
A:C:G:T
A.C.G.T
TALK DIRTY GENOMIC TO ME
A Journey From Sequences to Sentences



Martin Danner

Data Scientist & ML Engineer / scieneers GmbH
PhD Stud. / University Hospital RWTH Aachen
martin.danner@scieneers.de

Short Introduction

Looking forward to the exchange!

CV

- M.Sc. Neural Engineering
  2018-2019
- August-Wilhelm Scheer Institute
  2019-2021
- scieneers GmbH
  since 2021
- Center for Human Genetics and Genomic Medicine
  since 2024

Technological Focus

- Machine Learning / Deep Learning / Generative AI
- Machine Learning Operations
- Development Operations
- Cloud based Data Platforms
- Data Driven Products
- MS Azure & Databricks


scieneers
DRIVEN BY DATA
More about us on www.scieneers.de

Center for Human Genetics and Genomic Medicine

Uniklinik RWTH Aachen



Genome sequencing in rare diseases

Genome sequencing in rare diseases

Rare diseases

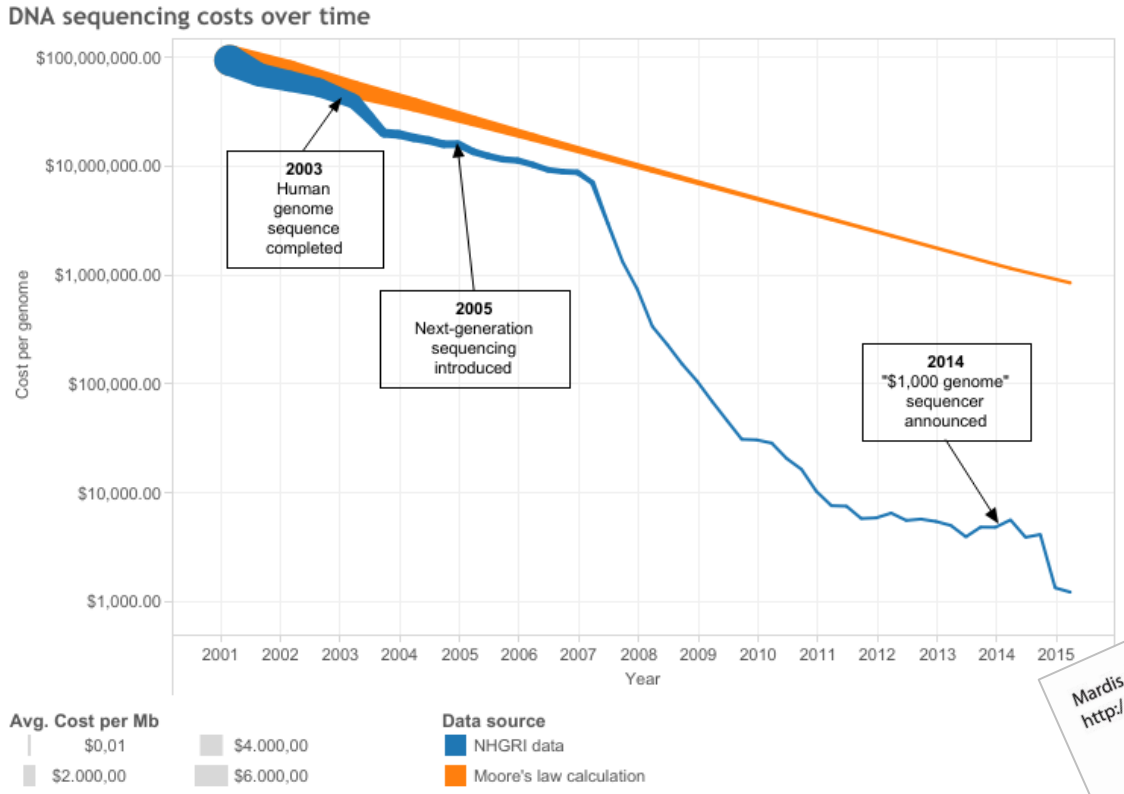
- 350 million, almost 5% of the world's population, live with a rare disease
- More than 7,000 rare diseases are known to date
- ~4 million affected individuals in Germany
- About 75% of rare diseases affect children
- On average, nearly 5 years pass before a correct diagnosis

80% of these diseases are caused by a single genetic change and can be diagnosed through genome analysis.

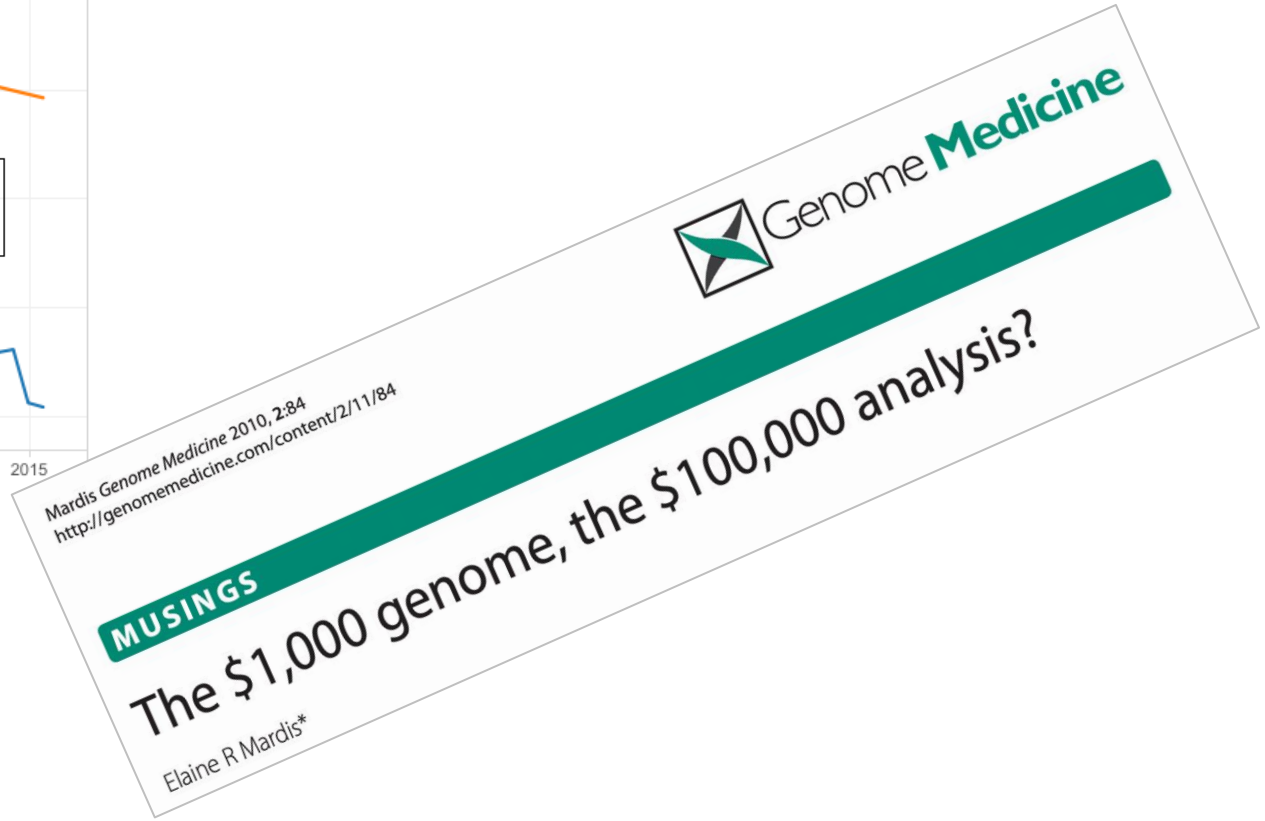


The cost of sequencing genomes is dropping

Analysis is the bottleneck

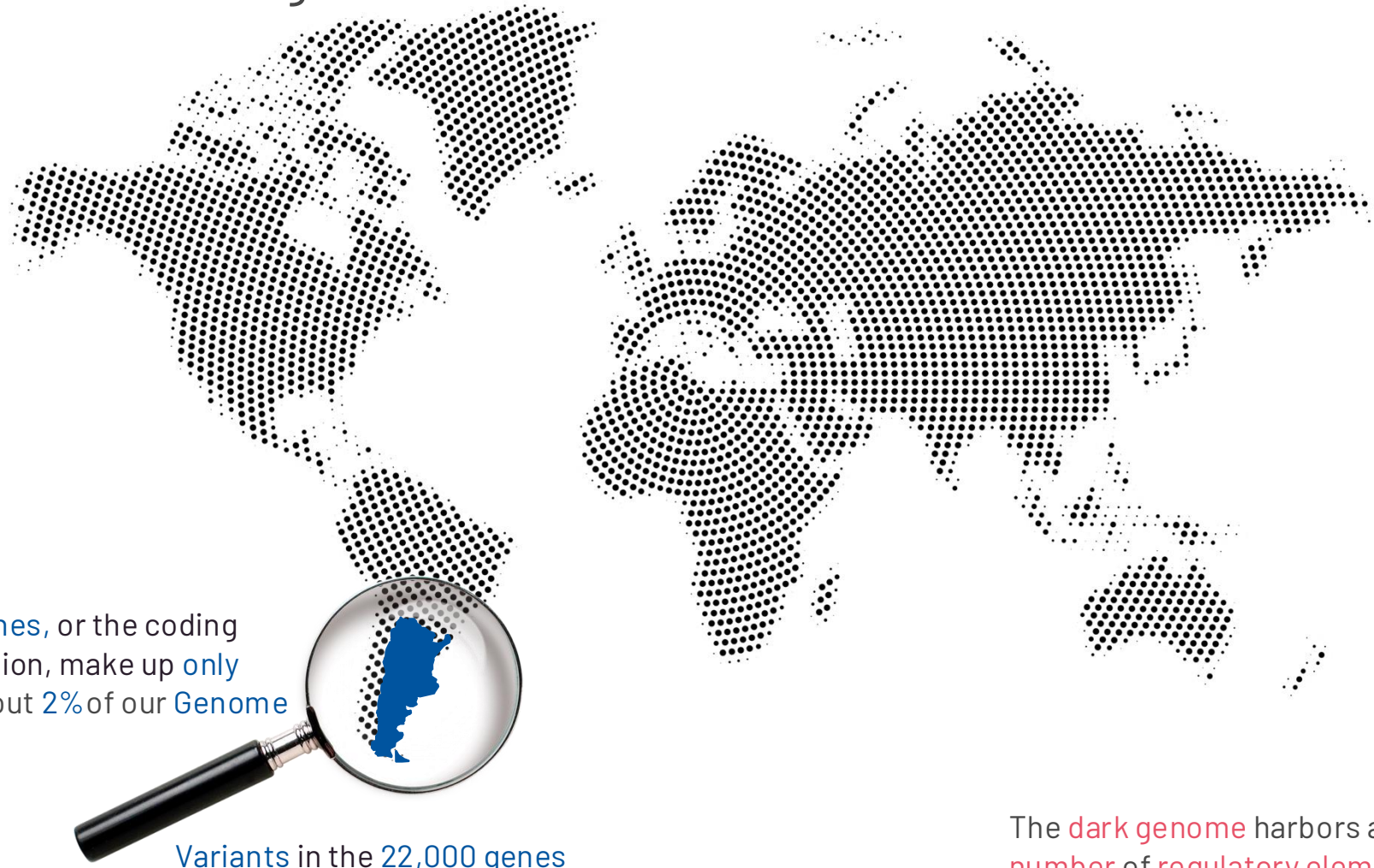


Decline in real costs compared to expected declines based on Moore's Law.
Trend line: Cost per human genome. Line width: Cost per megabase (Mb)
(*Data: NHGRI* https://www.genome.gov/27541954/dna-sequencing-costs-data/)



Genome research: The tip of the iceberg

Only a fraction of our genome is understood so far



Genes, or the coding region, make up only about 2% of our Genome

Variants in the 22,000 genes are comparatively well understood

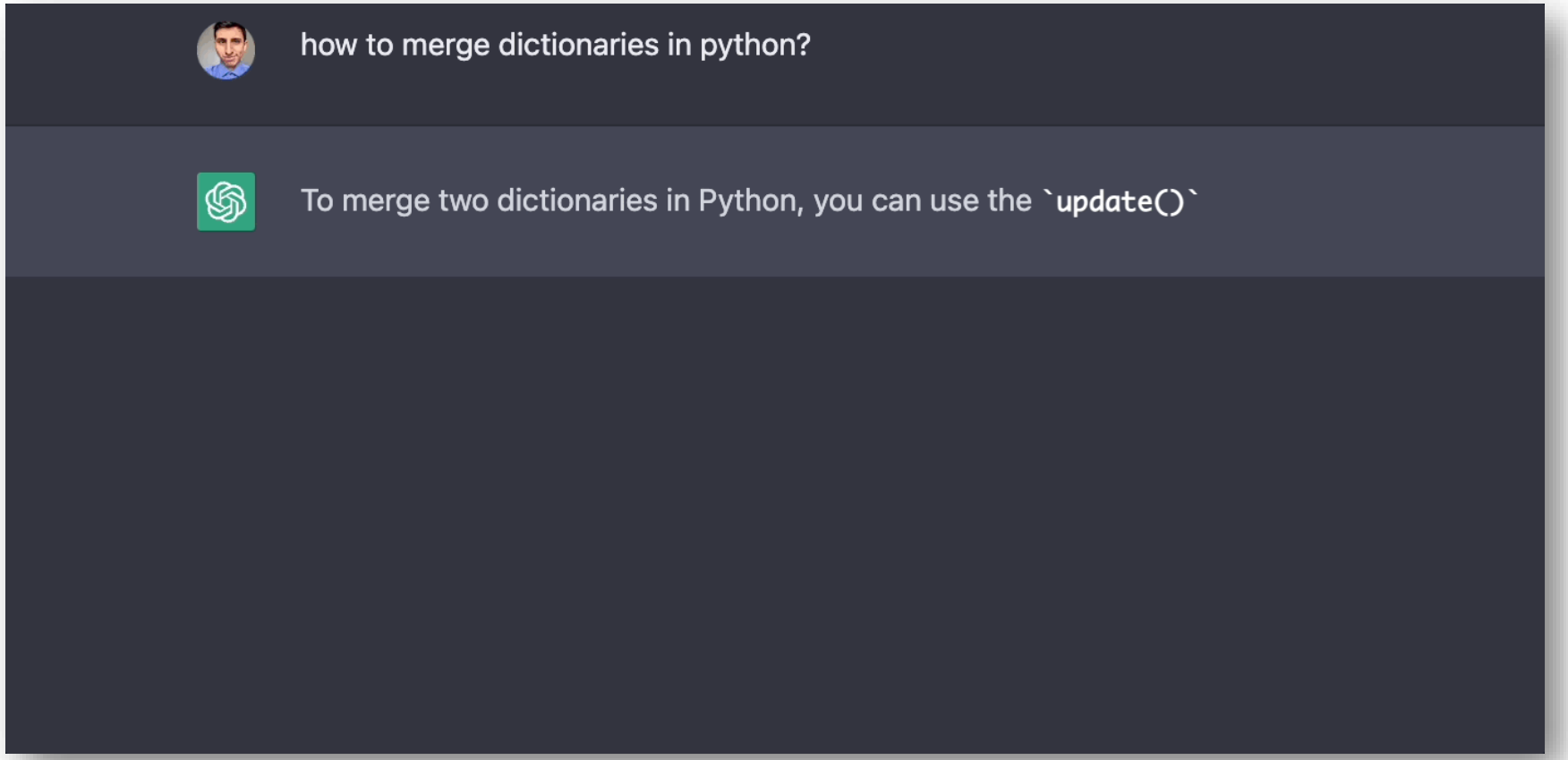
Variants in the dark genome (non-coding region) are so far not, or barely, interpretable

These many segments – about 98% of the genome – currently elude assessment!

The dark genome harbors a large number of regulatory elements, such as short open reading frames

We are all familiar with LLMs

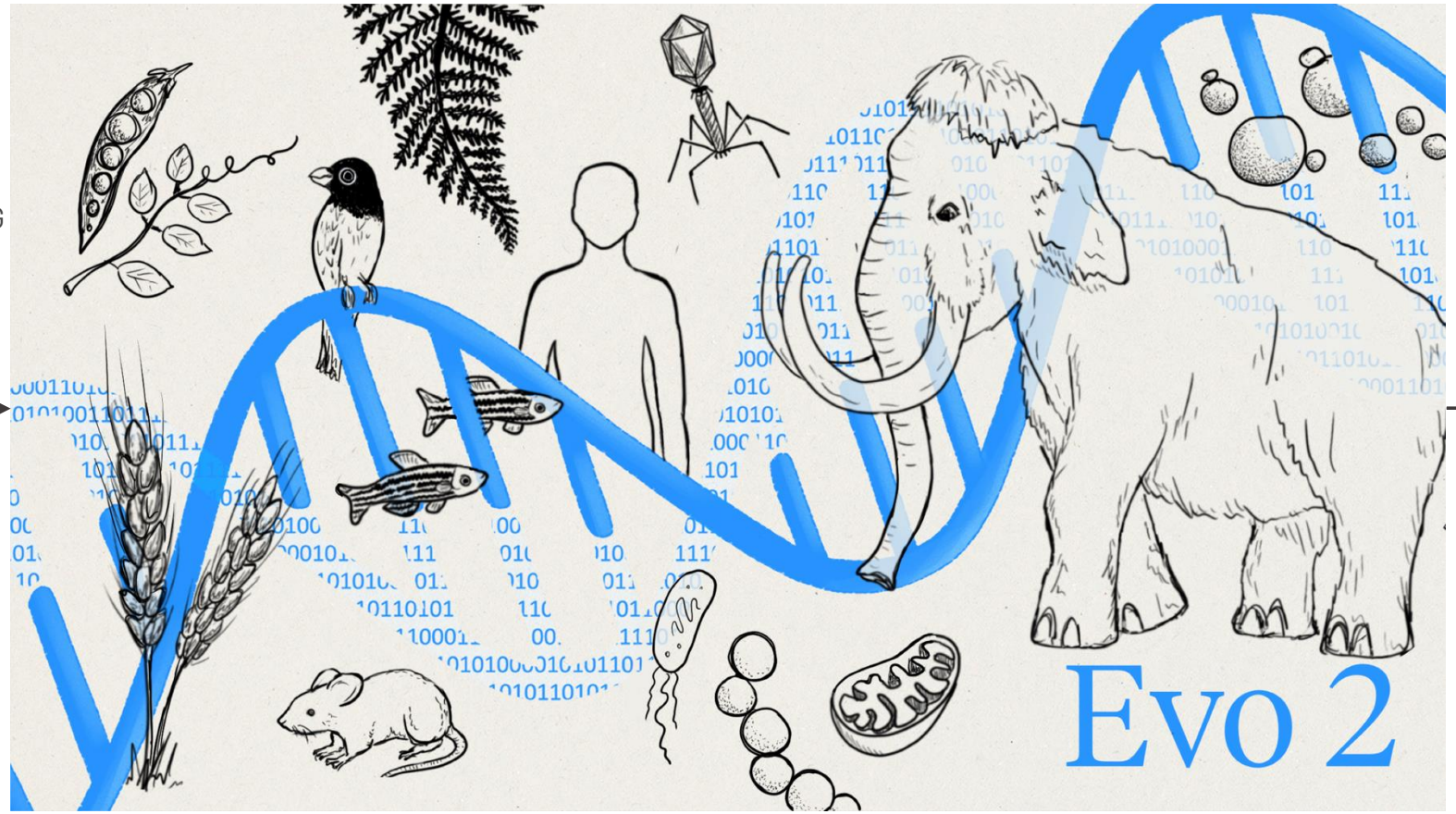
LLMs are already well established in our daily life 🤭



But have you come across GLMs yet?

Let´s talk about Genomic Language Models 🤯

ATGCGTACGTTAGGCTAACCTGGATTG



...TAACCTGGA
TT

Identified Gene

Pathogenicity Score
0.8

https://arcinstitute.org/news/blog/evo2

Evo 2 is the largest artificial intelligence model in biology to date, trained on over 9.3 trillion nucleotides and 100,000 species across the entire tree of life

Great! We identified a new Gene with GLMs!

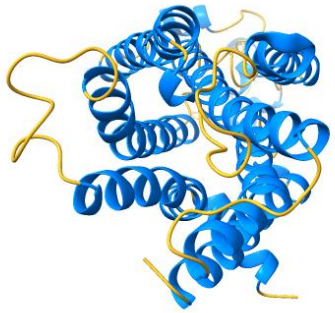
But how to determine it´s function?

Newly identified genomic element

GTTAGGCTAAATGCGTACGTTAGGCTAACCTGGATTGCGTACGTTAGGCTAAC



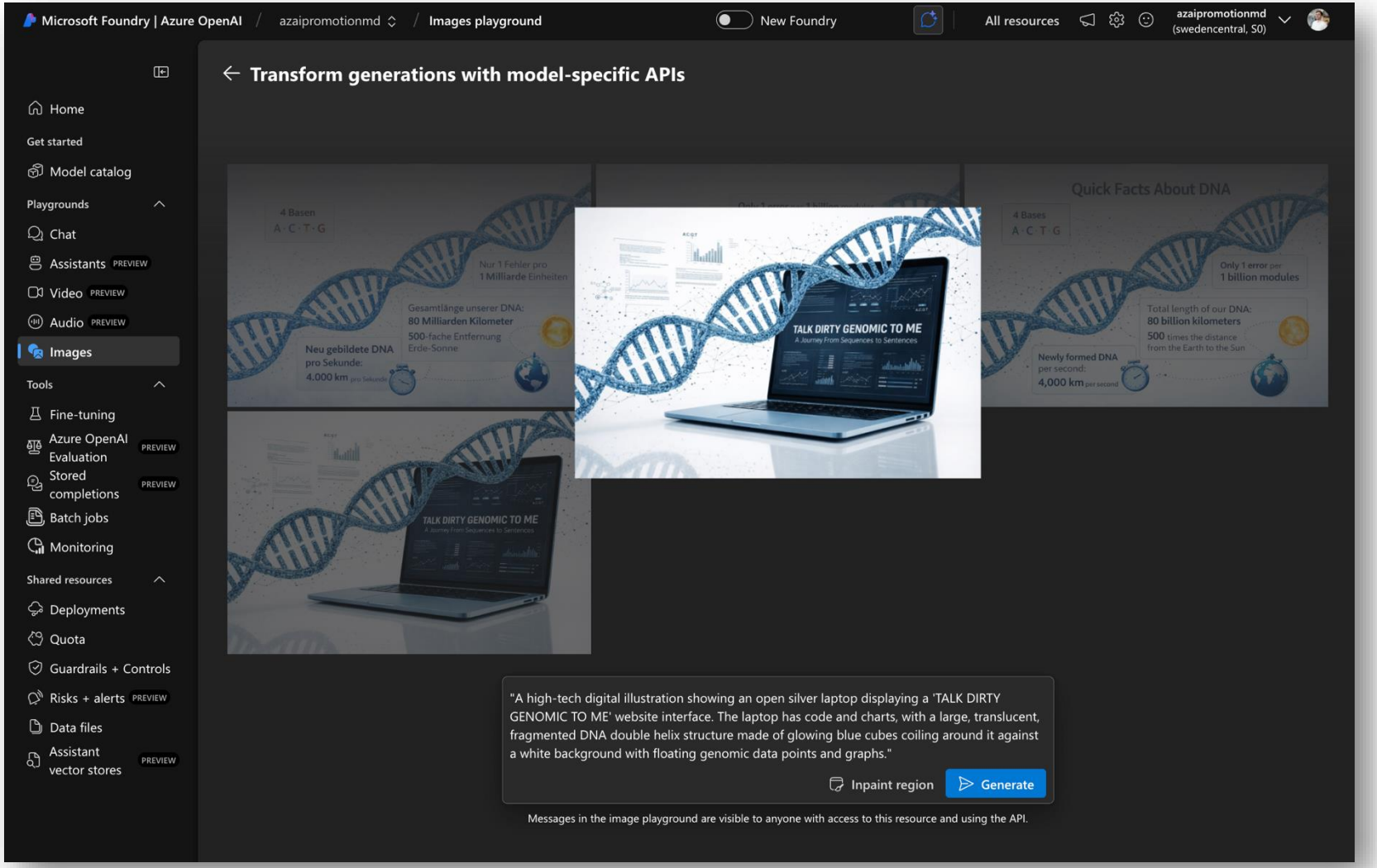
MEAQRRNQLPGD



Molecular Function? 🧐

Why not ask an LLM directly? 🤔

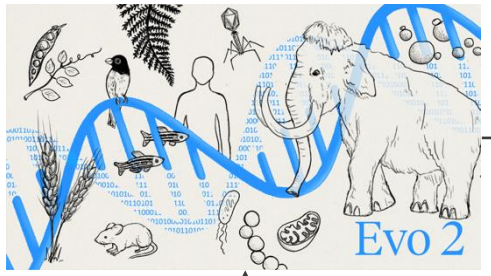
Inspired by other MLLMs – Translating between Text and Images



Why not ask an LLM directly? 🤔

Combining GLMs and LLMS - Translating Between Sequences and Sentences

Could the given protein structure contain sites compatible with direct interaction with RNA–protein complexes?





The protein structure reveals surface-exposed basic patches and conserved motifs that are frequently observed in proteins capable of binding RNA–protein complexes, supporting the possibility of such interactions.

ATGCGTACGTTAGGCTAACCTGGATTG

Genolator – From Sequences to Sentences

Translating between Life´s Code and Natural Language

ATGCGTACCTTAGCGAGTACCGTTAGCATGCCGATTAACG...

Could the given protein contain features that would allow it to associate with lipid-rich cellular boundaries?



User

The presence of hydrophobic stretches and predicted transmembrane domains in the given amino acid sequence suggests the protein may be capable of associating with lipid bilayers.



Genolator

ATGCGTACCTTAGCGAGTACCGTTAGCATGCCGATTAACG...

Could the features observed in the given protein suggest a role in metabolic enzymatic catalysis, such as carbohydrate breakdown?



User

No data indicate that the given protein contains features associated with metabolic enzymatic catalysis, such as carbohydrate breakdown.



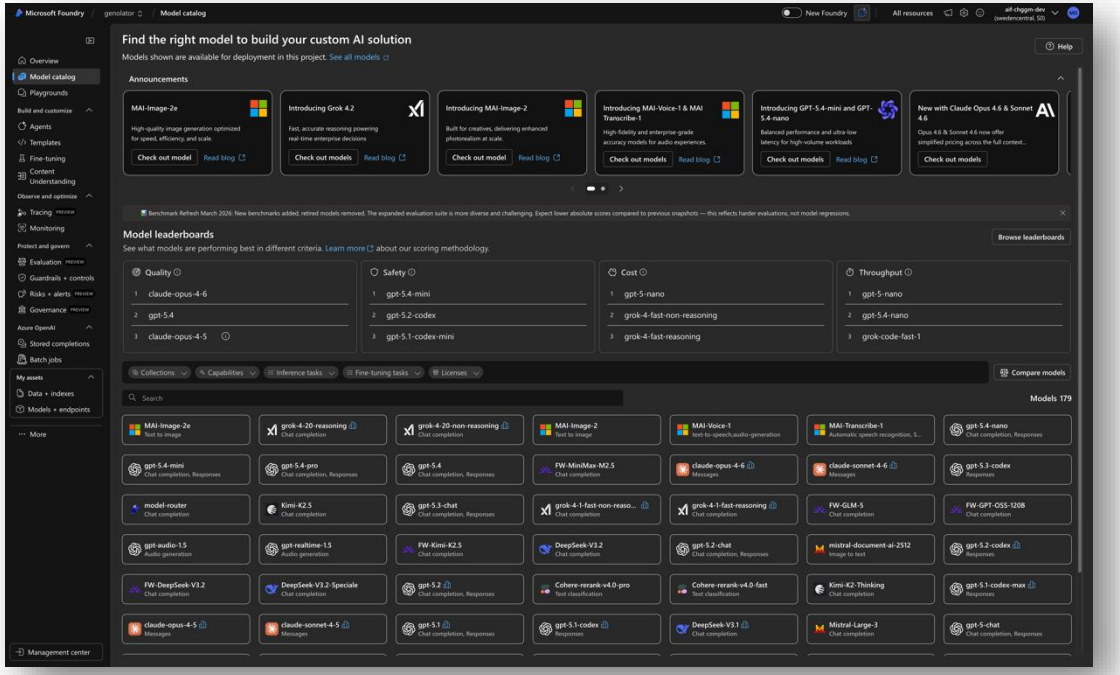
Genolator

Generation of a Gene Ontology QA-Dataset...

...with >360.000 QA pairs using Azure AI Foundry

This gene encodes the erythropoietin receptor which is a member of the cytokine receptor family. Upon erythropoietin binding, this receptor activates Jak2 tyrosine kinase which activates different intracellular pathways including...

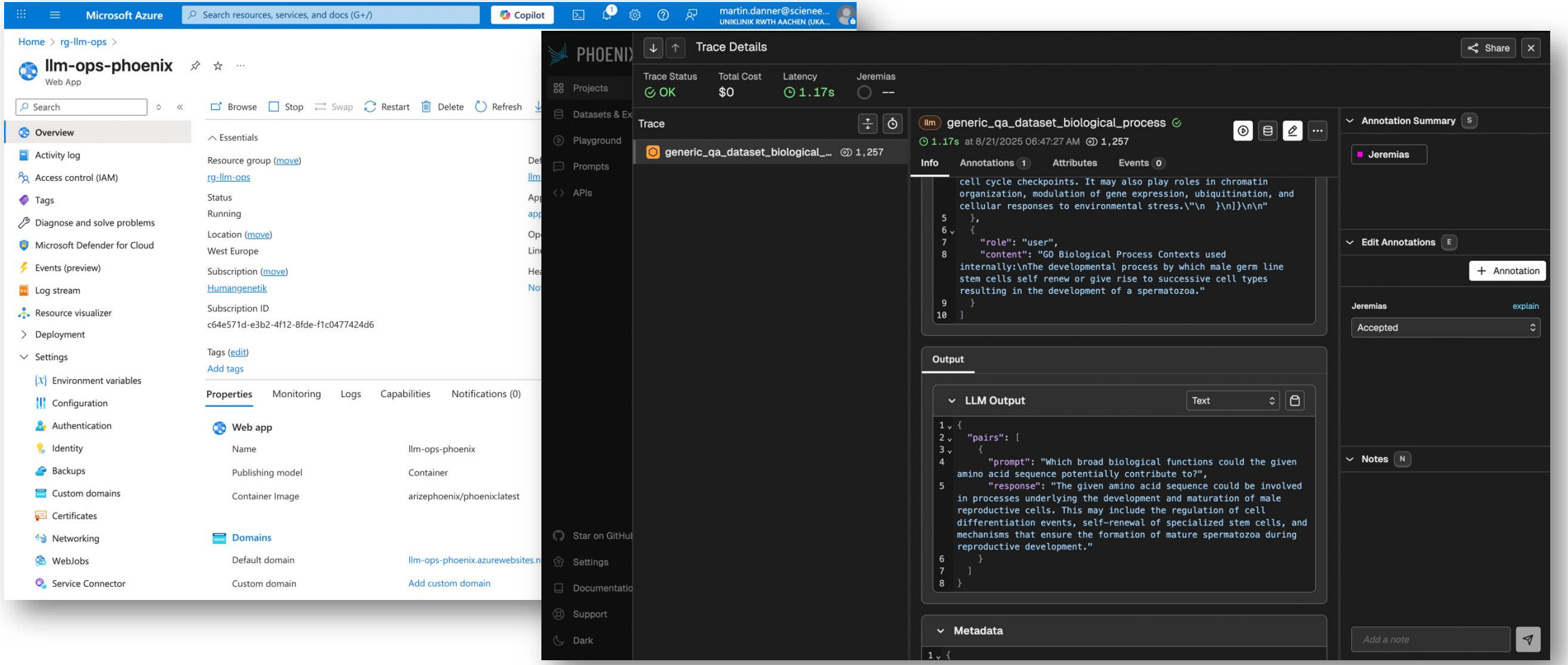
GPT-4.1




```
{"pairs": [
  {"prompt": "What cellular processes could the given protein
              be involved in?",
   "response": "The given protein likely participates in signaling
               cascades that regulate cell survival,
               proliferation, and differentiation, particularly
               within specific cell lineages such as those
               involved in blood formation."
  },
  {"prompt": "Could the given amino acid sequence indicate a
              role in receptor-mediated signal transduction?",
   "response": "Yes, features commonly found in the given
               amino acid sequence....
```

Large Language Model Operations

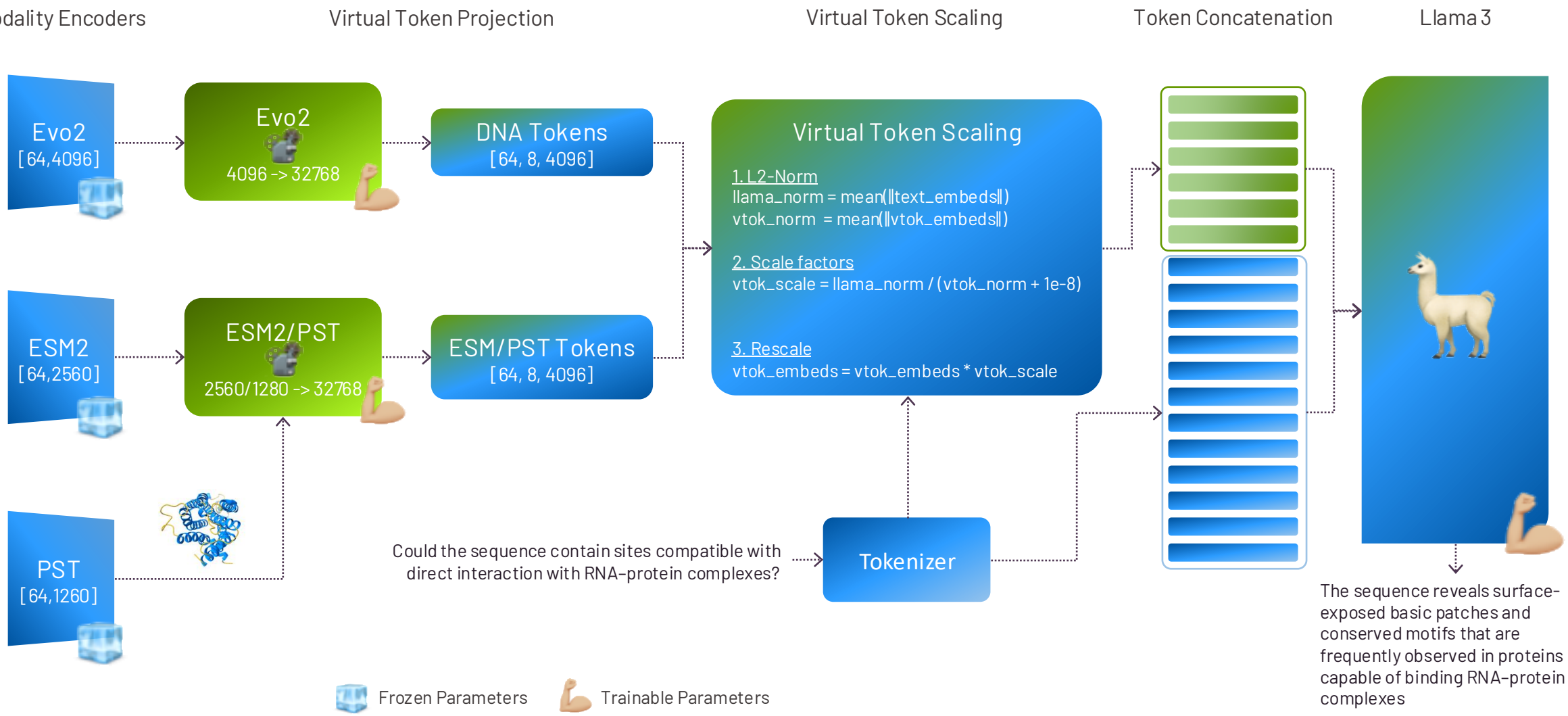
LLMOps using Azure Web Apps & Container Services



Tracing, Cost Tracking, Evaluation and Annotation of Millions of Generated QA Pairs

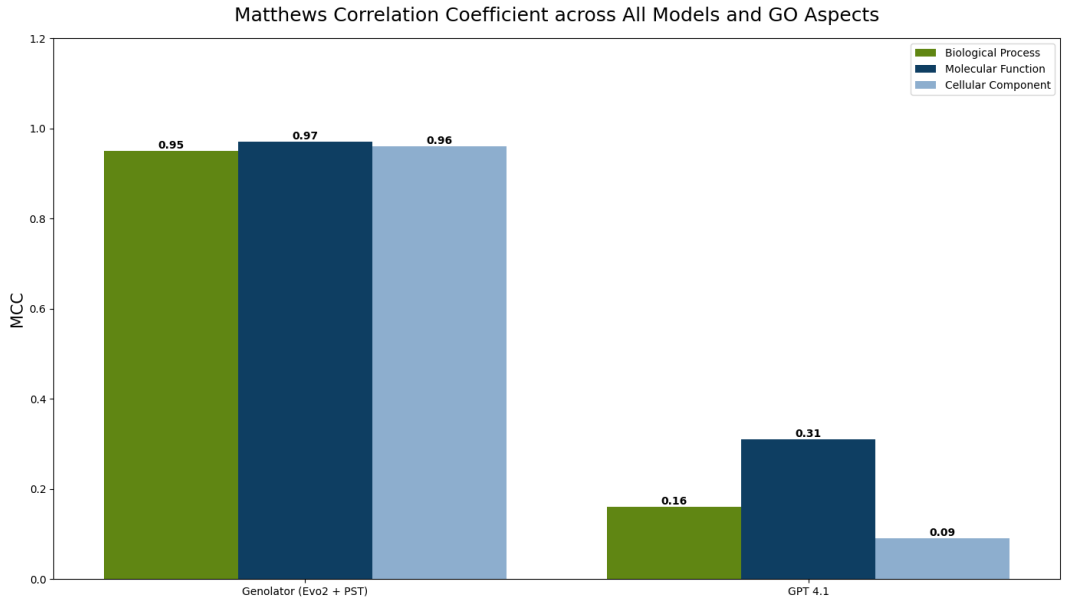
Genolator – From Sequences to Sentences

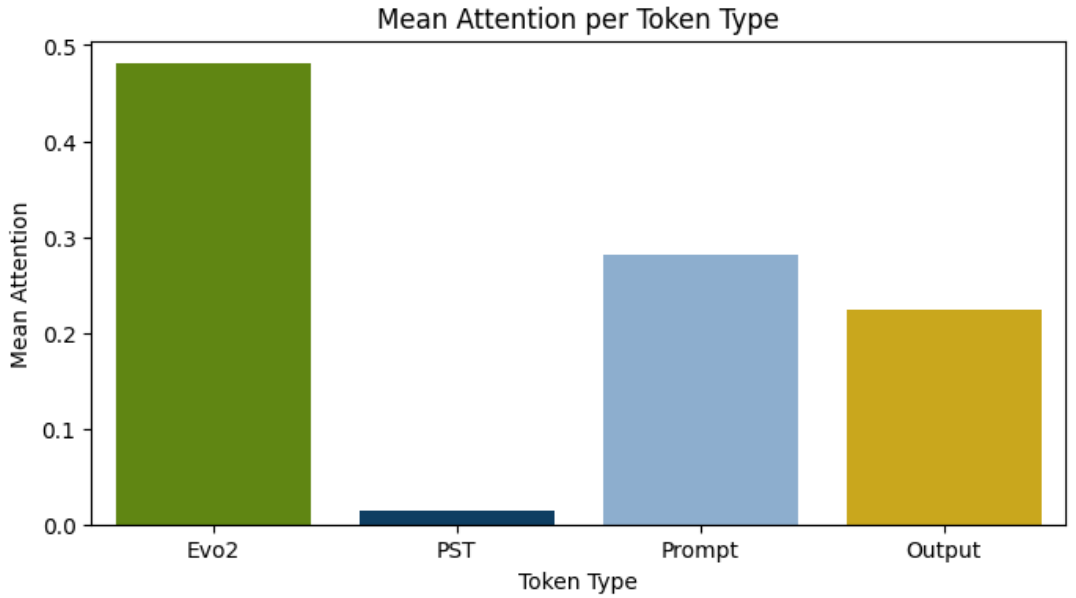
A Multimodal Large Language Model Fusing Natural Language, Genomic, and Structural Tokens



Genolator – From Sequences to Sentences

Genolator Achieves Unified Understanding of Language and Biology

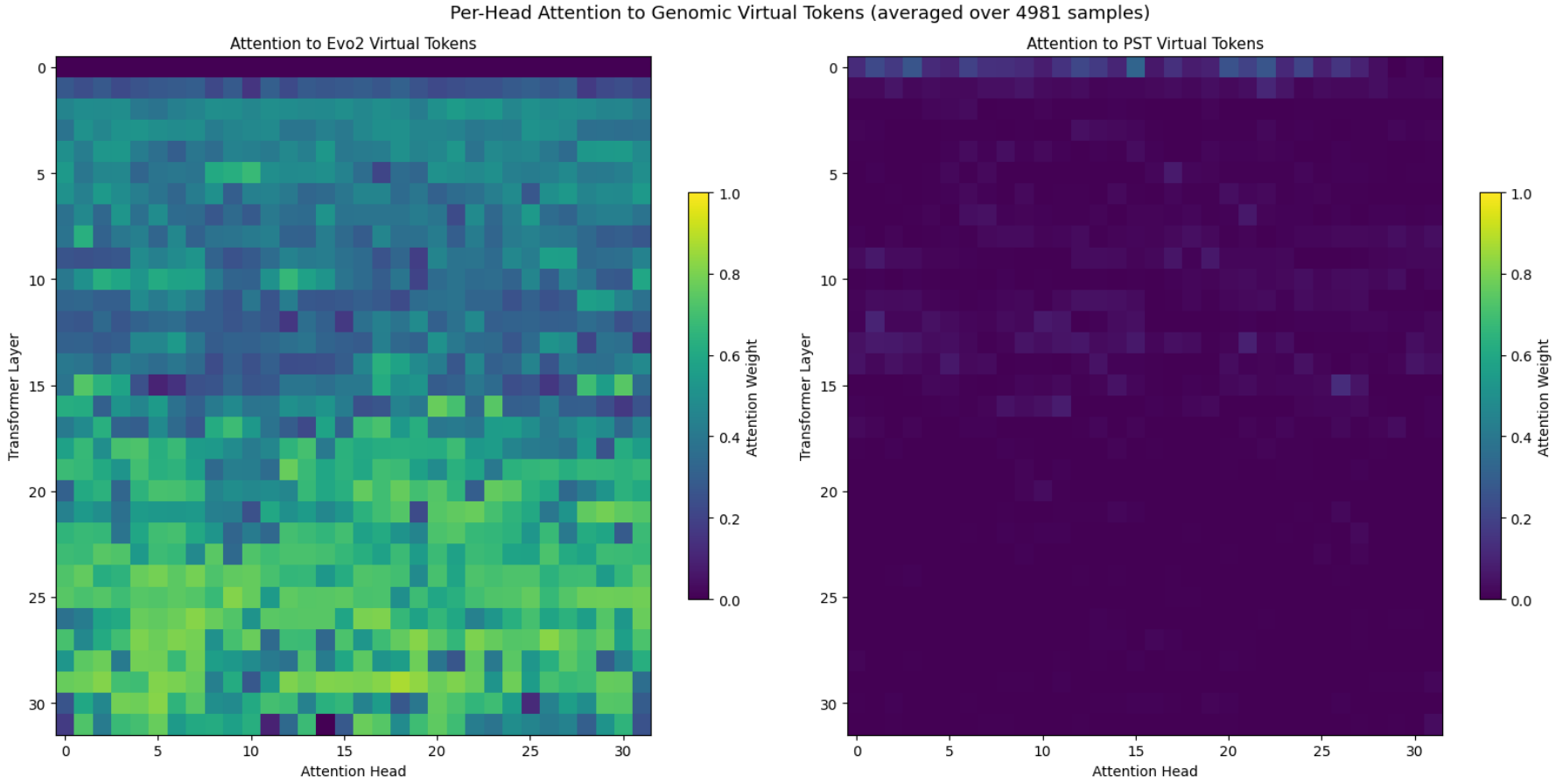




...however, attention to structure tokens did not reflect our observation that multimodality has a strong impact, that's why....

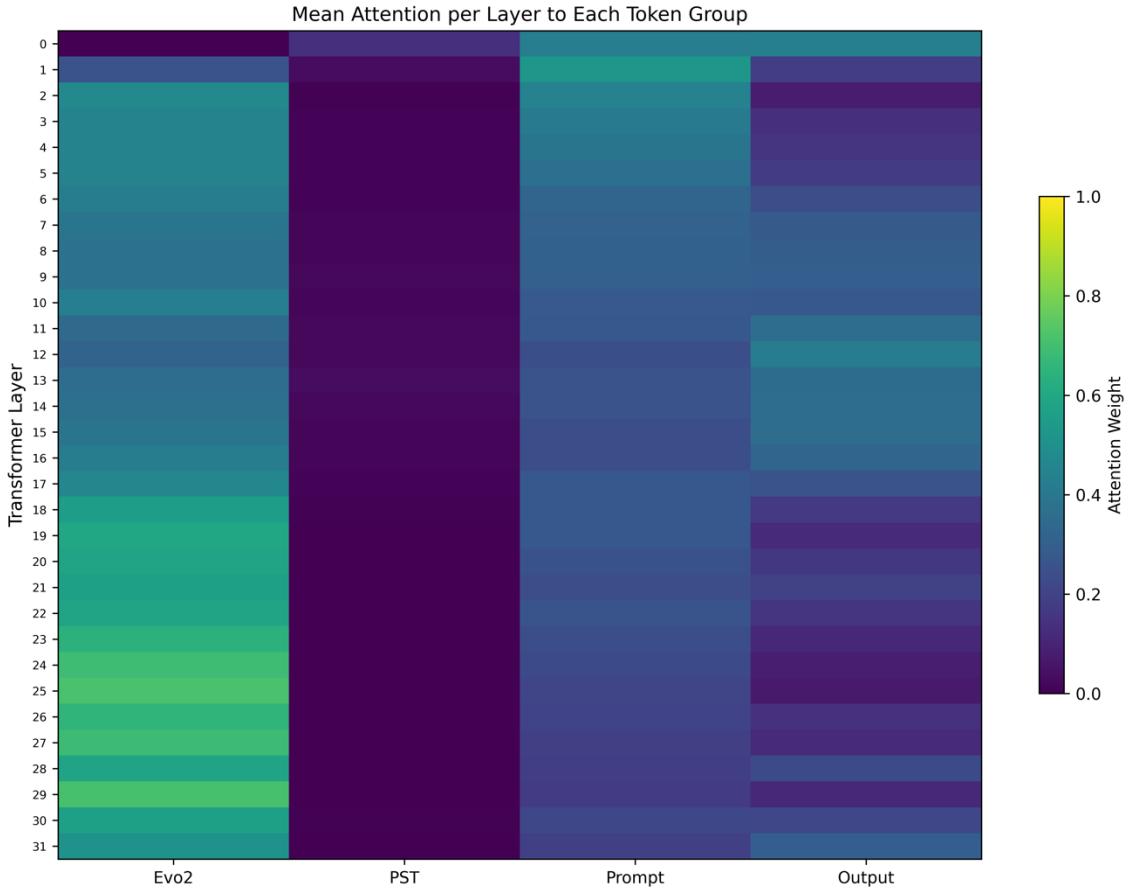
Attention is not Explanation

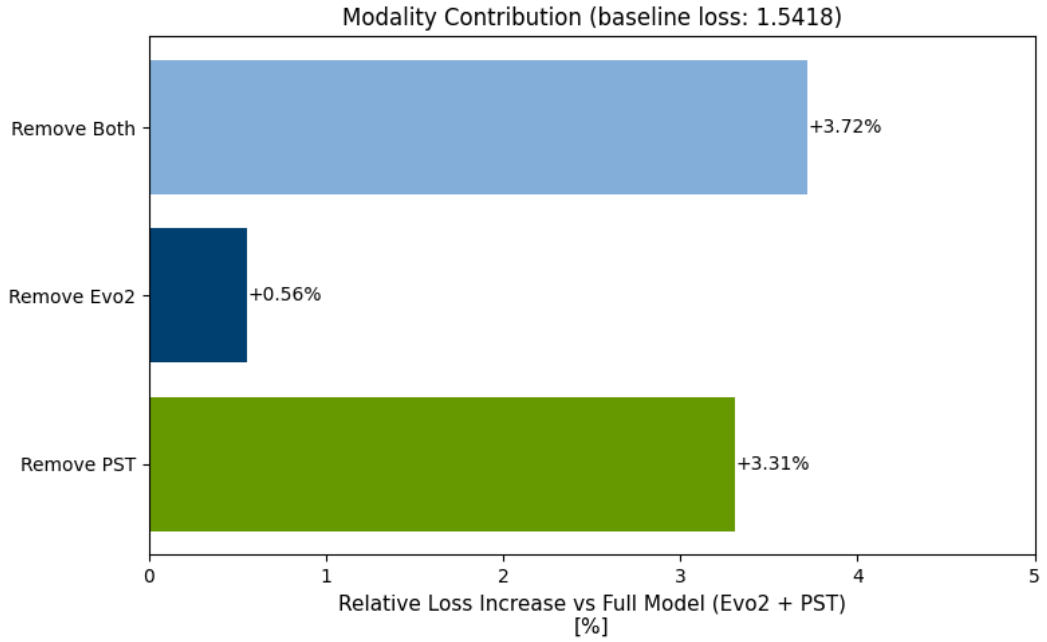
Deeper dive into transformer layers and attention heads



Attention is not Explanation

Ablation study as a causal test

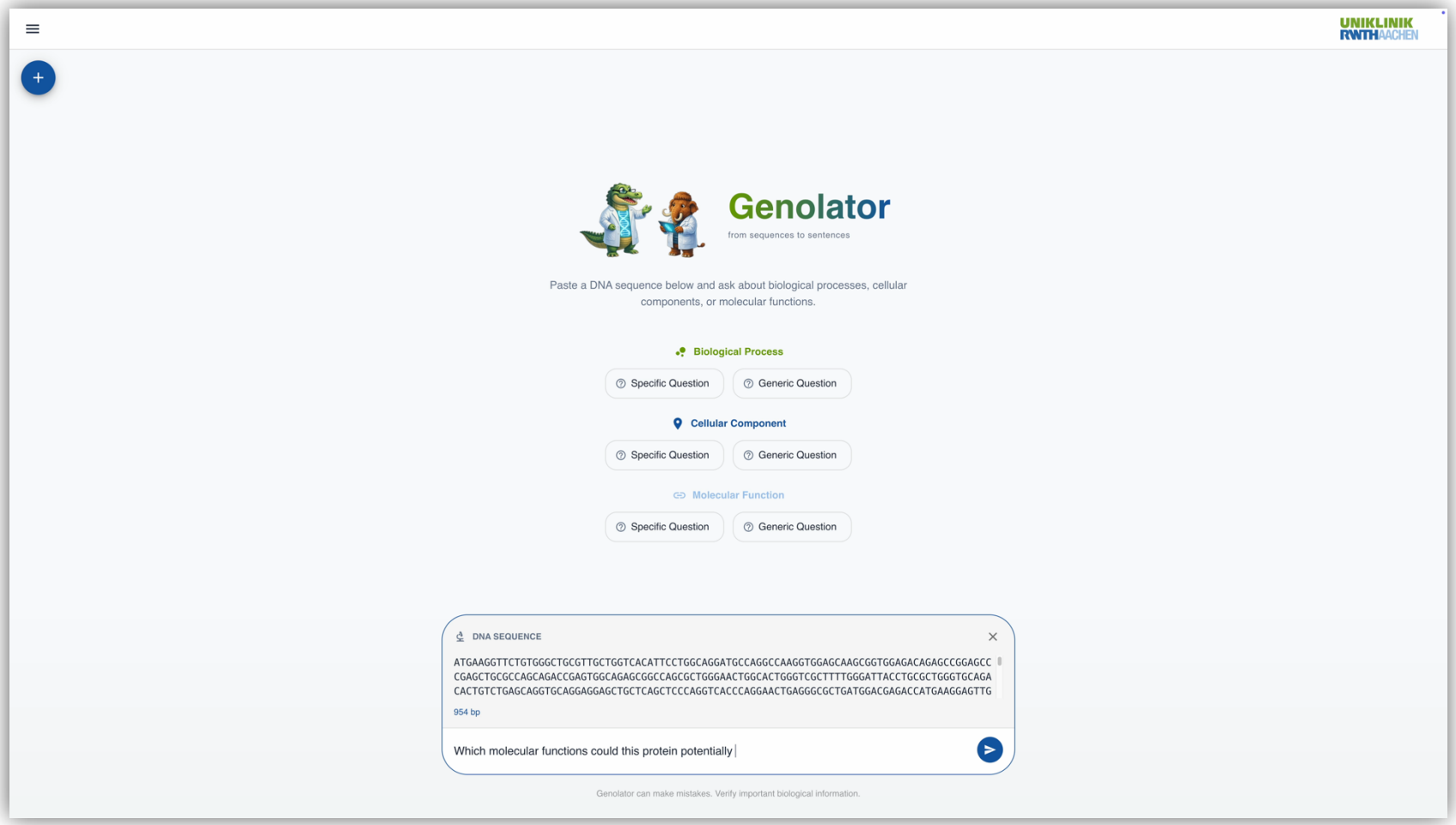




Attention points you where to look.
Causal tests tell you what matters.💡

Genolator – From Sequences to Sentences

Genolator - Translating between Life´s Code and Natural Language



What did we achieve so far?

Research Only Highlights

Data of > 1,875 unsolved cases processed

1.7% additional diagnostic yield that is > 32 cases

> 140.000 Protein Structures predicted

6 high interest not so far known variants identified in the Dark Genome

Congenital insensitivity to pain
(Exact match of variant in 4 patients)



https://www.researchgate.net/figure/Affected-extremities-a-superficial-inflamation-after-long-transfemoral-amputation-b_fig2_297730928

What comes next?

Multimodal language models as a new interface for genomic data analysis



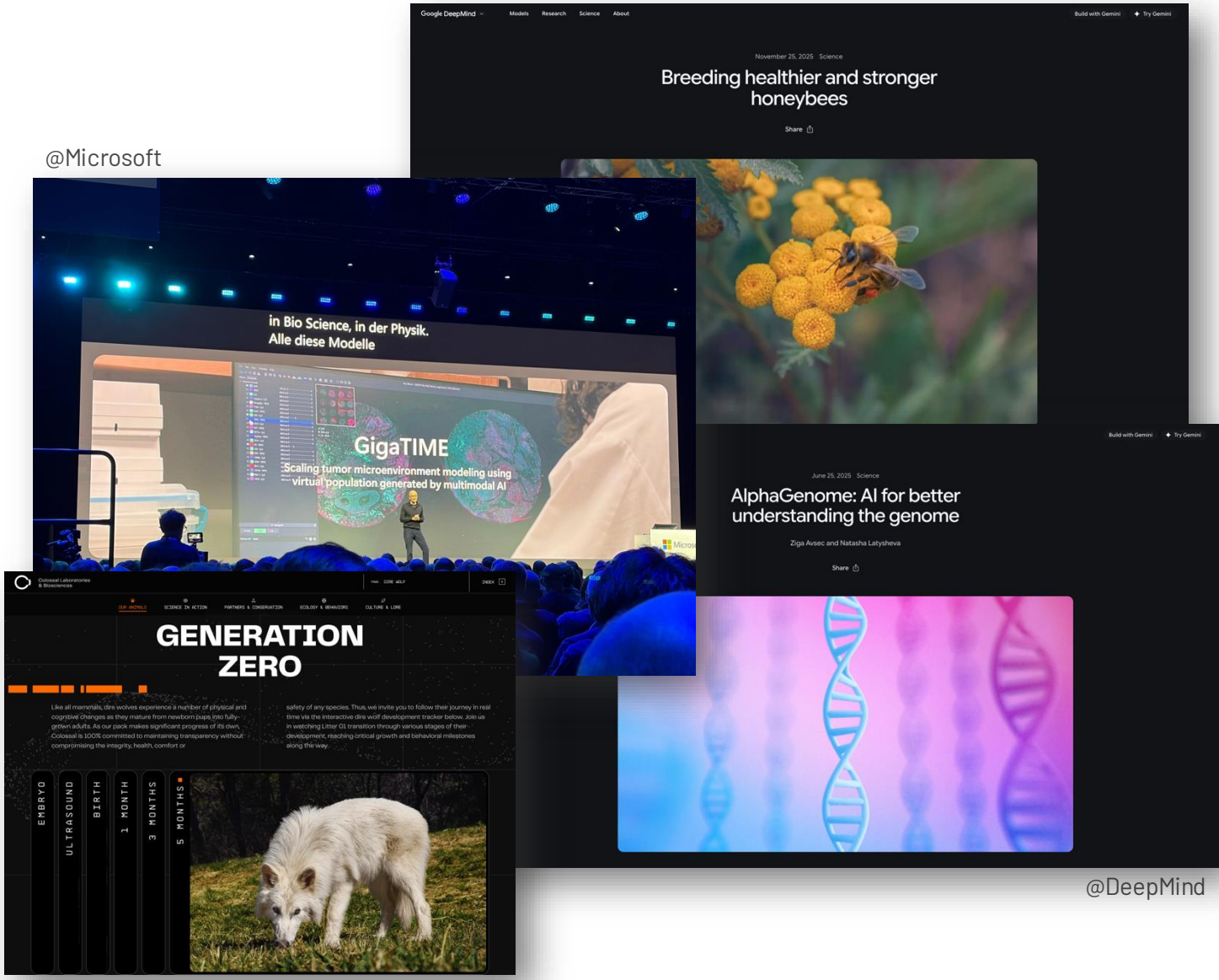
„Chatting with patient genomes“

AI in Healthcare is rapidly growing

AI in Healthcare – A Glimbse into a New Era

AI in Healthcare

- $35 - 40 billion in 2025
- $56 billion in 2026
- $700 billion in 2035



@Microsoft

@DeepMind

@ Colossal Laboratories & Biosciences

All great projects are based on team effort

and so is Genolator!



Get in contact!

Looking forward to the conversation

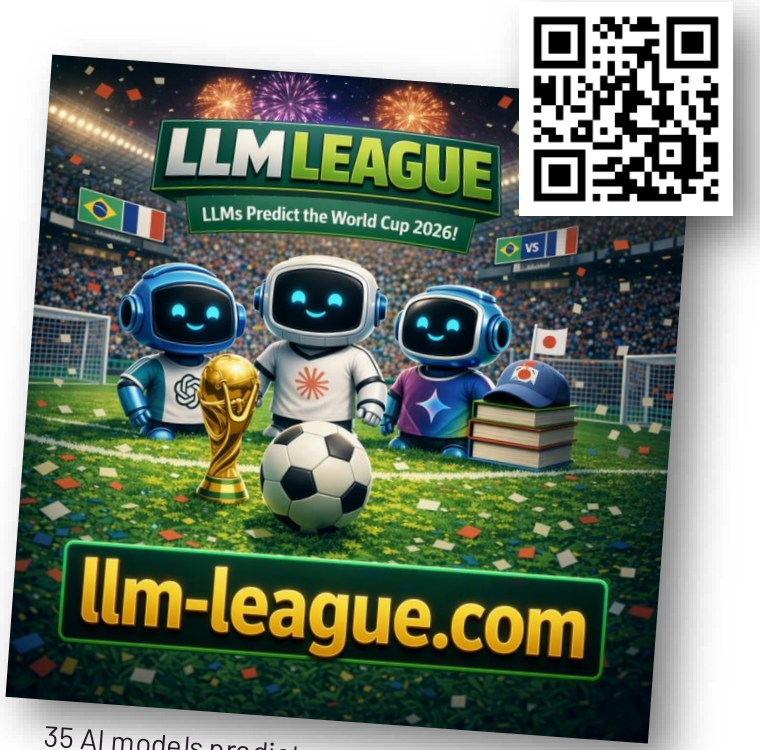


Martin Danner

Senior Data Scientist & ML Engineer / PhD Student

scieneers GmbH / Uniklinik RWTH Aachen

martin.danner@scieneers.de



35 AI models predict every World Cup match.
Who reasons best under genuine uncertainty?

