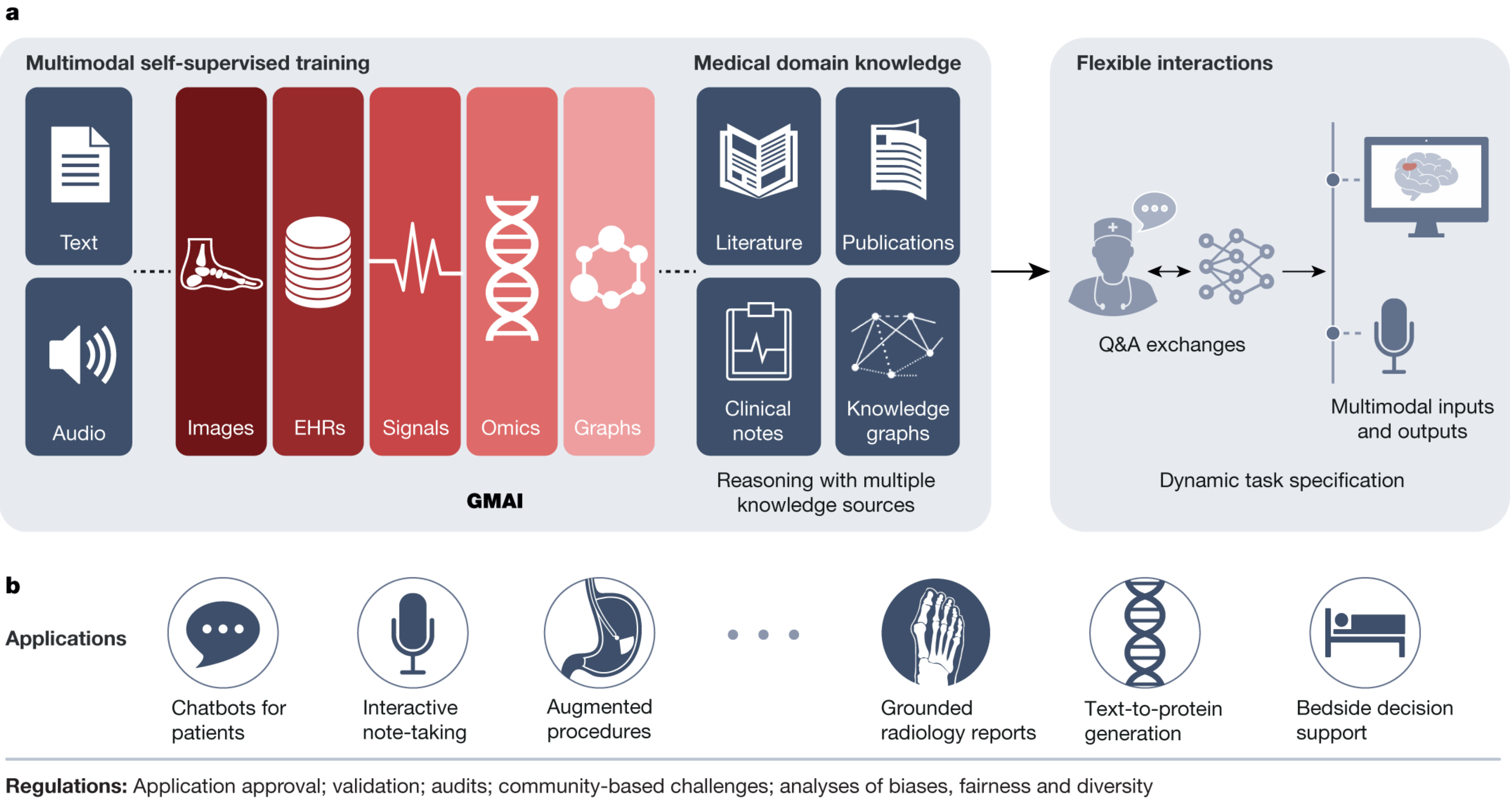
LLM in Healthcare

Foundation models

Large and reusable AI model trained on enormous quantities of unlabeled data and generalized to any tasks



What does Large Language Model (LLM) do?

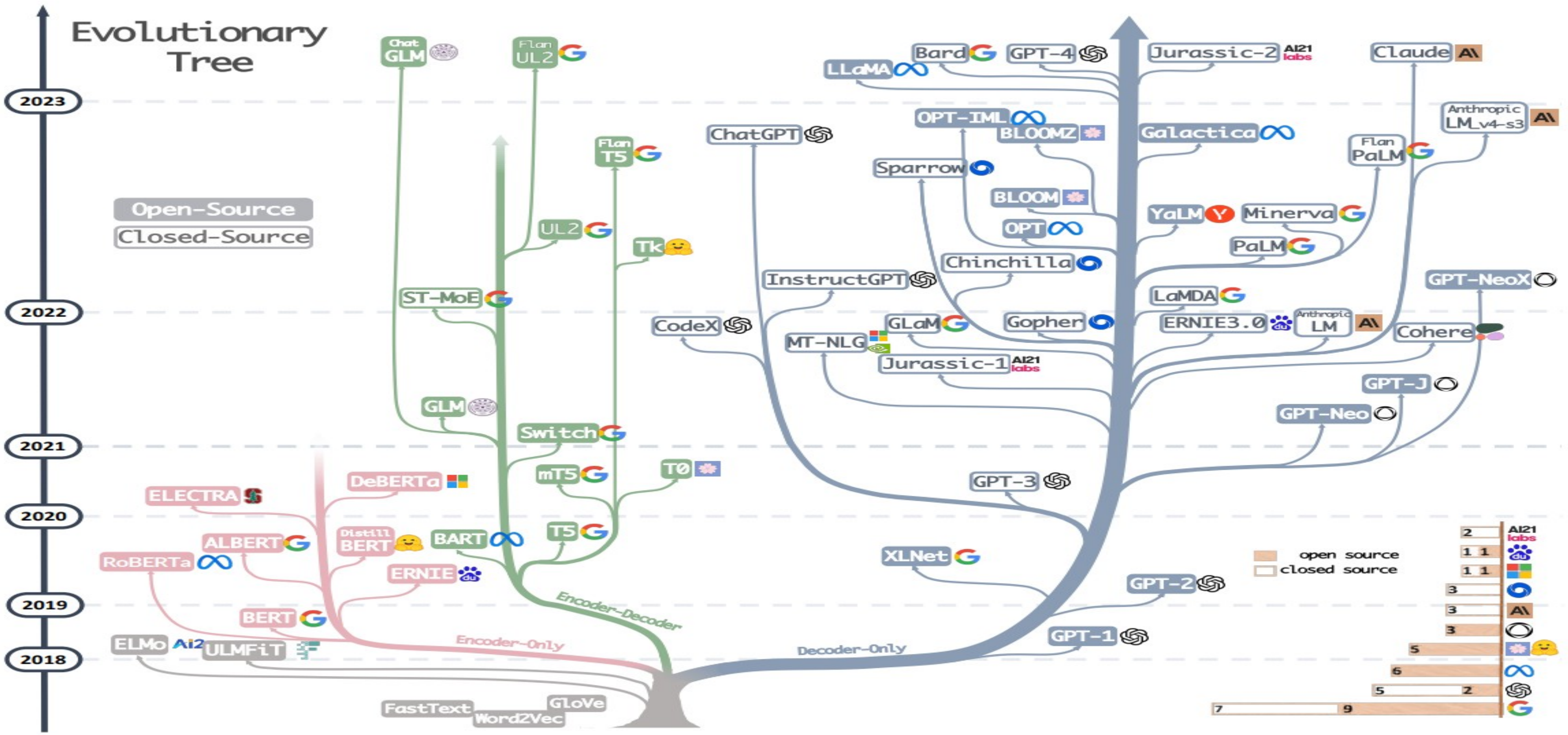
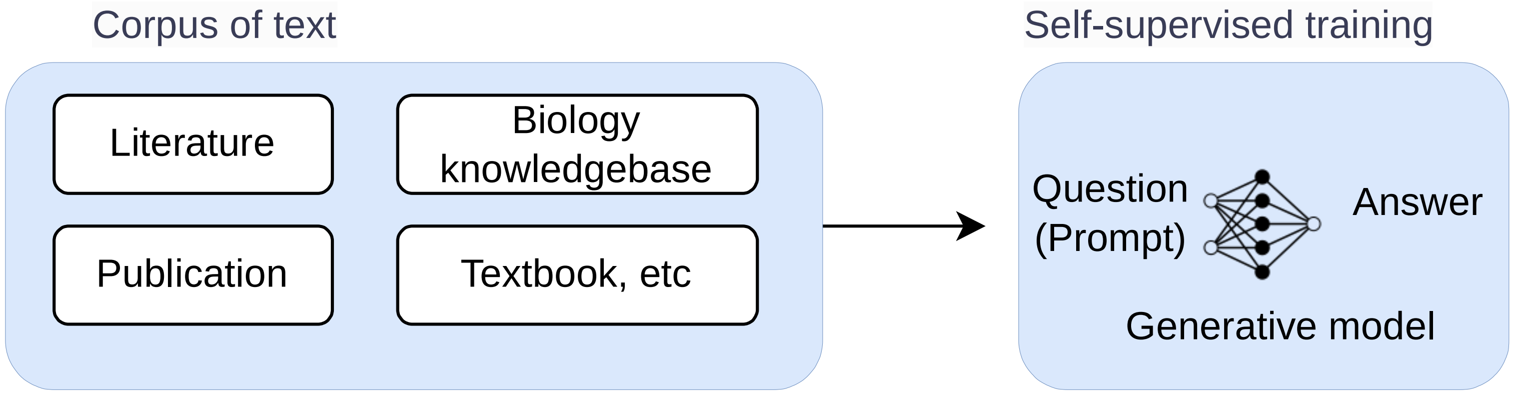


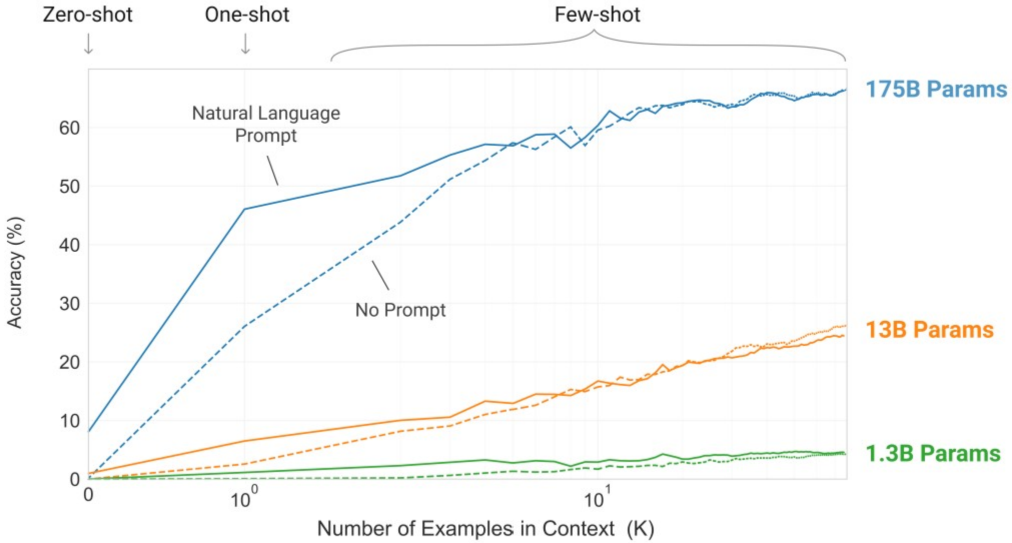
Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.

https://arxiv.org/abs/2304.13712

“Large Language Models are Few-shot Learners”

How it works





https://arxiv.org/abs/2005.14165

Are LLM good at using biomedical knowledge for reasoning?

Potential of LLM for biomedical prediction tasks

Challenge of biomedical prediction tasks

- Small or no data to train models
- Incorporating prior knowledge helps

What LLMs are good at:

Few-shot or zero-shot learning

Utilizing knowledge encoded in parameters

Let's evaluate whether LLM can do biomedical few-shot predication task

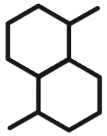
CancerGPT

CancerGPT: Predicting drug pair synergy in rare cancer types

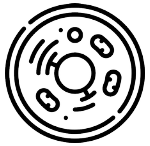
An example of biomedical prediction tasks with limited data

Predicting drug pair's synergistic effect

Drug 1



Drug 2



Cell line

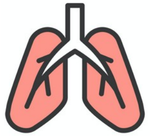
Prediction model

Synergistic effect?

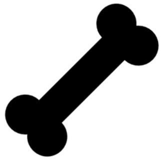
Challenges

Limited data

Data distributions are skewed



Lung cancer cell line



Bone cancer cell line

Soft tissue cancer cell line

https://arxiv.org/abs/2304.10946

1. Drug pair synergy data

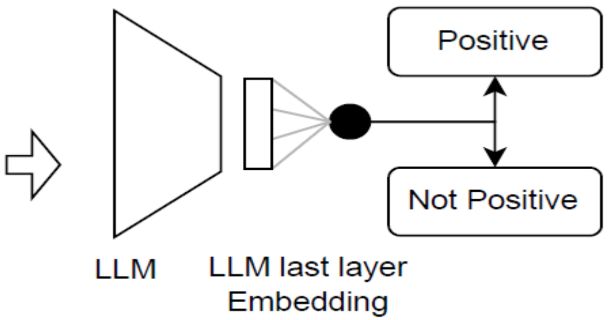
Drug1	Drug2	Cell line	Tissue	Drug1 sensitivity	Drug 2 sensitivity	Synergy
ABT-888	MK-8776	ES2	Bone	-1.625	48.756	<5
Ionidamine	717906-29-1	A-673	Bone	0.568	28.871	>=5
AZD1775	AZACITIDINE	EW-8	Bone	25.687	1.752	?

2. Convert tabular input and prediction task to natural text

Prompt: *"Decide in a single word if the synergy of the drug combination in the cell line is positive or not*

Converted string input: *"Drug combination and cell line: The first drug is Ionidamine. The second drug is 717906-29-1. The cell line is A-673. Tissue is bone. The first drug's sensitivity using relative inhibition is 0.568. The second drug's sensitivity using relative inhibition is 28.871. Synergy:"*

3. *k*-shot finetuning

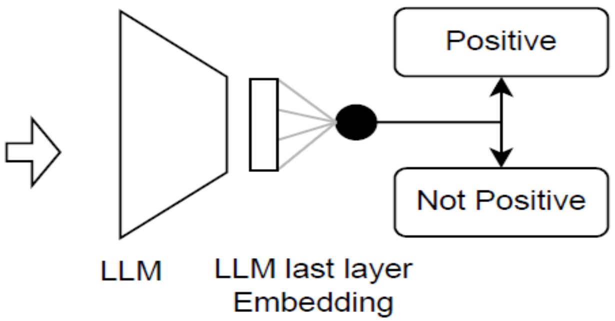


4. Predict drug pair synergy

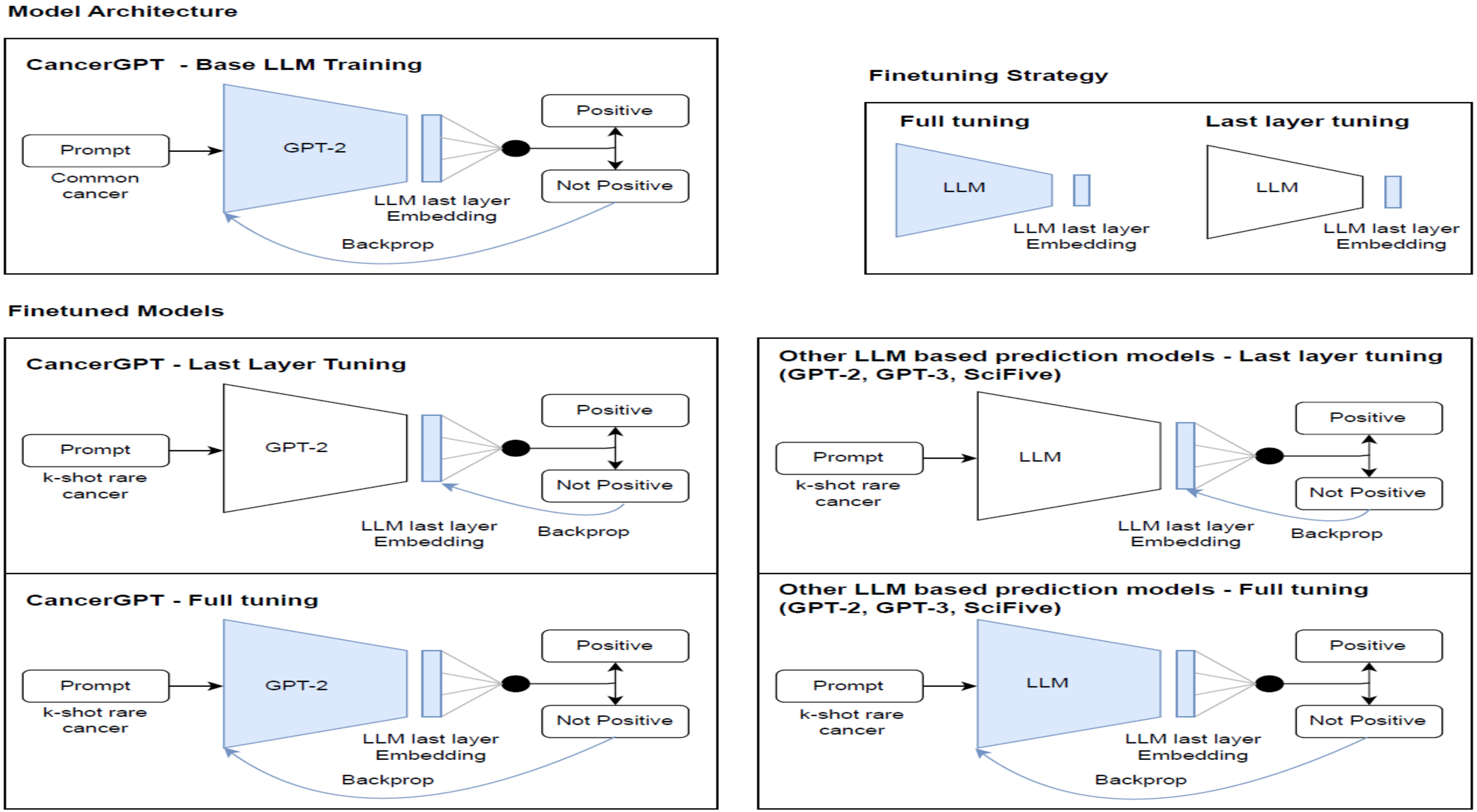
Prompt: *"Decide in a single word if the synergy of the drug combination in the cell line is positive or not"*

Converted string input: *"Drug combination and cell line: The first drug is AZD1775. The second drug is AZACITIDINE. The cell line is EW-8. Tissue is bone. The first drug's sensitivity using relative inhibition is 25.687. The second drug's sensitivity using relative inhibition is 1.752. Synergy:"*

5. Evaluate accuracy

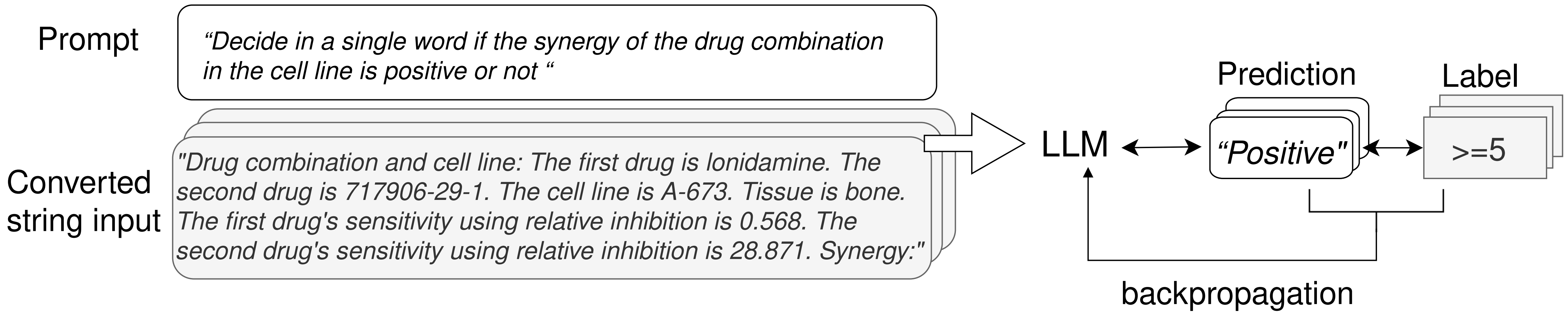


TabLLM, ATSTATS 2023, https://arxiv.org/abs/2210.10723


Model Architecture
CancerGPT - Base LLM Training
Prompt
Common cancer
GPT-2
LLM last layer Embedding
Positive
Not Positive
Backprop
Finetuning Strategy
Full tuning
LLM
LLM last layer Embedding
Last layer tuning
LLM
LLM last layer Embedding
Finetuned Models
CancerGPT - Last Layer Tuning
Prompt
k-shot rare cancer
GPT-2
LLM last layer Embedding
Positive
Not Positive
Backprop
CancerGPT - Full tuning
Prompt
k-shot rare cancer
GPT-2
LLM last layer Embedding
Positive
Not Positive
Backprop
Other LLM based prediction models - Last layer tuning (GPT-2, GPT-3, SciFive)
Prompt
k-shot rare cancer
LLM
LLM last layer Embedding
Positive
Not Positive
Backprop
Other LLM based prediction models - Full tuning (GPT-2, GPT-3, SciFive)
Prompt
k-shot rare cancer
LLM
LLM last layer Embedding
Positive
Not Positive
Backprop

Fine tune the pre-trained LLMs

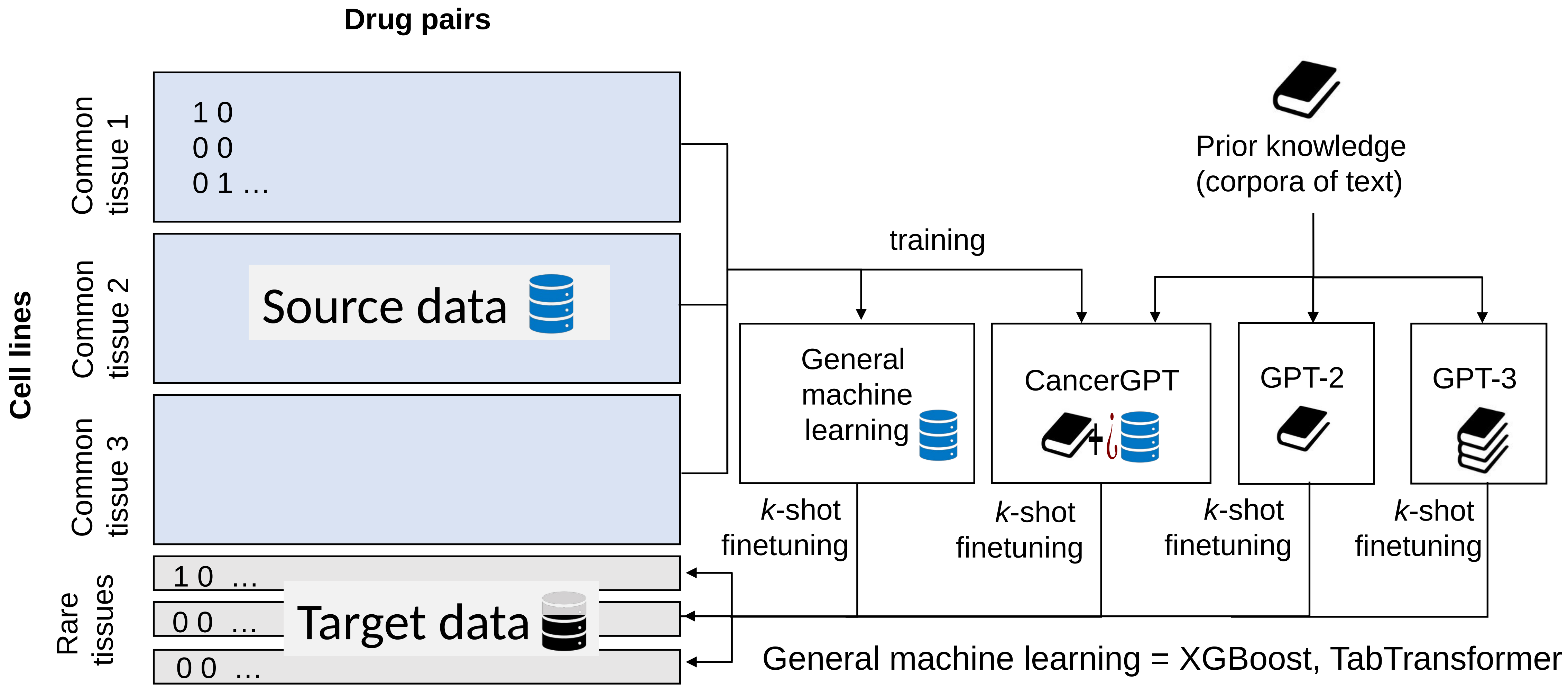
With k-shots of training data



- For GPT2, Add one linear layer on the last token of the GPT2 output and minimize cross entropy loss with binary label
- For GPT3, fine tune GPT3.5 *Ada* using OpenAI API

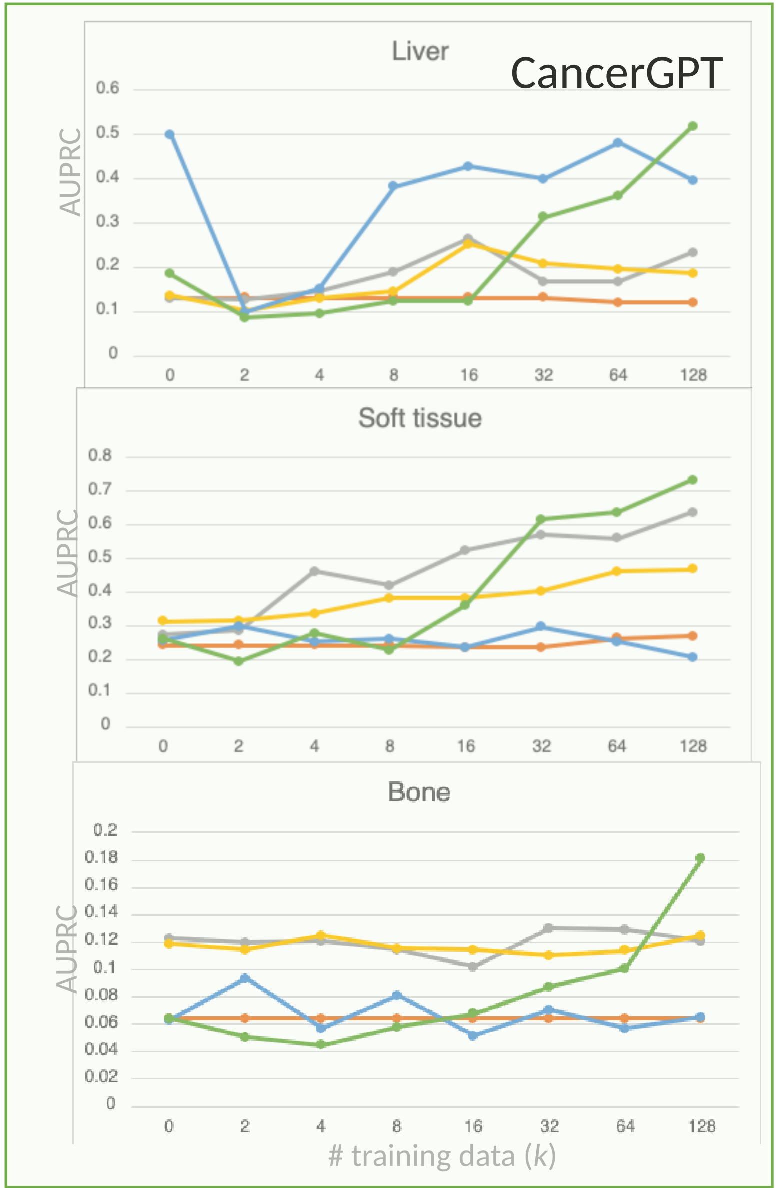
CancerGPT: Further fine tune the pre-trained LLMs

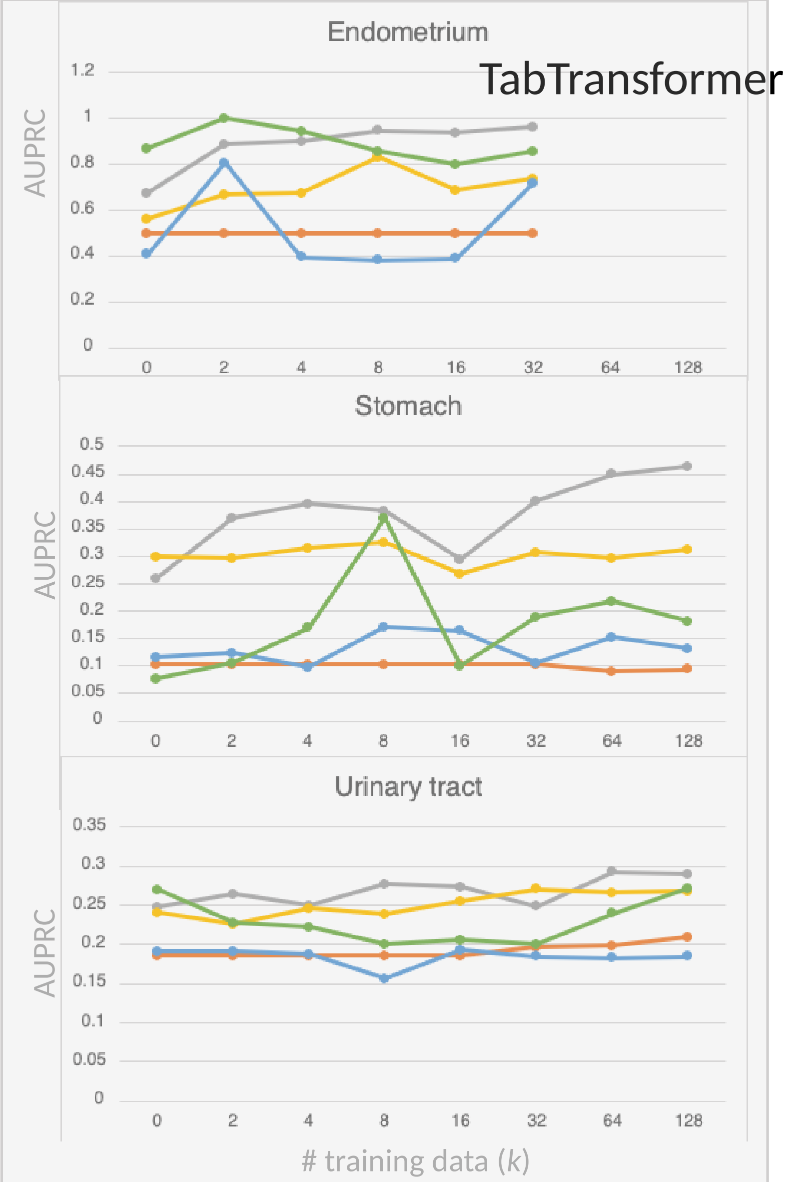
With large data from different sources



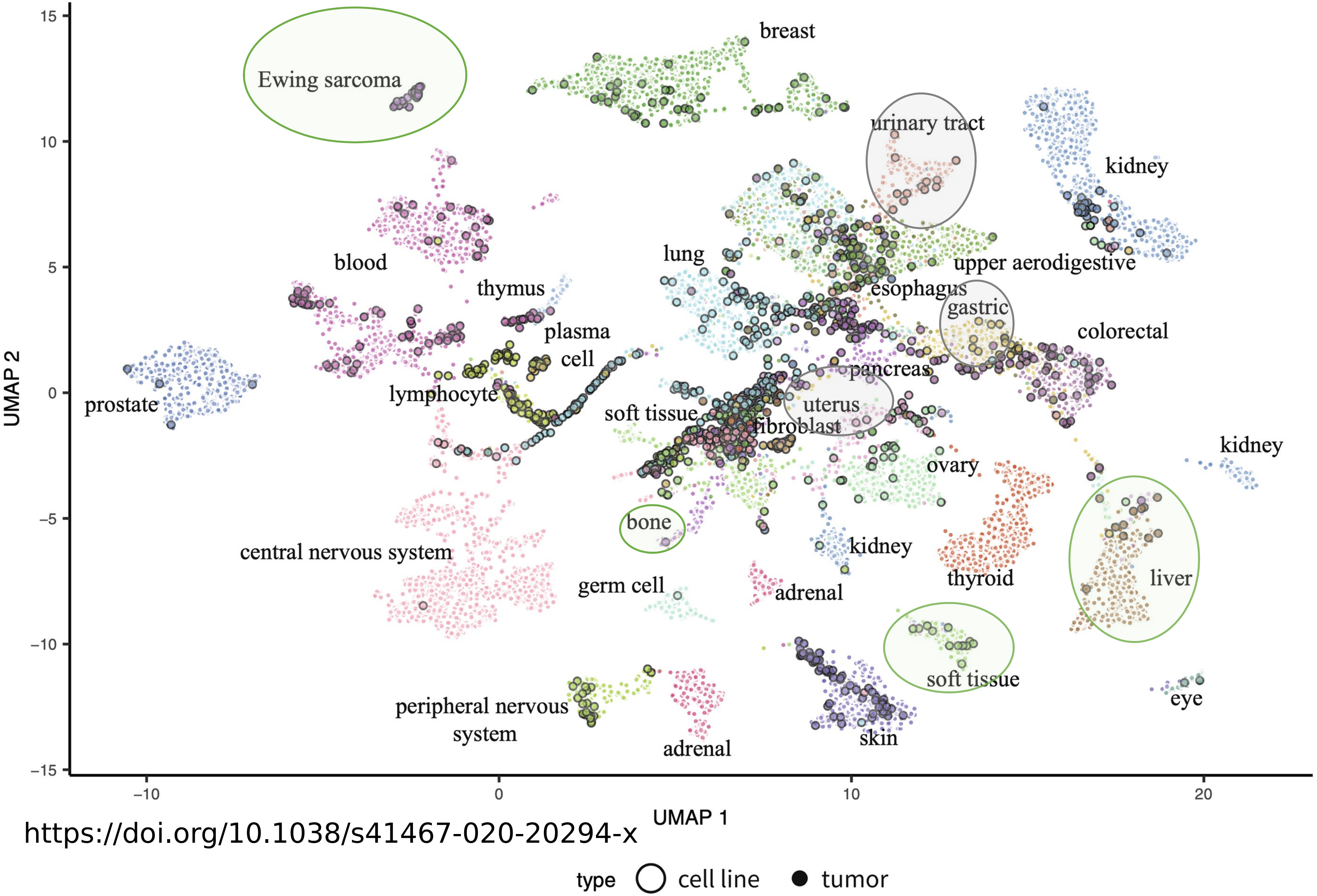
	Methods	0	2	4	8	16	32	64	128
		Number of shots							
Pancreas	XGBoost	0.5					-	-	-
	TabTransformer	0.211					-	-	-
	CancerGPT	0.132							
	GPT-2	0.211					-	-	-
	GPT-3	**0.789**							
Endometrium	XGBoost	0.5	0.5	0.5	0.5	0.5	0.5	-	-
	TabTransformer	0.327	0.571	0.816	**0.939**	**0.939**	**0.918**	-	-
	CancerGPT	0.551	0.571	0.571	0.571	0.673	0.714	-	-
	GPT-2	0.265	0.816	0.224	0.184	0.204	0.612	-	-
	GPT-3	**0.837**	**1**	**0.949**	0.898	0.878	0.898	-	-
Liver	XGBoost	0.587	0.587	0.587	0.587	0.587	0.587	0.574	0.574
	TabTransformer	0.76	0.753	0.76	0.747	0.837	0.824	0.76	0.74
	CancerGPT	**0.846**	**0.84**	**0.821**	**0.814**	**0.962**	**0.929**	**0.788**	0.814
	GPT-2	0.731	0.449	0.558	0.66	0.679	0.763	0.731	0.731
	GPT-3	0.615	0.49	0.542	0.583	0.474	0.731	0.737	**0.91**
Soft tissue	XGBoost	0.491	0.491	0.491	0.491	0.454	0.476	0.542	0.552
	TabTransformer	0.399	0.299	0.332	0.459	0.72	0.79	0.781	0.756
	CancerGPT	**0.814**	**0.738**	**0.795**	**0.804**	**0.801**	**0.88**	**0.899**	0.885
	GPT-2	0.546	0.535	0.519	0.56	0.427	0.577	0.456	0.384
	GPT-3	0.517	0.406	0.6	0.444	0.607	0.82	0.866	**0.889**
Stomach	XGBoost	0.529	0.529	0.529	0.529	0.529	0.529	0.476	0.508
	TabTransformer	0.731	**0.865**	**0.851**	0.796	0.724	0.75	0.785	0.781
	CancerGPT	**0.802**	0.805	0.82	**0.819**	**0.829**	**0.816**	**0.822**	**0.845**
	GPT-2	0.551	0.569	0.521	0.516	0.589	0.538	0.469	0.566
	GPT-3	0.419	0.575	0.724	0.769	0.534	0.69	0.742	0.724
Urinary tract	XGBoost	0.494	0.494	0.494	0.494	0.494	0.526	0.53	0.544
	TabTransformer	0.493	0.483	0.482	0.499	0.48	0.498	0.501	0.492
	CancerGPT	**0.601**	**0.591**	**0.6**	**0.601**	**0.615**	**0.611**	**0.633**	**0.639**
	GPT-2	0.526	0.528	0.532	0.397	0.515	0.452	0.469	0.566
	GPT-3	**0.645**	0.57	0.556	0.496	0.508	0.516	0.531	0.572
Bone	XGBoost	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499
	TabTransformer	0.547	0.587	**0.69**	**0.706**	**0.696**	**0.729**	**0.728**	0.746
	CancerGPT	**0.584**	**0.632**	0.659	0.631	0.619	0.686	0.602	0.667
	GPT-2	0.507	0.616	0.471	0.579	0.421	0.552	0.476	0.518
	GPT-3	0.498	0.415	0.341	0.429	0.485	0.605	0.62	**0.794**

Table 1: AUROC of k-shot learning on seven tissues sets.





Cancer cell line gene expression
(Nature Communication 2021)



CancerGPT works best

heterogeneous to other common cell lines

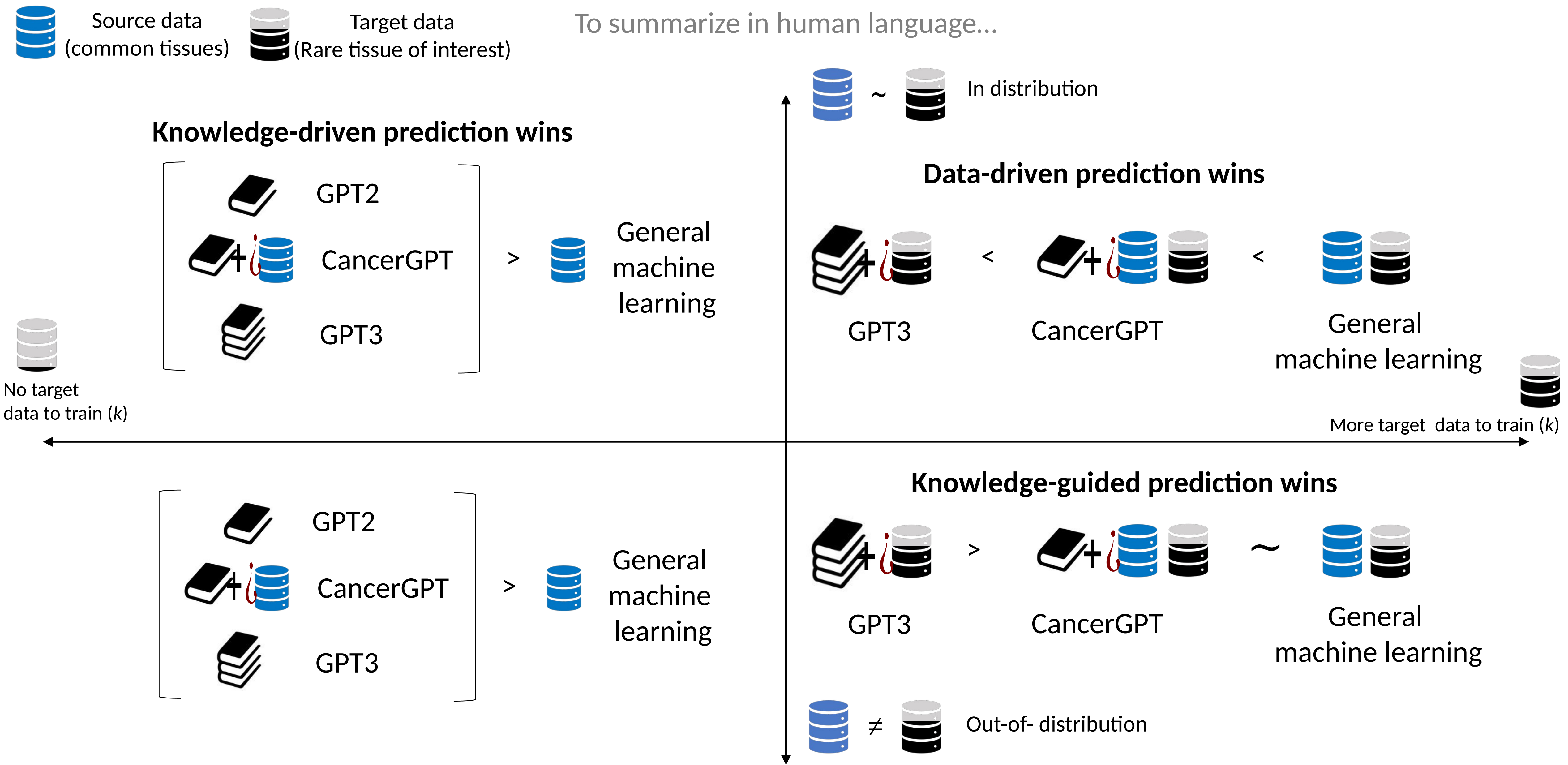
TabTransformer works best

Relatively homogeneous to other common cell lines

https://doi.org/10.1038/s41467-020-20294-x

So, what does it mean?

To summarize in human language…



Any hallucination, Any memorization?

Fact Check GPT3's reasoning

"Why is the drug pair AZD-4877 and AZD1208 synergistic in the cell line T24?"

"The combination of AZD-4877 and AZD1208 has been studied in T24 cells...to be synergistic in reducing bladder cancer cell growth and metastasis"

"The combination was also found to target multiple pathways involved in the growth and spread of bladder cancer cells."

"...Specifically, AZD-4877 was found to inhibit the activation of proteins involved in the promotion of tumor growth..."

"...AZD1208 was found to inhibit proteins associated with the inhibition of tumor growth."

"This combination was also effective at reducing the production of inflammatory mediators such as cytokines, which are known to con- tribute to tumor progression."

GPT3

"...these two drugs have been shown to reduce levels of apoptosis inhibitors, which can also play a role in tumor progression."

To conclude

Summary and contribution

General insight to machine learning

- When data size is limited, LLMs can be a good few-shot prediction tool based on knowledge
- But LLM’s reasoning sometimes contains hallucination

Biomedical contribution

- First prediction model for drug pair synergy in rare cancer types
- Can be extended to a wide range of application beyond cancer, particularly when data is limited

Other LLMs

Pathology with LLM

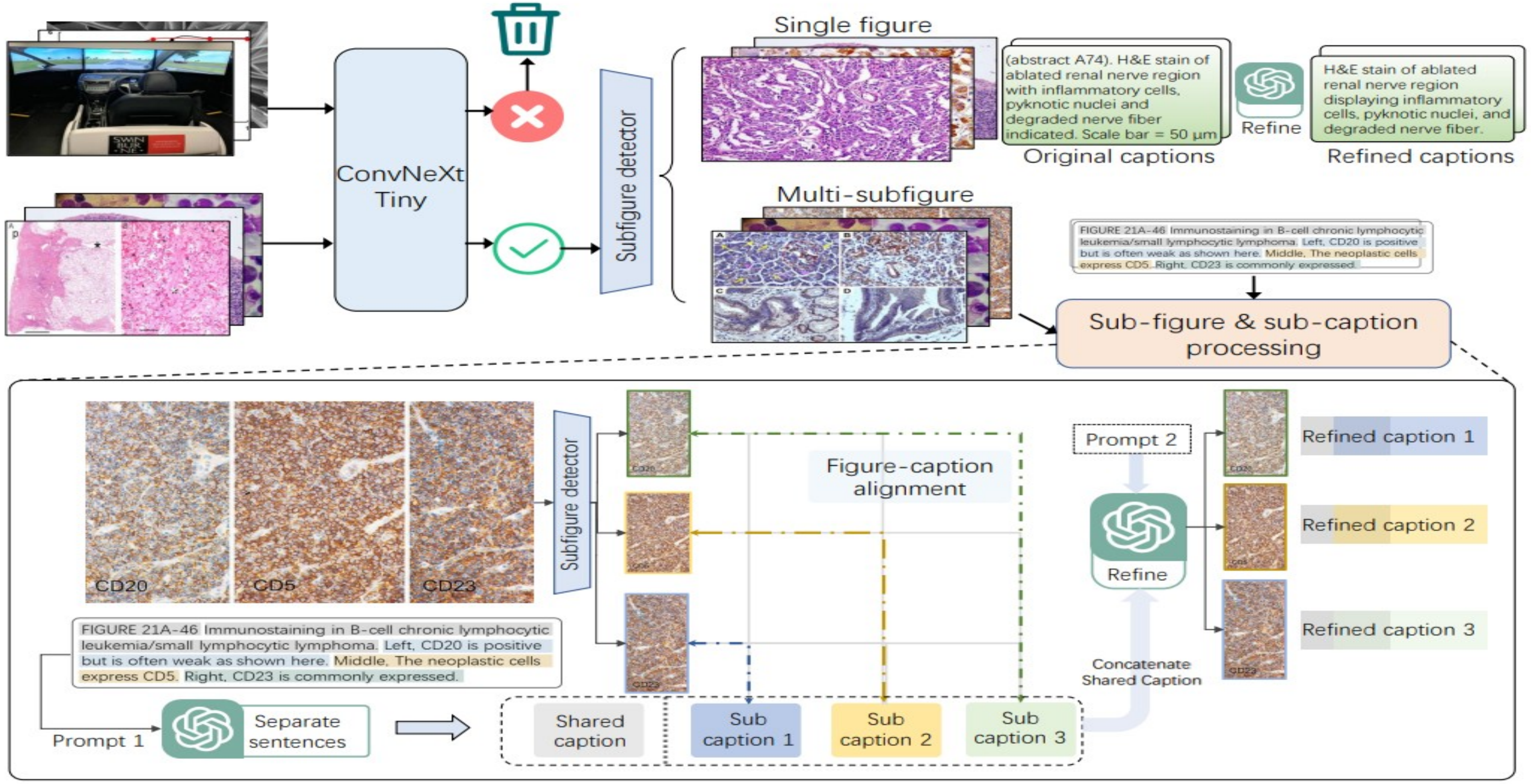


Figure 1: Illustration of processing the image with multiple sub-figures and its corresponding caption.

https://arxiv.org/pdf/2305.15072.pdf

Tree of Thoughts

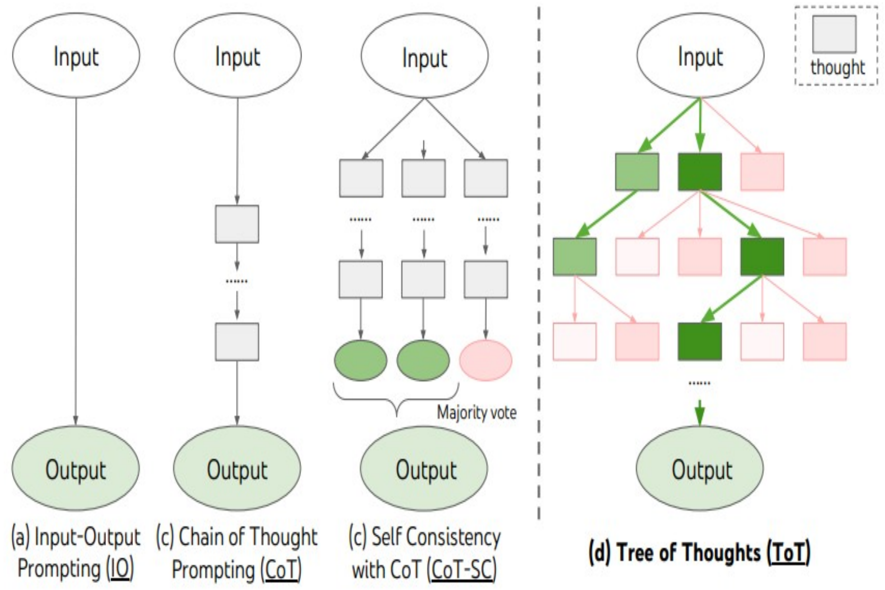


Figure 1: Schematic illustrating various approaches to problem solving with LLMs. Each rectangle box represents a *thought*, which is a coherent language sequence that serves as an intermediate step toward problem solving. See concrete examples of how thoughts are generated, evaluated, and searched in Figures 2,4,6.

https://arxiv.org/pdf/2305.10601.pdf

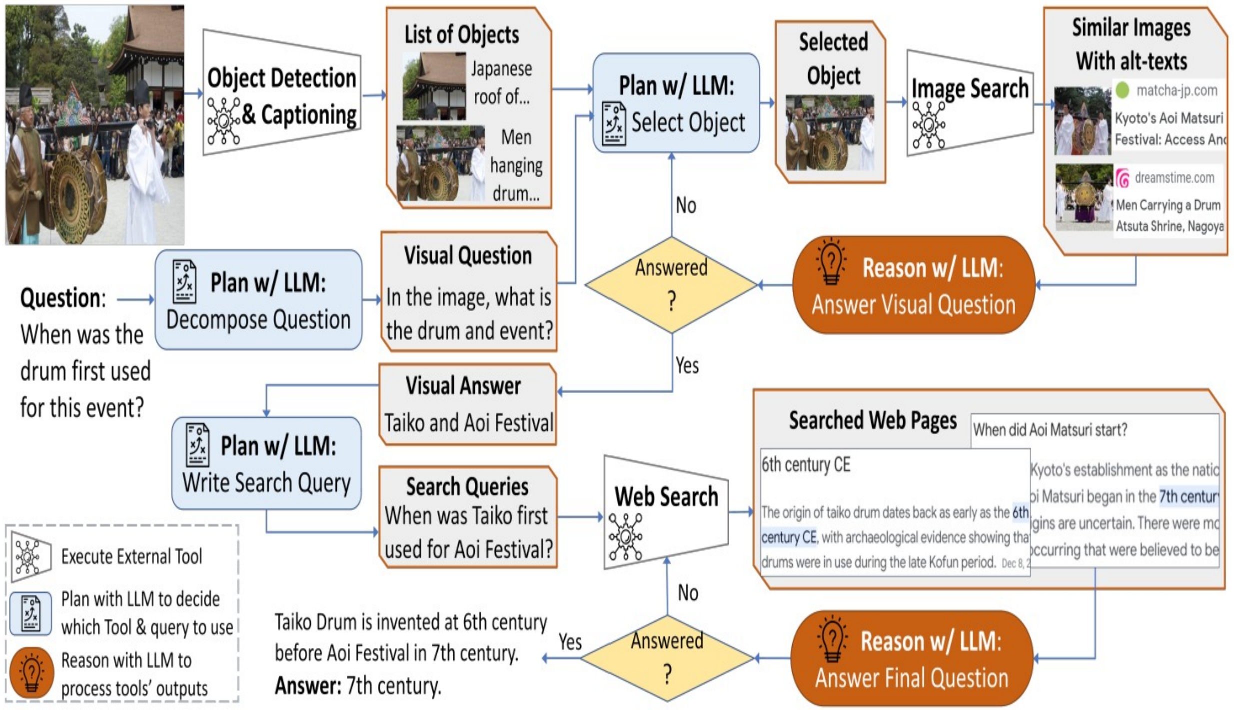


Figure 1: An example of AVIS's generated workflow for answering a challenging visual question. AVIS consists of an LLM-powered Planner that dynamically selects which tool to use and what query to send, executes the tool, and finally applies an LLM-powered Reasoner processing tool's outputs.

https://arxiv.org/pdf/2306.08129.pdf

CancerGPT Tutorial

An example in healthcare: Drug synergy prediction

- Drug synergy refers to the phenomenon where the combined effect of two drugs is greater than their individual effects.

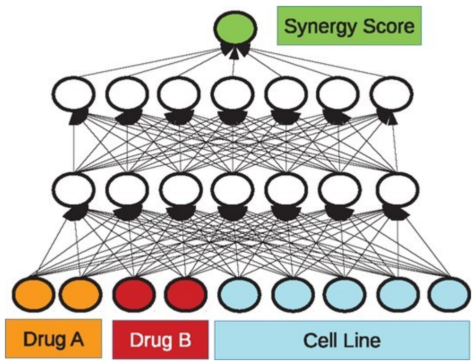
- Researchers usually use experimental methods to find potential drug combination for cancer treatment. However, it is expensive and time-consuming.

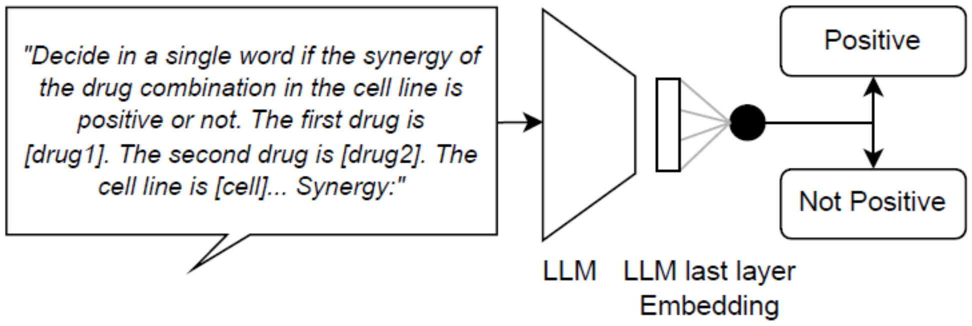
https://www.nature.com/articles/nchembio.1205

Deep learning in drug synergy prediction

- Some previous works use machine learning models to predict synergy. However, with limited data and features, the performance of rare cancers is poor.

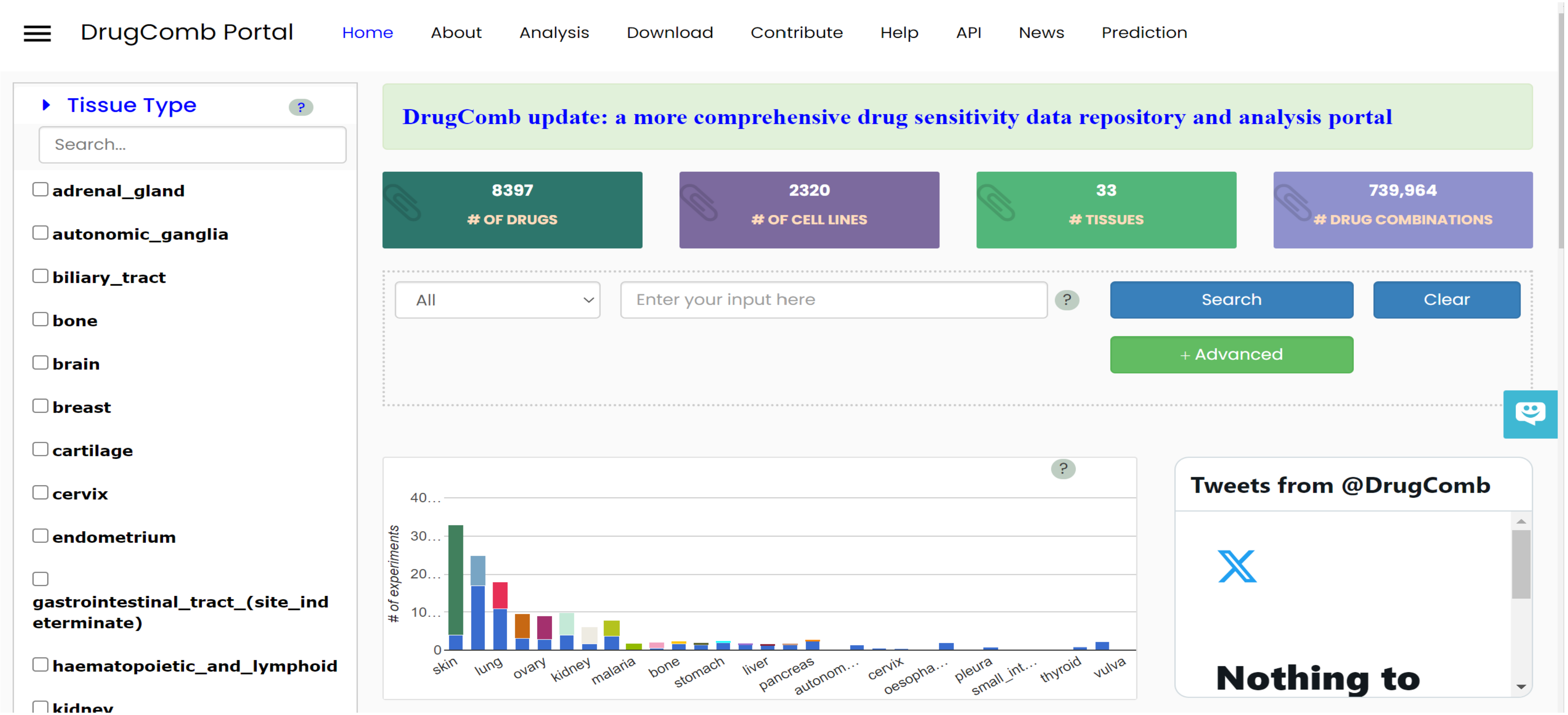
- Leveraging prior knowledge in scientific literature encoded in LLMs is a new way to predict drug synergy for cancers.





https://academic.oup.com/bioinformatics/article/34/9/1538/4747884

Dataset



https://drugcomb.fimm.fi/

Drug synergy prediction using LLM

- Task: Given two drugs and a cell line, determine whether the drugs have synergistic effect on the cell line, based on each drug's sensitivity to the cell line

- Data: DrugComb dataset

	drug_row	drug_col	cell_line_name	tissue_name	ri_row	ri_col	synergy_loewe
0	lonidamine	717906-29-1	A-673	bone	0.568	28.871	-11.702283
1	Ethyl bromopyruvate	717906-29-1	A-673	bone	4.282	26.716	-16.185120
2	Tranilast (trans-)	717906-29-1	A-673	bone	3.056	24.391	-16.588246
3	Lenalidomide	717906-29-1	A-673	bone	-4.751	23.131	-10.877569
4	Pomalidomide	717906-29-1	A-673	bone	2.972	19.578	-1.901326
...	...	...	...	...	...	...	...

https://drugcomb.fimm.fi/

ChatGPT API

- OpenAI's GPT models have been trained to understand natural language and code. GPTs provide text outputs in response to their inputs.

- To use a GPT model via the OpenAI API, you'll send a request containing the inputs and your API key, and receive a response containing the model's output.

```python
response = openai.ChatCompletion.create(
    model="gpt-3.5-turbo",
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The Los Angeles Dodgers won the World Series in 20
        {"role": "user", "content": "Where was it played?"}
    ]
)
```

https://platform.openai.com/docs/guides/gpt

Advantages using ChatGPT API

- Integration into Applications
  - The primary advantage of the API is that developers can integrate ChatGPT into their own applications, services, or platforms. This could be anything from a mobile app, a web service, a chatbot, to a research tool.
- Automation & Workflow
  - Through the API, businesses can automate certain tasks or services. For example, a company might use the API to provide automated customer support on their website.
- Customization
  - With the API, developers can create custom workflows, integrate the system with their databases, and generally tailor the experience to their needs.
  - E.g. Get embeddings from ChatGPT model
- Scalability
  - When integrated into a product, the API allows for scaling up the usage based on demand. If a developer wants to have multiple simultaneous conversations, they can do so using the API.

Zero shot learning by ChatGPT: Prompt Engineering

- Send a request to ChatGPT API using prompt

Prompt: Decide in a single word if the synergy of the drug combination in the cell line is positive (synergy >=5) or negative (synergy <5). Drug combination and cell line: The first drug is AZD4877. The second drug is AZD1208. The cell line is T24. Tissue is bone. The first drug's sensitivity using relative inhibition is 99.091. The second drug's sensitivity using relative inhibition is 3.803. Is this drug combination synergy positive or negative?

```
[48] response = openai.ChatCompletion.create(
       model="gpt-4",
       messages=[
             {"role": "system", "content": "You are an expert on drug discovery."},
             {"role": "user", "content": "Decide in a single word if the synergy of the drug combination in the cell line is positive (synergy >=5) or nega
         ]
     )

     print(response.choices[0].message['content'])
```

```
Positive
```

In-context learning: ChatGPT Prompt

- Adding a role to boost up ChatGPT performance
- Request explanation

```
[49] messages = [
        {"role": "system", "content": "You are an expert on drug discovery."},
        {"role": "user", "content": "Decide in a single word if the synergy of the drug combination in the cell line is positive
        {"role": "assistant", "content": response.choices[0].message['content']},
        {"role": "user", "content": "Can you provide details why the two drugs are synergistic in the cell line?"},
     ]

     response = openai.ChatCompletion.create(
       model="gpt-4",
       messages=messages
     )

     print(response.choices[0].message['content'])
```

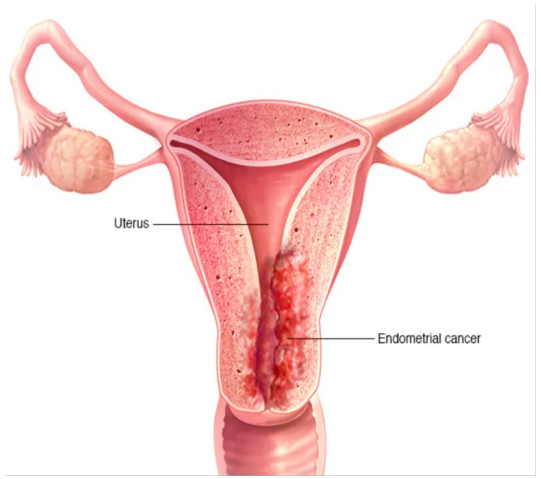
```
Yes. In drug synergy, the combined effectiveness of two drugs is determined not merely by their individual efficacy but also b

However, in drug synergy, even a low efficacy drug can contribute to significant improvements when combined with a high effica
```

https://promptengineering.org/role-playing-in-large-language-models-like-chatgpt/

Zero-shot ChatGPT synergy prediction on endometrium

```
[61] #manually change the positive result to 1 and negative result to 0. In this case [1,1,0,1,0,0,0,0,0,1,1,1,1,1]
     test_labels = list(df.iloc[test_indices]['synergy_class'])
     test_pred = [1,1,0,1,0,0,0,0,0,1,1,1,1,1]
     auroc = roc_auc_score(test_labels, test_pred)
     auprc = average_precision_score(test_labels, test_pred)
     print('\nAUROC:', auroc, '\nAUPRC', auprc)
```

```
AUROC: 0.6428571428571428
AUPRC 0.5892857142857143
```



Get embeddings from ChatGPT

- Embeddings are internal representations that the ChatGPT model learned during its training process, which are high-dimensional vectors
- The length of embedding for GPT-3.5 model “text-embedding-ada-002” is 1536
- Logistic Regression based on GPT-3.5 embeddings can significantly improve the performance

```
▾ LogisticRegression
LogisticRegression(max_iter=1000)
```

Test the performance

```
[67] test_embeddings = generate_embeddings(test_ds)
     test_labels = list(df.iloc[test_indices]['synergy_class'])

     test_pred = model.predict_proba(test_embeddings)[:,1]
     auroc = roc_auc_score(test_labels, test_pred)
     auprc = average_precision_score(test_labels, test_pred)
     print('\nAUROC:', auroc, '\nAUPRC', auprc)
```

```
100%|██████████| 14/14 [00:02<00:00,  6.42it/s]
AUROC: 0.8979591836734695
AUPRC 0.8736394557823128
```

https://platform.openai.com/docs/guides/embeddings

Get embeddings from CancerGPT

- ChatGPT is not an open-source model and cannot be used for large-scale finetuning
- However, finetuning on common cancer data boosts performance on rare cancer data
- We finetuned a GPT-2 model (CancerGPT) before and compared it with ChatGPT.

```
    train_dataset = test_ds,
    eval_dataset = test_ds,
    compute_metrics = compute_metrics,
    tokenizer = tokenizer,
    data_collator=data_collator,
)
eva = trainer.evaluate()
auroc = eva['eval_AUROC']
auprc = eva['eval_AUPRC']
print('\nAUROC:', auroc, '\nAUPRC', auprc)
```

```
[1/1 : < :]

AUROC: 0.7959183673469388
AUPRC 0.8468614718614718
```

https://arxiv.org/abs/2304.10946

Comparison of different models

- In Endometrium cancer, CancerGPT is better than basic ChatGPT, but not as good as ChatGPT embedding + classification.
- However, CancerGPT (~124M) is much smaller than ChatGPT (~175B)

https://arxiv.org/abs/2304.10946