

AI in Biomedicine: Trends and challenges

Guillem de Valles Ibáñez, PhD
Iván Díaz Álvarez, PhD

HPCKP'25, 3-5 June, Barcelona



Guillem de Valles

Computational Scientist

Guillem has a background in bioinformatics and currently works at DoITNow (NZ branch) deploying scientific software while we sleep



Iván Díaz

HPC Support Engineer

Iván, yours truly, works for DoITNow ES branch, and works on HPC and research support doing devops and sysadmin work

What is AI and why everyone is hyped about it?

What we really mean with AI?

- An umbrella term over the decades
 - Rule based systems, Neural Networks, ML, Deep Learning
- What it means today
 - Large Language Models (LLMs)
 - Transformer - multilayer perceptron paradigm
 - Predict the next most-probable response token sequentially
 - Diffusion models
 - Continuous data, images, video, music
 - Can have transformers / integrate with LLMs
 - Based on noising/denoising architecture
- Two phases:
 - Training phase - against labelled data, Reinforced Learning on Humans or other LLMs
 - Inference phase - usage phase, no “learning”



Midjourney

Why the sudden hype?

- Recurrent Neural Networks (RNN) widely used on NLP
 - *Word2vec* (2013) vector embeddings on highly dimensional manifolds
 - Generative capabilities, but gibberish when output grew in size
- Seminal 'Attention is all you need' (2017) Google paper
 - Transformer architecture, attention-based NN
 - Tokens can add meaning to following ones
 - Queries, Keys and Values all trainable
 - Large context windows, solves long-term memory
- Hardware GPU advancements
 - Transformers promising, but still not enough
 - Scaling LMs to billions of parameters (LLMs)
- Improved User Interaction
 - The spark needed for the AI boom to take place (from GPT → ChatGPT)
 - Chatbots changed fundamentally the user interaction, everyone could interact with them



AI on biomedicine milestones

2012



AlexNet

AlexNet wins ImageNet'12 computer vision challenge

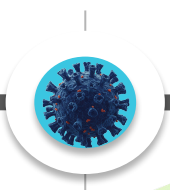
2018



IDx-DR

IDx-DR 1st FDA-authorized AI tool independently diagnose diabetic retinopathy with no human oversight

2020



COVID

ML instrumental in the research and manufacture of vaccines, clinical trials (e.g AstraZeneca)

2021



AlphaFold

DeepMind's AlphaFold2 accurately predicts protein folding

2022



Foundation models

Domain specific dedicated foundational models like BioGPT and PubMedBERT

2024



Generative AI drugs & therapies

AI generates novel molecules, antibodies, and gene therapies.

Current developments

DeepSeek



- DeepSeek R1 took the AI world by surprise on Jan, 2025
 - State-of-the-art reasoning model from China
 - Orders of magnitude lower training price (\$5M vs \$100M+)
 - Dramatically reduced inference compute and memory requirements (e.g. KV cache optimizations)
 - Based on unsupervised RL techniques with minimal human intervention
 - Open Source weights and very open paper publications.
- What it means for the field
 - Lower training and inference costs significantly, small institutions to have on-prem AI
 - Enabling a push for more openness, innovation and competition on the AI field.
 - Baked-in conformance with Chinese regulations

DeepSeek (II)



- Issues
 - Security vulnerabilities, leaks and conformance issues with GDPR/HIPAA
 - Lacks robust ethical guardrails, comparatively easy to jailbreak
 - Possible issues by being trained against other competitor modes
- Upcoming DeepSeek R2 will offer improvements on
 - Multilingual reasoning, with clearer separation
 - Multimodal reasoning
 - Better source code generation performance
 - Generative Reward Modelling, a method to generate its own feedback on training
 - Self-principled Critique Tuning

Reasoning models

- LLMs basically work by predicting the most probable sequence of tokens to follow the prompt and attached context sequentially
- Fine for small problems, issues with larger problems since the computation effort is the same
- Chain of Thought (CoT), early prompt engineering hack
- Reasoning models
 - Generate intermediate CoT outputs
 - Final answer uses intermediate as input
 - Reasoning time controlled by length of CoT, can be tuned on training
 - This elevates cost both in training and in inference.

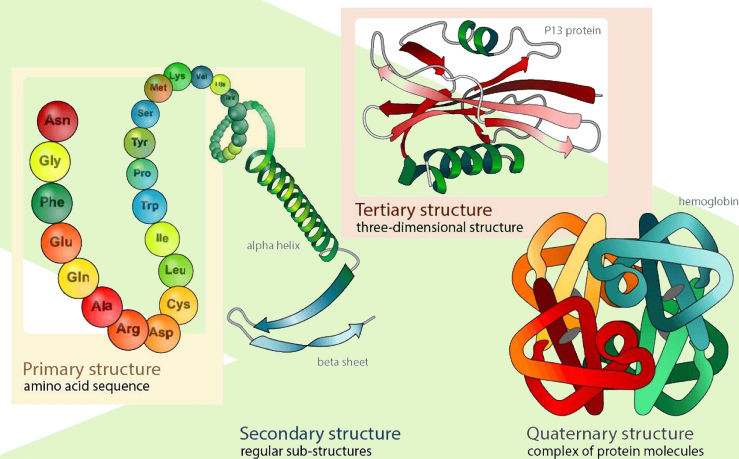
Retrieval Augmented Generation

- Context windows typically in the 100K token range (75-100 pages).
- Reference documentation bounded by context window
- RAG ingests documents using “Embedding model”
 - Convert and store those as vector embeddings in a special vector database
 - The embedding model is independent to the LLM used
- RAG Retriever
 - Search the DB for vectors matching the user prompt
 - Retrieve original text snippets, attach them to the prompt context window
- Access to enormous amounts of data,
 - Update and curate them without having to retrain the model
 - Crucial for the field
- For smaller document corpus, CAG can also be used
 - All the documents converted into preprocessed attention KV
 - Preloaded as the initial state before the prompt

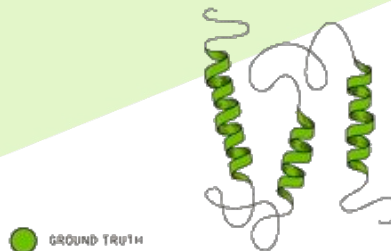
Some use cases in Biomedicine

Protein Folding

- AlphaFold is the most widely successful application of AI in the field
- It is not an LLM, more of a transformer-diffusion custom model with specific cross-attention for both chemistry-physics and existing biology data.
- It predicts accurately the 3D folded structure of proteins starting only from a sequence of amino acids
- Won the CASP 13 challenge and then exploded the number of known protein structures from ~150K to 200M and merited the 2024 Nobel Prize in Chemistry



Global Distance Test



Drug and protein discovery

- **Relevant protein identification:** AI analyzes vast biological datasets to pinpoint disease-relevant proteins
- **Molecule Virtual Screening:** Rapidly evaluate billions of molecules for binding affinity (e.g., Atomwise, Schrödinger's ML-guided discovery).
- **De Novo Molecular Generation:** GNoME (Google DeepMind) and Chroma (Generate Biomedicines) create novel drug candidates.
- **Generative Drug Design:** Predict new protein structures, enabling rational drug design



BPGTM bio



Recursion



Exscientia



Atomwise



iktos



RELAY



insitro



AnimaBiotech



Insilico
Medicine



Generate:Biomedicines



Benevolent^{AI}



Isomorphic
Laboratories



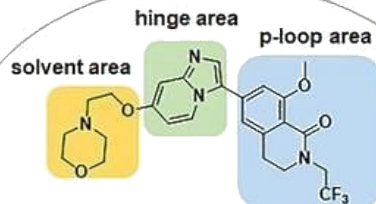
Latent Labs



Cradle

Drug and protein discovery (II)

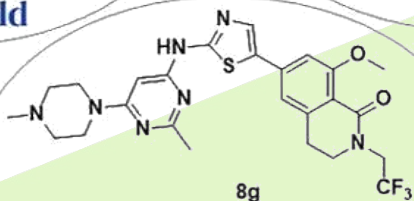
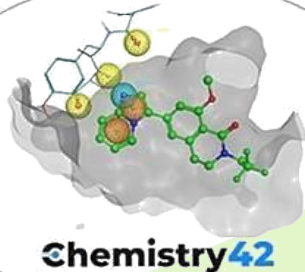
Binding Pose Prediction



3, GLPG-3970
*SIK1/2/3, IC₅₀ = 5048/114/35 nM

AlphaFold

Hinge Cores Generation



*SIK1/2/3, IC₅₀ = 140/0.7/17 nM
✓ Good AMPK kinases selectivity
✓ TNFα, IC₅₀ = 30 nM
✓ IL-10, EC₅₀ = 148 nM

*[ATP] = 1 mM

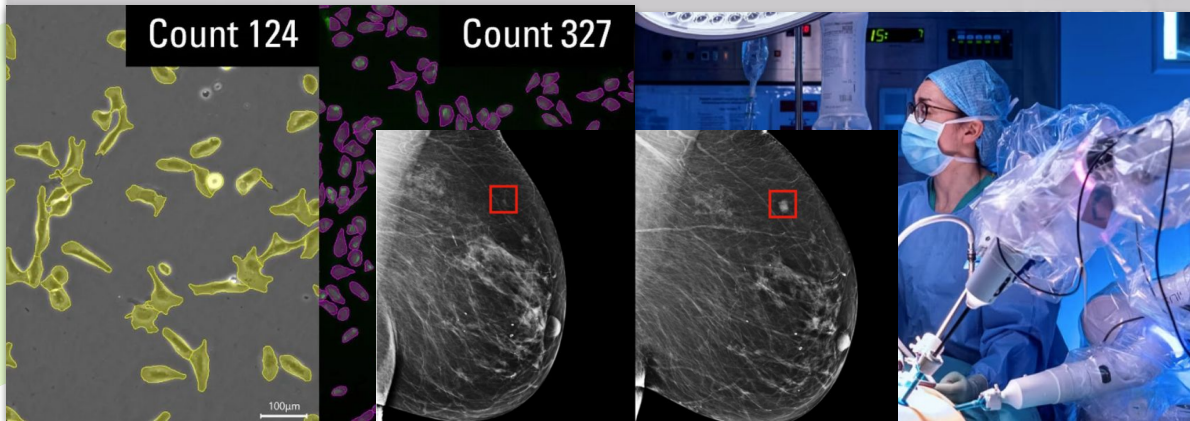
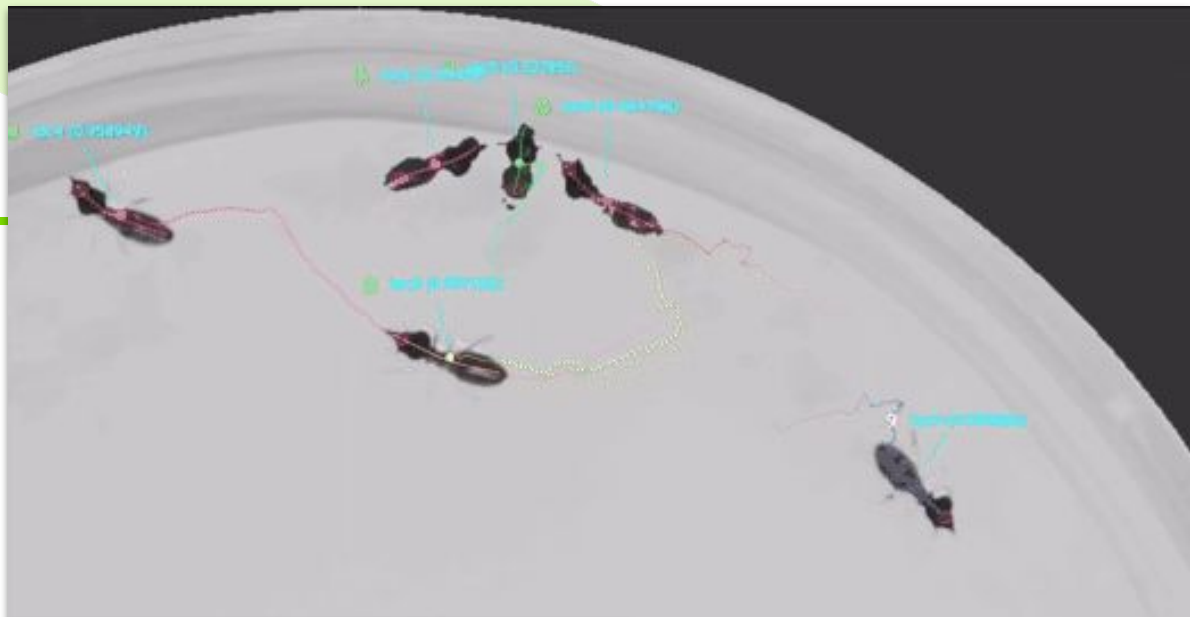
SAR Exploration

Literature example of finding enzyme inhibitor for SIK2 Kinase with several generation tools

- Involved in glucose metabolism
- AlphaFold to predict structure
- Chemistry-42 to generate hinge cores → 7f hit molecule
- SAR → 8g molecule with superior potency against SIK2

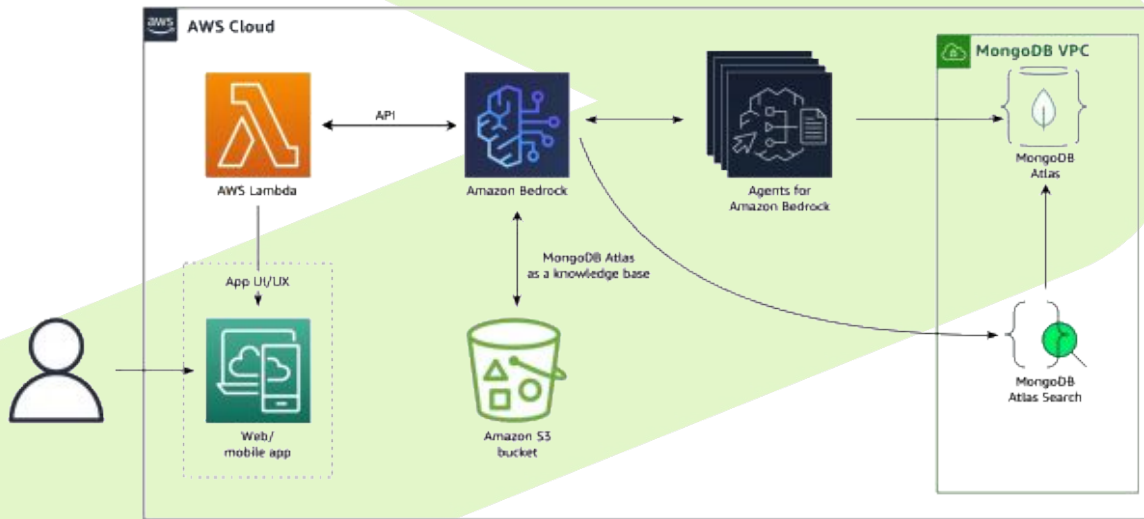
Image Processing

- Medical image analysis
 - Tumor detection
 - Fracture identification
 - Disease detection
 - Retinal analysis
- Patient id & monitoring
- Monitor hygiene
- Automated cell counting
- Surgical guidance
- Animal behaviour



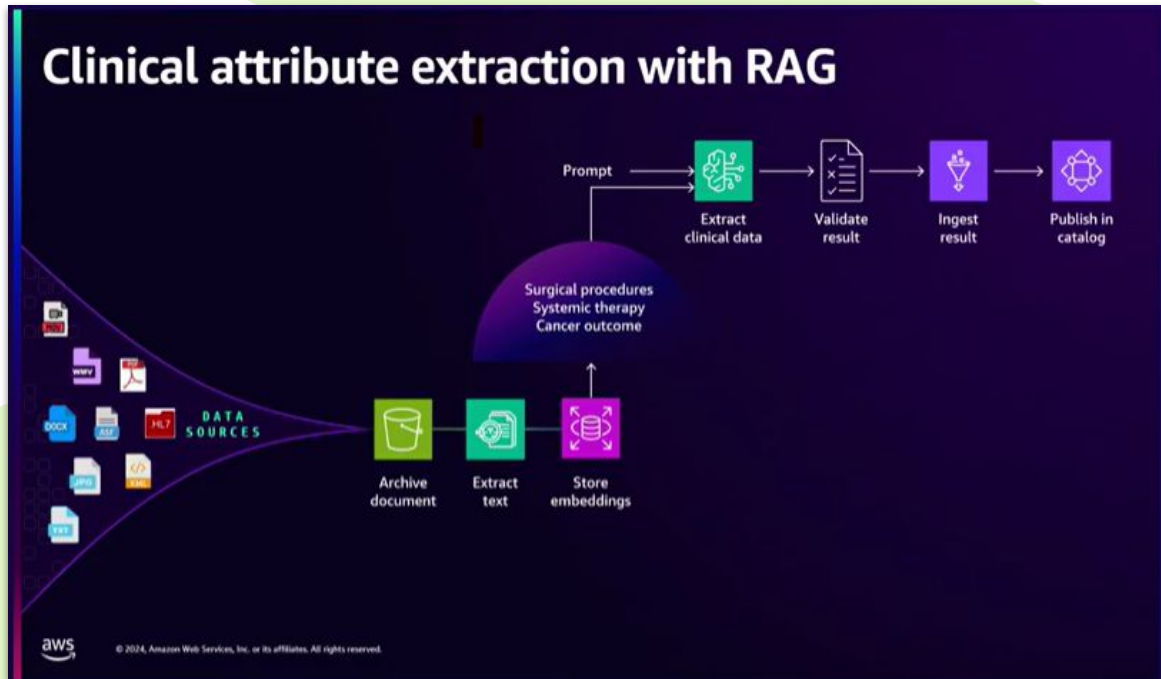
Clinical Documentation

- Novoscribe (Novo Nordisk)
 - MongoDB Atlas with RAG
 - Compile Clinical Study Reports,
 - Makes them in hours instead of weeks like previously
- Signatera (Natera)
 - Personalized patient MRD testing
 - RAG to extract metadata from raw clinical data documents
 - LLMs to extract the relevant data to a catalog
 - WDL workflows on AWS HealthOmics to create



Clinical Documentation

- Novoscribe (Novo Nordisk)
 - MongoDB Atlas with RAG
 - Compile Clinical Study Reports,
 - Makes then in hours instead of weeks like previously
- Signatera (Natera)
 - Personalized patient MRD testing
 - RAG to extract metadata from raw clinical data documents
 - LLMs to extract the relevant data to a catalog
 - WDL workflows on AWS HealthOmics to create



Scientific Literature Review

- Literature review & discovery
 - AI-powered web search engine
 - Summarization and analysis
 - Filling research gaps
- Writing and editing
- Data analysis
 - Tabular data ingestion and analysis
 - Ingestion patent documents
- Citation management
- Collaboration



IRIS AI



Semantic **Scholar**



Coding aids

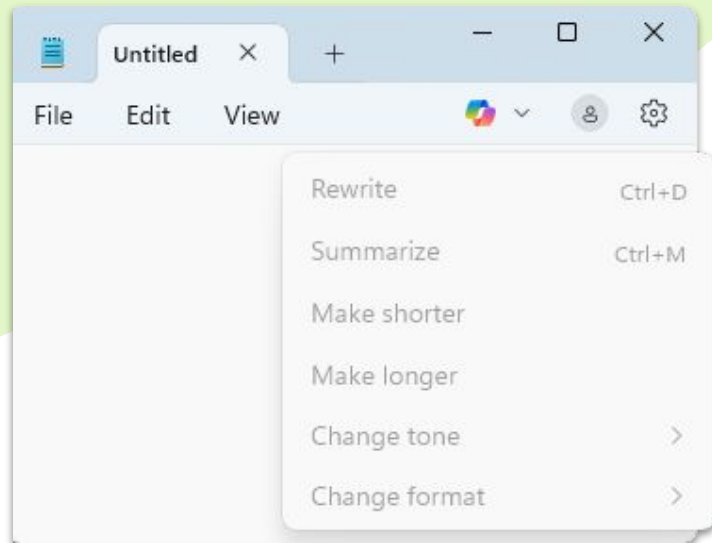
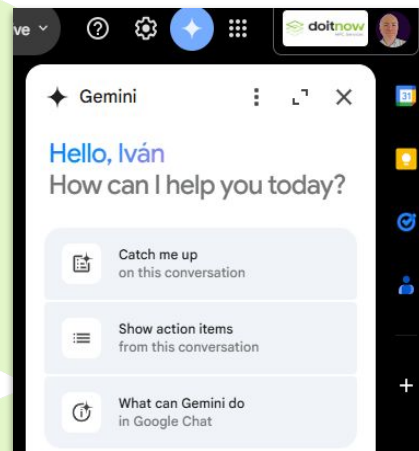
- LLMs trained on source code
 - Github Copilot, CodeWishperer, Code LLama, Anthropic Claude...
- Benefits
 - Avoids wasted time on boilerplate code
 - Spot common errors and help on debugging them
 - Programmers can concentrate on creative and high-level tasks
- Risks
 - Reliability, security, vulnerabilities.
 - Can increase technical debt
 - Can become a crutch
- Augmentation vs Automation
- Code Agents
 - Copilot Agent Mode, Open AI Codex

```
1
2 def common_prefix(a, b):
3     """Return the common prefix of two lists."""
4     if len(a) < len(b):
5         return common_prefix(b,a)
6     for i in range(len(a)):
7         if a[i] != b[i]:
8             return a[:i]
9     return a
10
11
12
```

```
0 references | 0 changes | 0 authors, 0 changes
39 public static void CreateTables()
40 {
41     using (var context = new TaskContext())
42     {
43         context.Database.ExecuteSqlRaw("CREATE TABLE tasks (id INT PRIMARY KEY, title VARCHAR(50), priority INT)");
44     }
45 }
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
```

Productivity

- Not all flashy stuff
- LLMs also used to generate internal documentation, meeting notes, drafts, spreadsheets, etc
- Integrated on existing tools
 - Google Workspace integrating Gemini
 - Even Windows Notepad has Copilot now
- Big incentive for internal vetted, secure implementation
- Automated meeting minutes
 - Full transcripts can be limited
 - Participant consent issues
 - Stifling discourse



Concerns and Regulation

AI: Great opportunities carry great concerns

- **Security:**
 - Everyday biggest concern in the field
 - Risk of data leakage, prompts and context data sent to third parties.
 - AI can promote bad security practices (partial information, no organization alignment)
 - Jailbreaks, no strong security compartmentalization guarantee
- **Privacy**
 - Pervasive surveillance (example Microsoft Recall)
 - Extensive personal data collection
 - No data minimization
 - No transparency
- **Lack of Human Oversight:**
 - Risk eliminating human oversight from the loop and arrive at bad outcomes
 - AIs are black boxes, so there is an intrinsic lack of transparency
 - Can over-optimize without having into account human values

AI: Great opportunities carry great concerns (II)

- **Systemic Biases:**

- Data biases, social ones, perpetuating discrimination.
- Very insidious since they are intrinsic to the training data.
- Ethic guardrails don't totally prevent those.

- **Hallucinations:**

- General problem, specially troubling in the field
- Controllable via temperature, personality prompt and human oversight.

- **Sycophancy:**

- AIs can be “too nice”
- Very bad UX
- GPT-4o update rollback
- Causes and promotes delusional thinking
- Shut downs any critical or negative responses

AI Regulation

EU AI Act :

- August 2024, 4 risk category
 - **Unacceptable** (banned): - “Dystopia Tier”, manipulate humans, public biometric id, profiling, facial emotion reading, social score
 - **High** (transparency, oversight, safety): health and safety concerns, impact assessment, evaluated before commercial life. All medical systems here
 - **General purpose**: Foundational models, transparency, high systemic risk ones (e.g. >100 PetaFlops) must present assessment and be registered in a High-risk AI DB.
 - **Limited**: Mostly generative AI, basic transparency
 - **Minimal**: Not regulated
 - More lenient with Open Source models
- Governance institutions (AI Office, AI Board, Advisory Forum)
- Penalties of 35M€ or 7% annual global turnover for worst cases.

AI Regulation (II)

- **US:** NO Federal Laws at this moment, only state laws
 - Colorado Senate Bill 24-205 “Consumer AI Act”, limited scope, Texas and Illinois have similar laws
 - Covers “algorithmic discrimination” of “consumers and workers”
 - Transparency, minimal documentation
 - Small 20k\$ fines no consumer right to private legal action
 - Tennessee ELVIS Act (Ensuring Likeness Voice and Image Security Act) :
 - Extremely limited scope, protects musicians from the unauthorized cloning of their voice
- **China:** Interim Measures for the Management of Generative AI Services (2023)
 - Only for public-facing AIs
 - Transparency, content moderation, anonymization, minors time control
 - Training data from legitimate sources respecting IP, authentic, accurate and diverse
 - Small 100K RMB (~14K\$) fines, more a legal frame to spur AI growth



doitnow

HPC Services

Thanks!
Time for Q & A

Bonus-1: Embedding space

- Input tokens converted to vectors in a high dimensional manifold
- GPT3 has a 12k dimension embedding space, each dimension can represent a semantic continuum
- Words that are close in one set of dimensions are related
- Vectors ops
 - Sum to compose meaning
 - Dot product to compare them

