt SPARQL endpoint SPARQL access to UniProt and related data	  Rhea SPARQL endpoint SPARQL access to the Rhea knowledgebase	  SwissLipids SPARQL endpoint SPARQL access to the SwissLipids knowledgebase	  MetaNetX SPARQL endpoint SPARQL access to MetaNetX
  GlyConnect SPARQL endpoint SPARQL access to GlyConnect	  Cellosaurus SPARQL endpoint SPARQL access to Cellosaurus	  STRING SPARQL endpoint SPARQL access to the STRING knowledgebase	  SIBiLS SPARQL endpoint SPARQL access to SIBiLS, the SIB Literature Services

What is a Knowledge Graph?

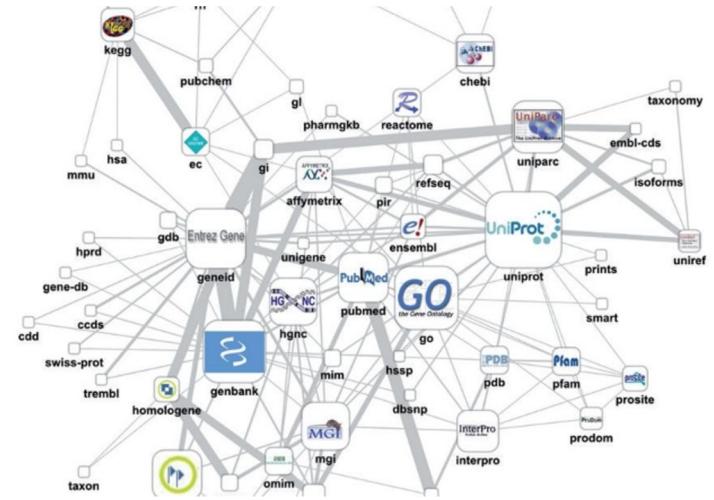
- Graph data model
 - Easy to **interlink** and extend
 - Data usually stored as triples
 - Triple = <subject, predicate, object>
 - E.g. "<gene X> isExpressedIn <anatomic entity Y>"
- Can be queried using SPARQL:
 - Select * where {
?gene isExpressedIn ?anatEntity.
}
- Play an important role in **Semantic Interoperability**



Why are these useful?

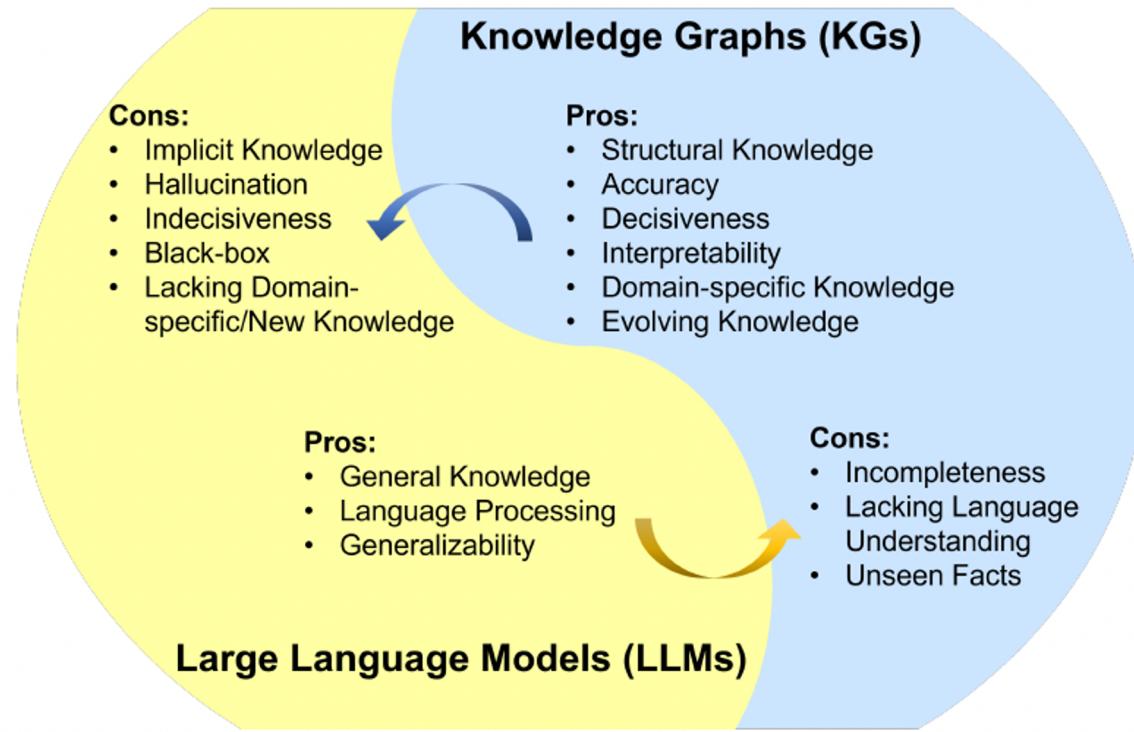
- Explicit semantics through:
 - Standard identifiers
 - Shared vocabularies
 - Ontologies
- Ask complex questions that span across disciplines and across datasets

Data Integration
and FAIRification



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

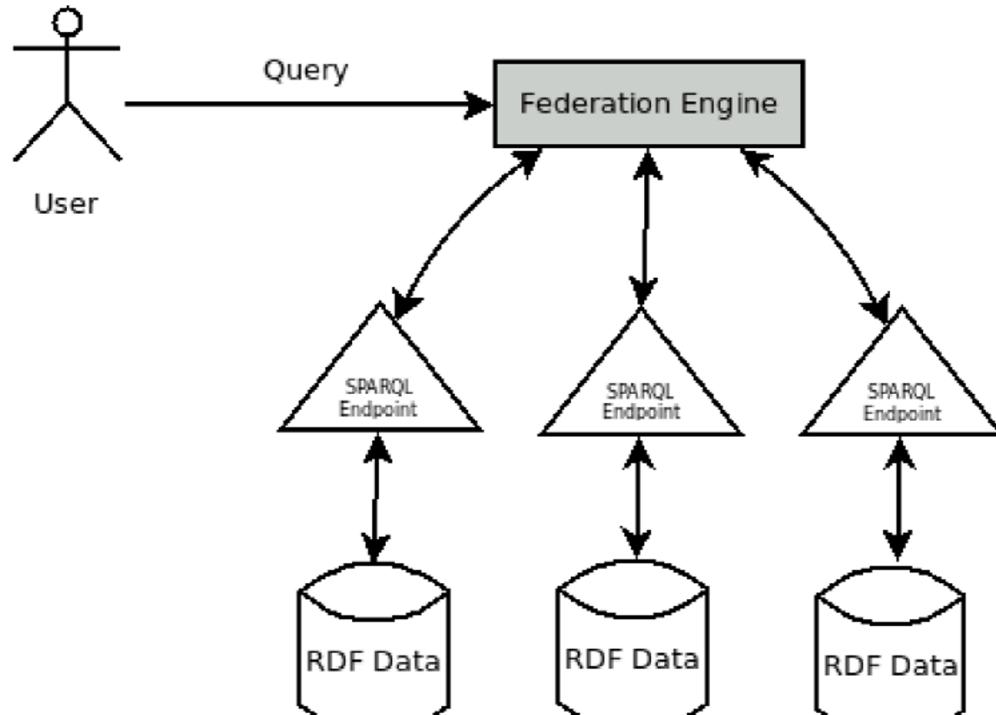
Research question: LLMs + KGs = ?



Retrieval Augmented Generation (RAG)

Fig. 1. Summarization of the pros and cons for LLMs and KGs. LLM pros: *General Knowledge* [11], *Language Processing* [12], *Generalizability* [13]; LLM cons: *Implicit Knowledge* [14], *Hallucination* [15], *Indecisiveness* [16], *Black-box* [17], *Lacking Domain-specific/New Knowledge* [18]. KG pros: *Structural Knowledge* [19], *Accuracy* [20], *Decisiveness* [21], *Interpretability* [22], *Domain-specific Knowledge* [23], *Evolving Knowledge* [24]; KG cons: *Incompleteness* [25], *Lacking Language Understanding* [26], *Unseen Facts* [27].

How do we extract information from RDF data?



```
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

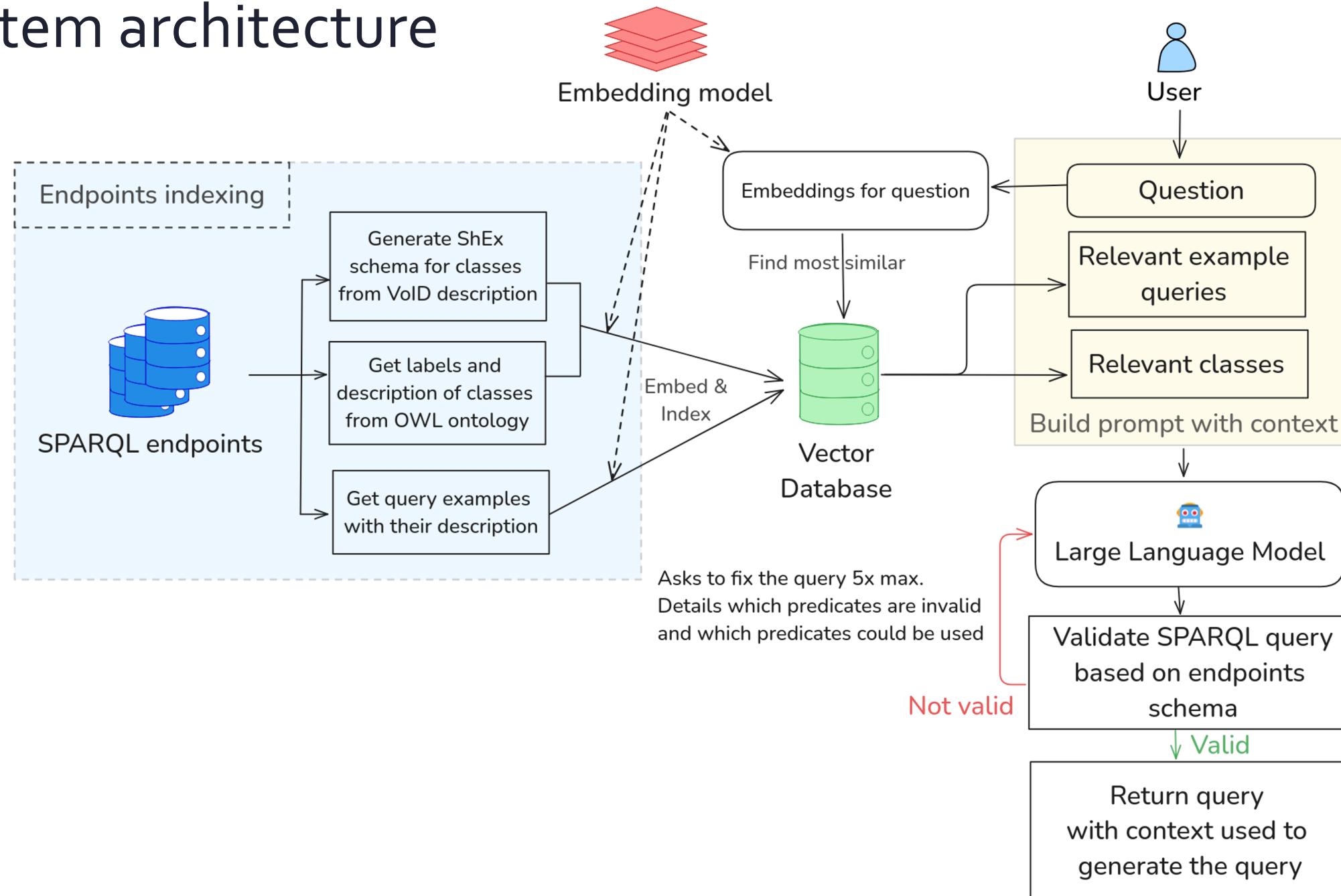
Writing SPARQL queries is hard and time-consuming.
LLMs are great at it, but they need context

For complex questions finding the right context is not trivial.

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



System architecture





ExpasyGPT Beta

ExpasyGPT is an AI-driven tool that facilitates queries across life science databases in Expasy. It also provides accurate information on bioinformatics resources developed by SIB. This allows you to easily:

- identify resources for your use case
- compile biological information from different databases*
- ask complex questions of one database*

The tool either answers your question directly, or converts it into a SPARQL query that you can then automatically run across relevant databases on Expasy and/or further edit.

* See [how ExpasyGPT works](#), including which databases it can query.

ExpasyGPT is in Beta version and may make mistakes or experience issues. We are continually improving its performance, and **highly value your input!** Please send your suggestions or questions to the [Expasy helpdesk](#).

Ask your question



Which SIB resources are supported by ExpasyGPT?

Where is the ACE2 gene expressed in humans?



Swiss Institute of
Bioinformatics

[Services](#) [Training](#) [Community](#) [What's on](#) [About](#) [Intranet](#) [Careers](#) [Contact](#)

[Home](#) > [What's on](#) > [News](#)

Enabling faster answers to complex biological questions

ExpasyGPT: enhanced biological discovery with AI and knowledge representation

08 May 2025

Expasy
Swiss Bioinformatics Resource Portal

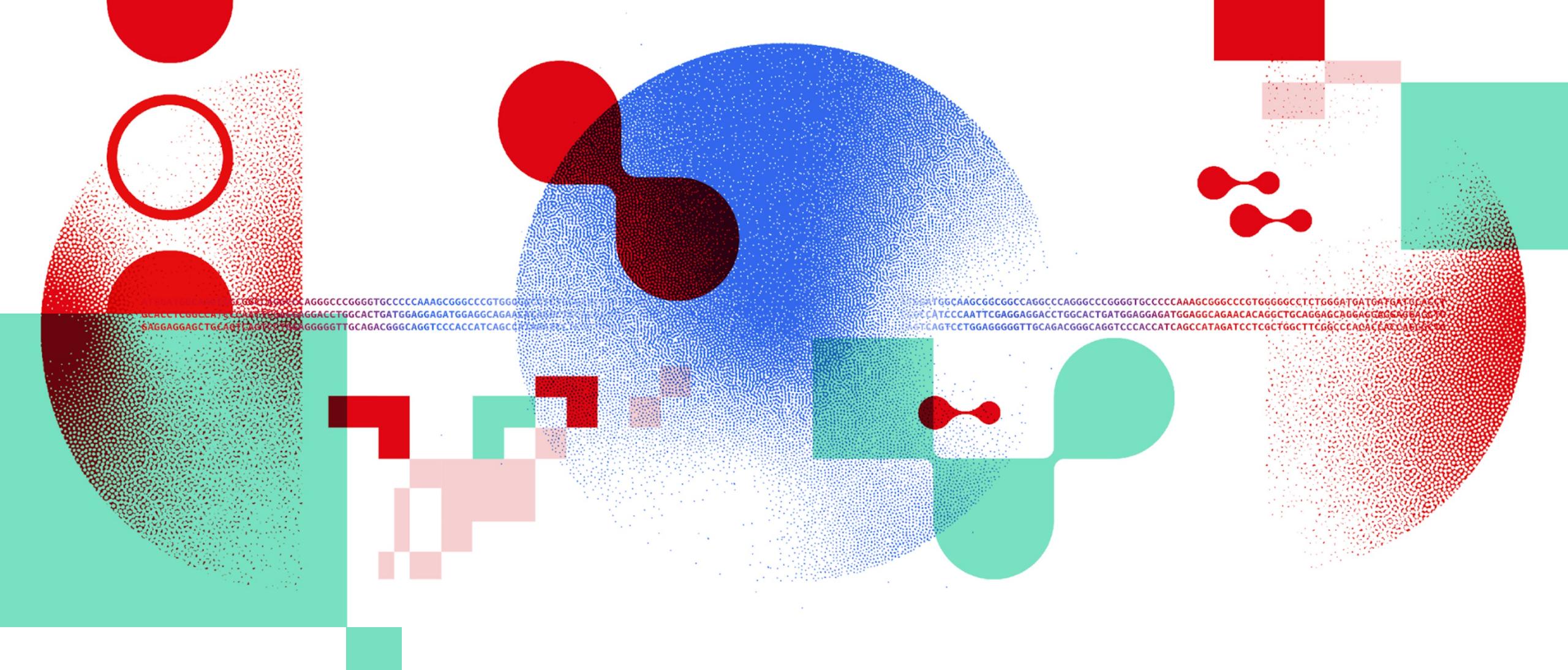


Some Intermediate Conclusions...

- 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-langs-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.pdf>, 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., Schaeckermann, 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. "Emotional Prompting Amplifies Disinformation Generation 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