Comparative Performance Evaluation of Large Language Models for Extracting Molecular Interactions and Pathway Knowledge

Gilchan Park, Xihaier Luo, Vanessa Lopez-Marrero, Shinjae Yoo, Byung-Jun Yoon, Shantenu Jha

Computational Science Initiative, Brookhaven National Laboratory

10/25/2023

Introduction

- Fine-tuning model for every task is costly, labor-intensive, and inefficient.
- We want more generalized *all-in-one* models.
- Large language models (LLMs) built on enormous amounts of text data with hundreds of billions of parameters.
- Evaluate LLMs for the following biological tasks:
  - ✓ Recognizing protein-protein interactions (PPIs)
  - ✓ Identifying genes related to human pathways effected by low-dose radiation
  - ✓ Finding gene regulatory relations

LLMs for the evaluation

Model	Release Date	Developer	Parameters	Context length	Features
BioGPT-Large	Feb 2023	Microsoft	1.5B	1024	• Domain-specific foundation model • Trained on biomedical literature for biological tasks
BioMedLM	Jan 2023	Stanford	2.7B	1024	• Domain-specific foundation model • Trained on biomedical literature for medical question answering
Galactica	Nov 2022	Meta	120M, 1.3B, 6.7B, 30B, 120B	2048	• Trained on scientific literature • Designed data for scientific tasks
Alpaca	March 2023	Stanford	7B	2048	• Instruction fine-tuned version of the LLaMA 7B model on 52K instruction-following demonstrations
RST	Sep 2022	CMU	11B	input: 1024 output: 256	• ReStructured Pre-training (RST) • Transformer encoder-decoder framework • Designed data for various NLP tasks
Falcon	March 2023	TII	7B, 40B	2048	• Trained on high-quality data filtered by Falcon RefinedWeb
MPT-Chat	July 2023	MosaicML	7B, 30B	2048 8192	• Chatbot-like MPT model for dialogue generation
Llama-2-Chat	July 2023	Meta	7B, 13B, 70B	4096	• 40% more data than Llama 1 and has double the context length • Fine-tuned version of Llama 2 that is optimized for dialogue use cases

Recognizing protein-protein interactions (PPIs)

LLMs evaluation on human protein interactions

PPI Task 1

- Task - List proteins binding to a protein (Generative question)
- Data: **STRING DB** human (Homo Sapiens) protein network
- Compared generated a list of proteins with ground truth using 10,000 PPI pairs from 1,000 STRING DB human protein list.
- Prompt:

```
Question: Which proteins interact with C12orf74?
Answer: AGXT, CELSR1, FASTK, GRK5, LCORL, PLEKHG7, RIMS1, RIMS2, SFTA3, ZNF280B

Question: Which proteins interact with OR2T7?
Answer: ACTL9, ADRBK1, ADRBK2, ANO2, ARRB1, ARRB2, CNGA2, FTCD, GNAL, GNB1
```

PPI Task 1 – Evaluation Metrics

- Micro F1: measure the matches in all 10K pairs.
- Macro F1: measure the matches for each label (protein used in a query like USP32 below)

```
Question: Which proteins interact with USP32?
Answer (true): USP54, USP41, USP42, USP34, USP38, USP50, CACNA1H, ACTC1, DHX32, MAGI3
Answer (pred): USP54, USP41, USP42, USP34, USP38, USP50, USP52, USP32, USP55, USP56
```

- The number of full matched proteins out of 1K like EED below.

```
Question: Which proteins interact with EED?
Answer (true): HDAC1, SMARCA4, HMGB2, CBX5, HDAC2, EZH2, CBX3, GATA2, STAG2, RB1
Answer (pred): HDAC1, SMARCA4, HMGB2, CBX5, HDAC2, EZH2, CBX3, GATA2, STAG2, RB1
```

PPI Task 1 – Results

- LLaMA-2-Chat (70B) generated the most correct protein interactions followed by Galactica (30B), and Falcon (7B) performed the worst.
- The larger models potentially possess a greater reservoir of in-depth information pertaining to specific proteins as seen in full match count.
- The models predicted better for proteins with similar names, such as IKZF4 and RFC5.

* 5-shot prompting used

Model	Micro F1	Macro F1	# Full Match out of 1K
BioGPT-Large (1.5B)	0.1220	0.1699	10
BioMedLM (2.7B)	0.1584	0.1992	61
Galactica (6.7B)	0.2110	0.2648	75
Galactica (30B)	0.2867	0.3516	110
Alpaca (7B)	0.1573	0.2211	23
RST (11B)	0.0987	0.1523	10
Falcon (7B)	0.0435	0.0632	7
Falcon (40B)	0.1124	0.1492	31
MPT-Chat (7B)	0.1307	0.1688	40
MPT-Chat (30B)	0.2926	0.3467	144
LLaMA-2-Chat (7B)	0.2768	0.3436	99
LLaMA-2-Chat (70B)	**0.3517**	**0.4187**	**159**

PPI Task 2

- Task - Determine whether two proteins bind to each other (yes/no question)
- Data:
  - ✓ Positive PPI: **STRING DB** human (Homo Sapiens) protein network
  - ✓ Negative PPI: **Negatome 2.0** human (Homo Sapiens) protein interactions
- Evaluation on 2000 samples (1000 positive + 1000 negative).
- Prompt:

Question: Do TMEM43 and POTEI interact with each other?
Answer: yes

Question: Do Q5JTD0 and A5JSJ9 interact with each other?
Answer: no

PPI Task 2 – Results

- MPT-Chat (7B) exhibited the highest score followed by LLaMA-2-Chat (70B) and MPT-Chat (70B).
- BioGPT-Large and Falcon (7B) manifested almost zero capability in responding to questions, and Falcon (40B) and RST also exhibited higher rates of false negatives (i.e., almost all ‘no’) while Galactica (6.7B) model showed a higher rate of false positives.

Model	Micro F1
BioGPT-Large (1.5B)	0.5005 (1-shot)
BioMedLM (2.7B)	0.8500 (2-shot)
Galactica (6.7B)	0.5320 (1-shot)
Galactica (30B)	0.8585 (5-shot)
Alpaca (7B)	0.8615 (0-shot)
RST (11B)	0.6990 (0-shot)
Falcon (7B)	0.5000 (1-shot)
Falcon (40B)	0.6570 (1-shot)
MPT-Chat (7B)	**0.9840** (5-shot)
MPT-Chat (30B)	0.9350 (5-shot)
LLaMA-2-Chat (7B)	0.8695 (5-shot)
LLaMA-2-Chat (70B)	0.9545 (5-shot)

Identifying genes related to human pathways

LLMs evaluation on human pathways effected by low-dose radiation exposure

Genes in LD related Pathways

- Task - List genes involved in a human pathway (Generative question)
- Data: **KEGG DB** human pathways affected by low-dose (LD) radiation exposure.
- Evaluated on 998 genes from the top 100 pathways effected to low-dose radiation exposure [1].
- Prompt:

```
Question: Which genes are associated with “Caffeine metabolism”?
Answer: NAT2, CYP1A2, CYP2A6, CYP2A7, XDH, NAT1, RAET1E, ...

Question: Which genes are associated with "Basal cell carcinoma"?
Answer: CDKN1A, APC2, GADD45G, FZD10, CTNNB1, DDB2, ...
```

[1] Luo, Xihaier, et al. "Pathway-based analyses of gene expression profiles at low doses of ionizing radiation” *2022 the Journal of Frontiers in Bioinformatics (under review)*

Genes in LD related Pathways – Results

- Galactica (30B) showed the best performance followed by MPT-Chat (30B) and LLaMA2-Chat (70B), and Falcon (7B) performed the worst.
- The overall performance of the models surpassed PPI Task1 results, which might be explained by low-dose radiation are often mentioned in narrower and specific sections or categories within the literature. (the significant improvement of BioMedLM and BioGPT-Large)

Model	Micro F1	Macro F1	# Full Match out of 100
BioGPT-Large (1.5B)	0.2435 (3-shot)	0.3131 (3-shot)	5
BioMedLM (2.7B)	0.4279 (2-shot)	0.5040 (2-shot)	19
Galactica (6.7B)	0.3136 (5-shot)	0.3874 (5-shot)	8
Galactica (30B)	**0.4609** (5-shot)	**0.5304** (5-shot)	**24**
Alpaca (7B)	0.1293 (5-shot)	0.1715 (5-shot)	2
RST (11B)	0.0741 (5-shot)	0.0837 (5-shot)	4
Falcon (7B)	0.0491 (5-shot)	0.0685 (5-shot)	2
Falcon (40B)	0.1844 (5-shot)	0.2367 (5-shot)	5
MPT-Chat (7B)	0.1824 (5-shot)	0.2368 (5-shot)	4
MPT-Chat (70B)	0.3978 (5-shot)	0.4550 (5-shot)	18
LLaMA-2-Chat (7B)	0.2535 (5-shot)	0.3106 (5-shot)	8
LLaMA-2-Chat (70B)	0.3908 (5-shot)	0.4577 (5-shot)	18

Finding gene regulatory relations

LLMs evaluation on human gene regulatory relations

Gene Regulatory Relation Task

- Evaluate the LLMs on human gene regulatory relations.
- Data: human gene regulatory relation from **INDRA DB**.
- Unlike the earlier experiments, INDRA has context information.
- There are total 23 relation types and used the top 6 relation types by occurrence.
- For each relation type, 1,000 instances were sampled.
- Task - Choose a relation of two genes given a statement (multiple choice question)

INDRA DB – Classes

# classes	Classes
2 class	Activation, Inhibition
3 class	Activation, Inhibition, Phosphorylation
4 class	Activation, Inhibition, Phosphorylation, Dephosphorylation
5 class	Activation, Inhibition, Phosphorylation, Dephosphorylation, Ubiquitination
6 class	Activation, Inhibition, Phosphorylation, Dephosphorylation, Ubiquitination, Deubiquitination

Gene Regulatory Relation – Prompt

- Prompt with 6 classes:

```
Context: In 2006, we demonstrated that activation of TRPM2 appeared to induce
insulin secretion.

Question: Given the options: "Activation", "Inhibition", "Phosphorylation",
"Dephosphorylation", "Ubiquitination", "Deubiquitination", which one is the
relation type between TRPM2 and insulin in the text above?

Answer: Activation
-------------------------------------------------------------------------
Context: WRN was shown to genetically interact with topoisomerase 3 and restore
the slow growth phenotype of sgs1 top3.

Question: Given the options: "Activation", "Inhibition", "Phosphorylation",
"Dephosphorylation", "Ubiquitination", "Deubiquitination", which one is the
relation type between WRN and top3 in the text above?

Answer: Inhibition
```

Gene Regulatory Relation – Results

- Overall, the larger models outperformed the smaller models such as BioGPT-Large and BioMedLM. This suggests that models trained on larger and more diverse datasets possess a stronger ability to comprehend the meaning of text compared to models trained on narrower and smaller datasets.
- The improved linguistic understanding associated with the size of the training data is further supported by the superior performance of the largest model, LLaMA-2-Chat (70B), MPT-Chat (70B), Galactica (30B).

† zero-shot ‡ 1-shot $ 2-shot

Model	2 class	3 class	4 class	5 class	6 class
BioGPT-Large (1.5B)†	0.4740	0.3897	0.2933	0.3276	0.2878
BioMedLM (2.7B)†	0.5420	0.4083	0.3070	0.2298	0.1950
Galactica (6.7B)$	0.7040	0.6053	0.5670	0.5852	0.5970
Galactica (30B)$	0.7385	**0.7347**	0.5982	0.6672	0.6678
Alpaca (7B)‡	0.7355	0.6447	0.5560	0.6362	0.5347
RST (11B)†	0.6395	0.7177	0.5972	0.6666	0.6137
Falcon (7B)$	0.6200	0.4717	0.3370	0.3550	0.2648
MPT-Chat (7B)$	0.7725	0.7313	0.5180	0.6132	0.5865
MPT-Chat (70B)$	0.7540	0.7070	**0.6590**	**0.6912**	0.6432
LLaMA-2-Chat (7B)$	0.7730	0.6527	0.5643	0.4072	0.4093
LLaMA-2-Chat (70B)$	**0.7995**	0.7260	0.5905	0.6858	**0.6693**

Discussion & Ongoing

- The larger models such as Llama-2-chat (70B), MPT-Chat (30B), and Galactica (30B) performed the best, and they hold promise for selective tasks involving the extraction of biological knowledge.
- Contextual information might improve the model's performance.
  - ✓ Leverage external knowledge sources (e.g., biological databases)
  - ✓ Utilize auxiliary queries to the model to extract additional context to be fed into main questions. (e.g., list papers about "human papillomavirus infection" pathway.)
- Prompts are an important role, which needs to be further elaborated.
  - ✓ Adopting state-of-the-art prompt design/tuning methods
  - ✓ Selecting good examples in a prompt for a model

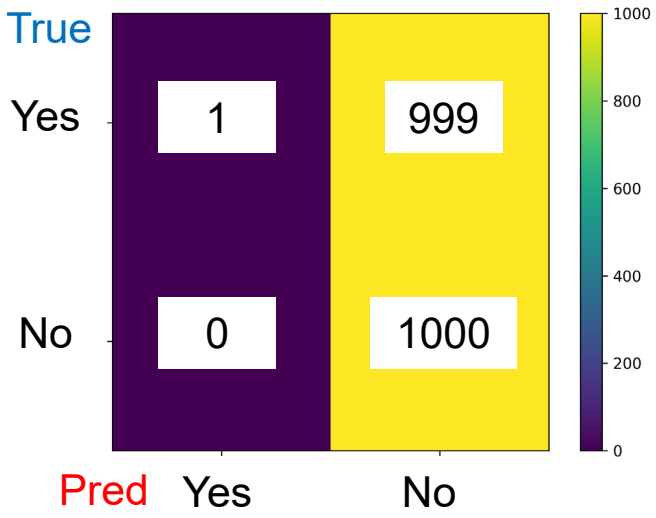
Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

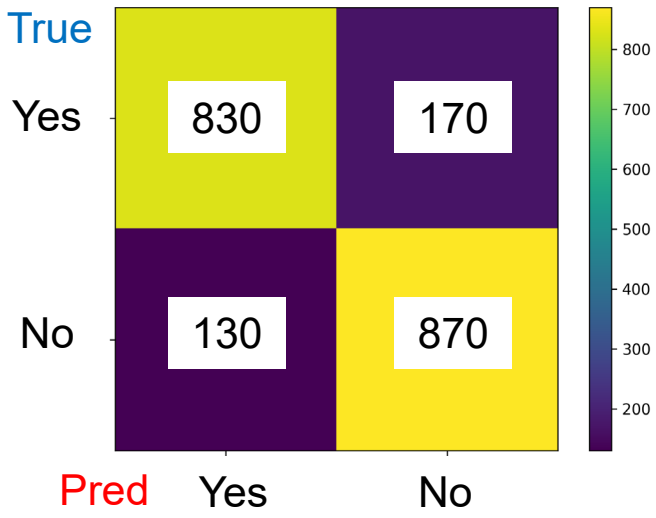
Thanks

@BrookhavenLab

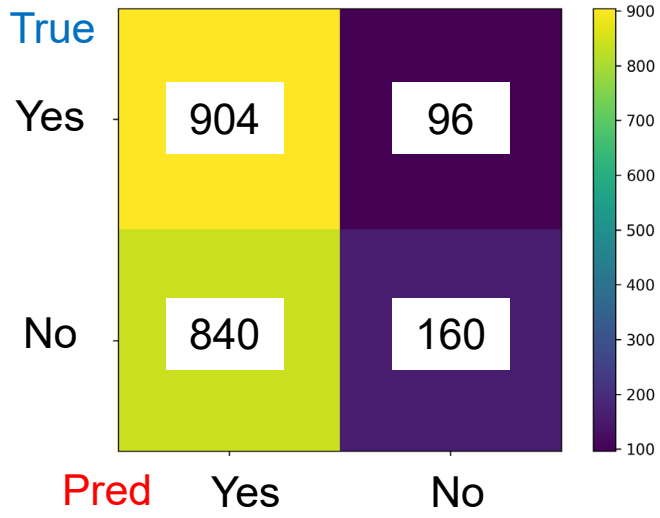
BioGPT-Large (1.5B)



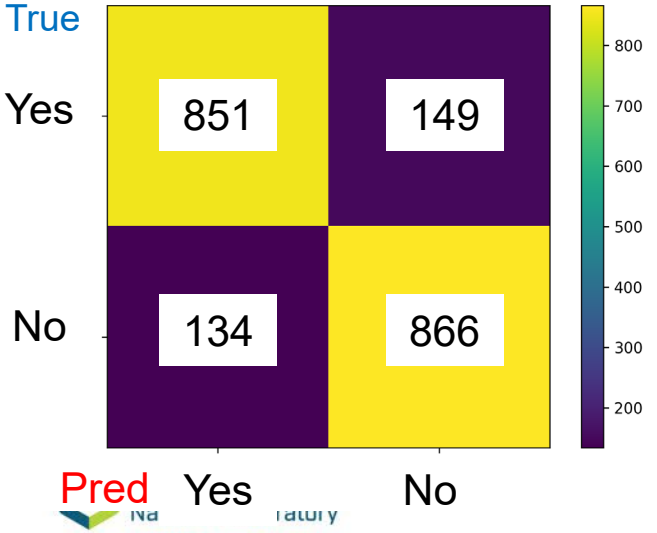
BioMedLM (2.7B)



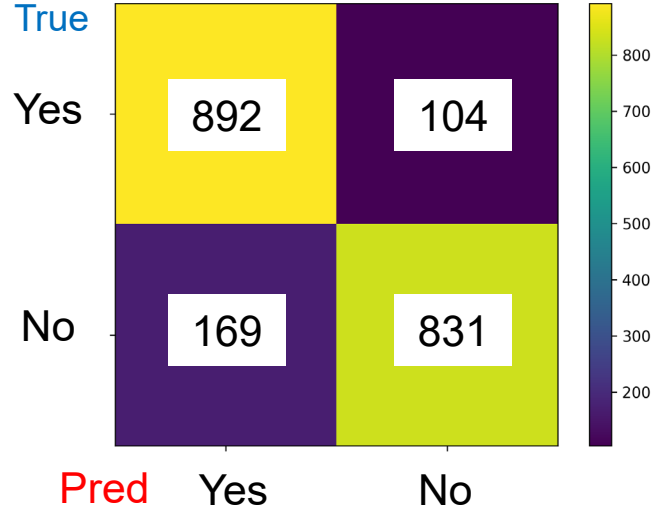
Galactica (6.7B)



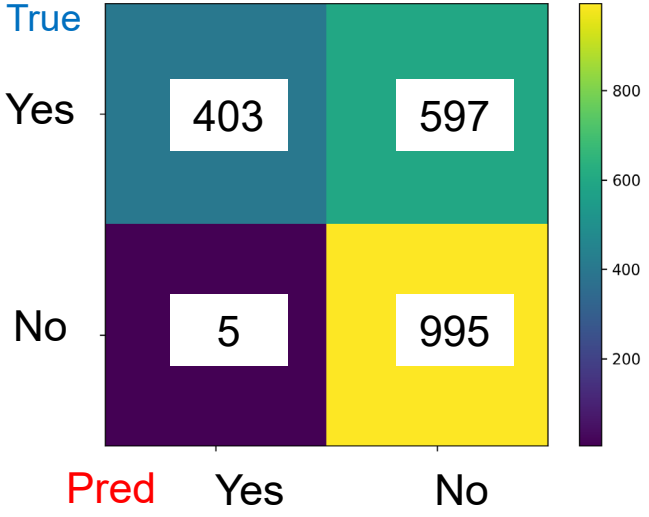
Galactica (30B)



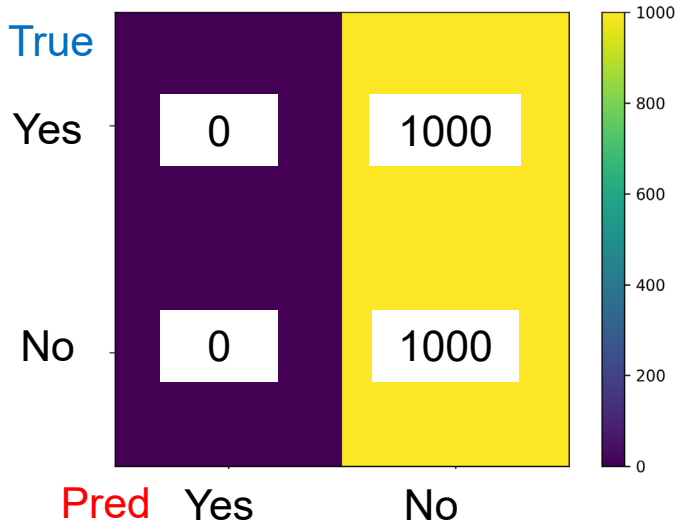
Alpaca (7B)



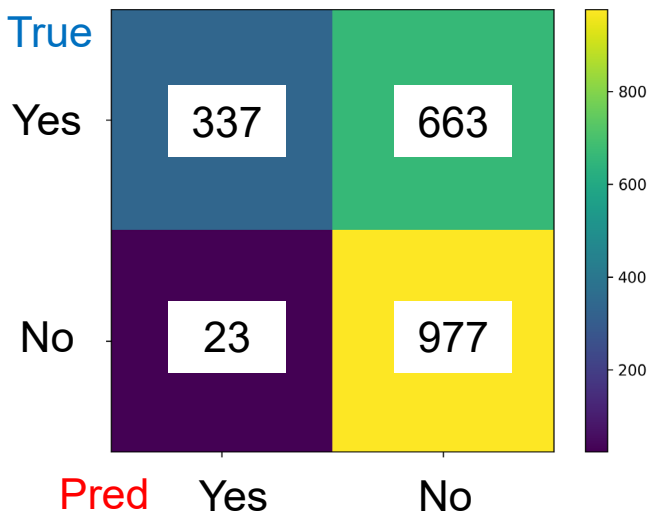
RST (11B)



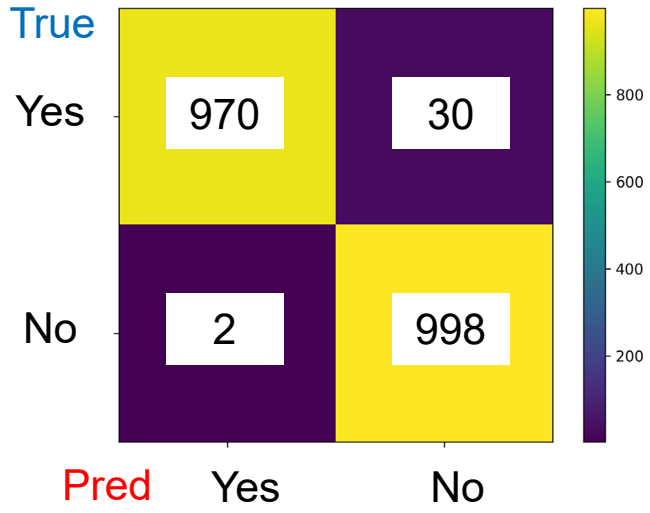
Falcon (7B)



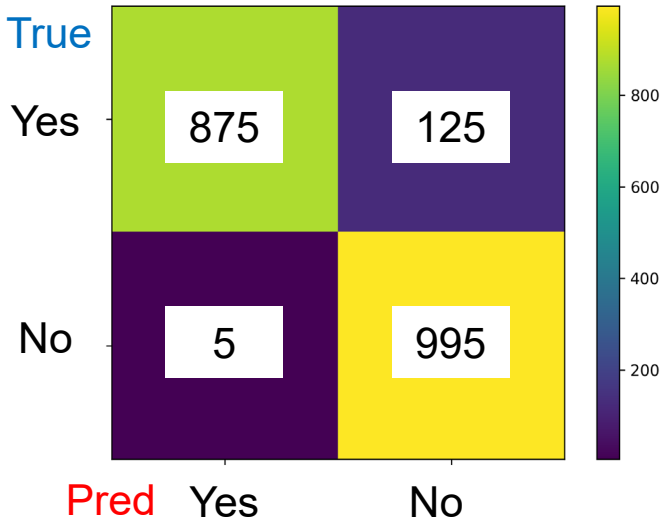
Falcon (40B)



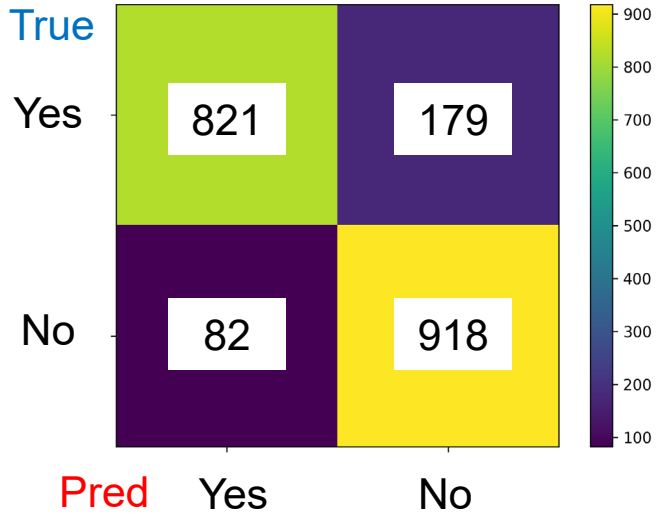
MPT-Chat (7B)



MPT-Chat (30B)



LLaMA2-Chat (7B)



LLaMA2-Chat (70B)