Vaccine Large Language Model (VaxLLM)

Presentor: Laurel (Xingxian) Li

12/09/2024

https://www.biorxiv.org/content/10.1101/2024.11.25.625209v1

Motivation

The Challenge

- Development of database (ex. ontologies) needs manual annotation
- Rapid increase in scientific literature overwhelms manual annotation efforts.
- Annotated information is essential to advance vaccine development

↓

Urgent need to automate the vaccine annotation process from resources

VIOLIN

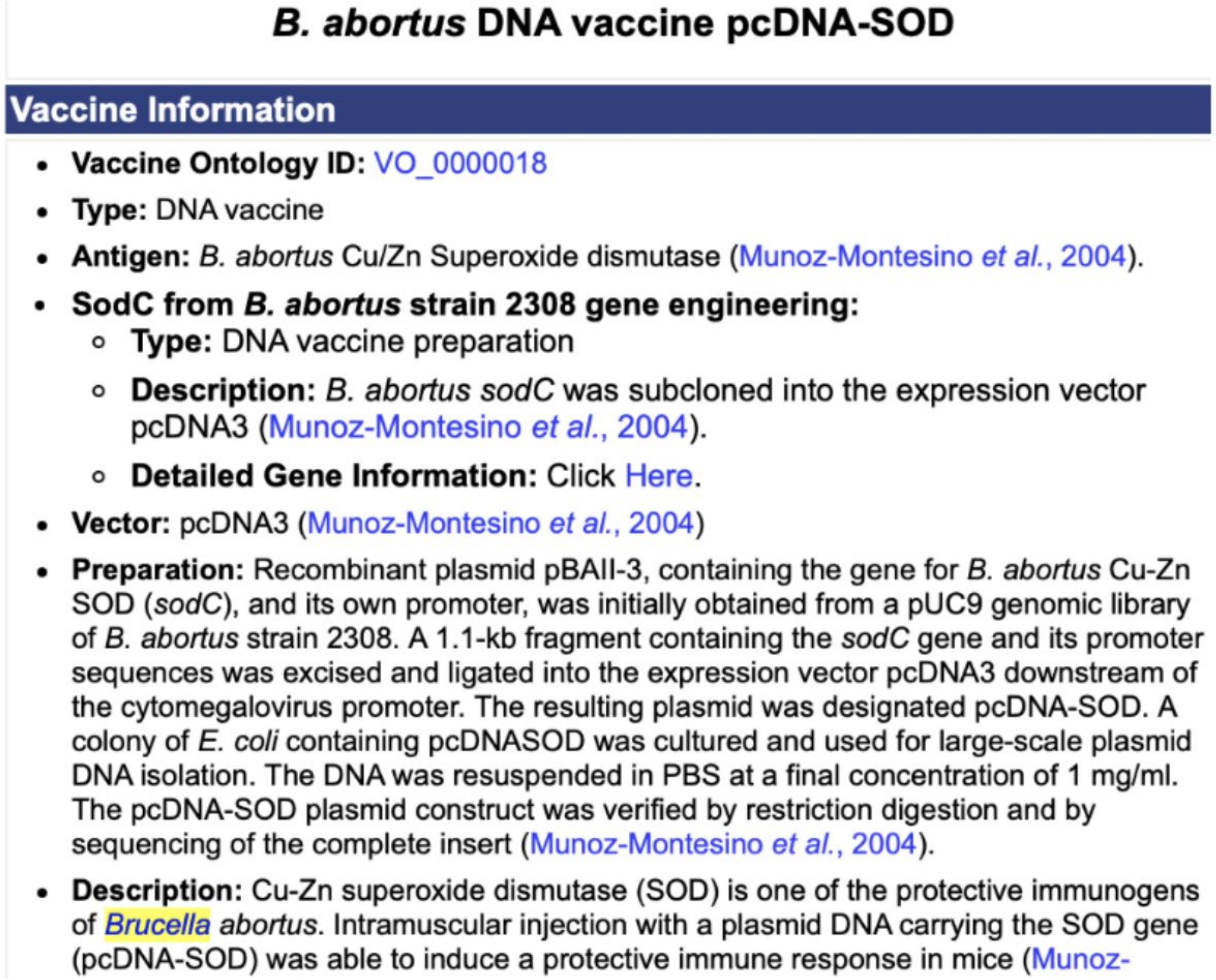
Vaccine Investigation and Online Information Network

VIOLIN database

- vaccine research data curation and storage

- include vaccine candidates developed against various pathogens

https://violinet.org/



B. abortus DNA vaccine pcDNA-SOD

Vaccine Information

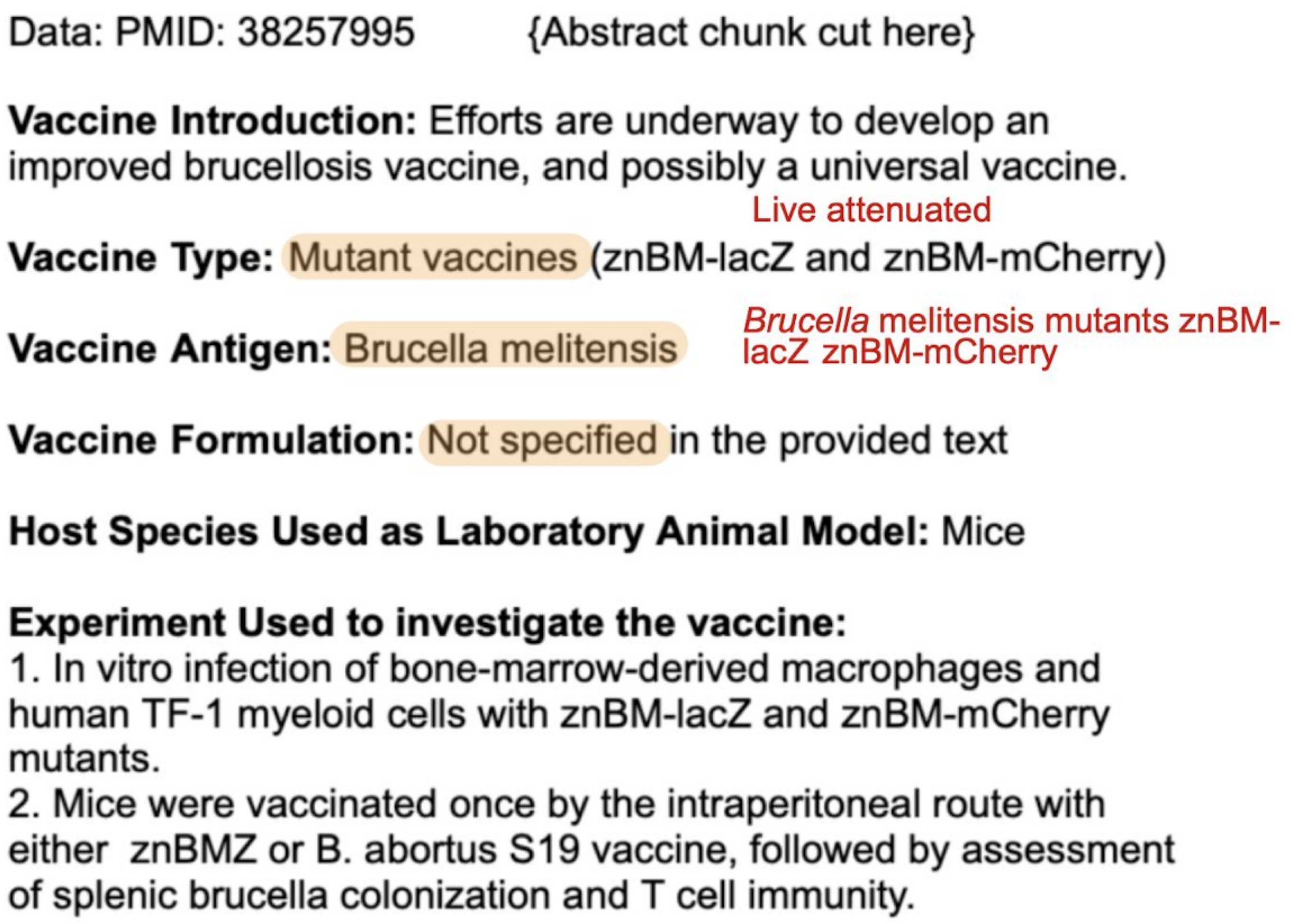
- **Vaccine Ontology ID:** VO_0000018
- **Type:** DNA vaccine
- **Antigen:** *B. abortus* Cu/Zn Superoxide dismutase (Munoz-Montesino *et al.*, 2004).
- **SodC from _B. abortus_ strain 2308 gene engineering:**
  - **Type:** DNA vaccine preparation
  - **Description:** *B. abortus sodC* was subcloned into the expression vector pcDNA3 (Munoz-Montesino *et al.*, 2004).
  - **Detailed Gene Information:** Click Here.
- **Vector:** pcDNA3 (Munoz-Montesino *et al.*, 2004)
- **Preparation:** Recombinant plasmid pBAII-3, containing the gene for *B. abortus* Cu-Zn SOD (*sodC*), and its own promoter, was initially obtained from a pUC9 genomic library of *B. abortus* strain 2308. A 1.1-kb fragment containing the *sodC* gene and its promoter sequences was excised and ligated into the expression vector pcDNA3 downstream of the cytomegalovirus promoter. The resulting plasmid was designated pcDNA-SOD. A colony of *E. coli* containing pcDNASOD was cultured and used for large-scale plasmid DNA isolation. The DNA was resuspended in PBS at a final concentration of 1 mg/ml. The pcDNA-SOD plasmid construct was verified by restriction digestion and by sequencing of the complete insert (Munoz-Montesino *et al.*, 2004).
- **Description:** Cu-Zn superoxide dismutase (SOD) is one of the protective immunogens of *Brucella abortus*. Intramuscular injection with a plasmid DNA carrying the SOD gene (pcDNA-SOD) was able to induce a protective immune response in mice (Munoz-

LLM Annotation Focus:

using PubMed abstract for itemized vaccine annotation

1. Vaccine type (live attenuated, recombinant, etc)
2. Vaccine antigen
3. Vaccine formulation (antigen, protein, gene, adjuvant）
4. Host species use as a specific lab animal model
5. Experiment used to investigate the vaccine

From Baseline Llama 3 to VaxLLM



Data: PMID: 38257995 {Abstract chunk cut here}

Vaccine Introduction: Efforts are underway to develop an improved brucellosis vaccine, and possibly a universal vaccine.

Live attenuated

Vaccine Type: Mutant vaccines (znBM-lacZ and znBM-mCherry)

Brucella melitensis mutants znBM-lacZ znBM-mCherry

Vaccine Antigen: Brucella melitensis

Vaccine Formulation: Not specified in the provided text

Host Species Used as Laboratory Animal Model: Mice

Experiment Used to investigate the vaccine:
1. In vitro infection of bone-marrow-derived macrophages and human TF-1 myeloid cells with znBM-lacZ and znBM-mCherry mutants.
2. Mice were vaccinated once by the intraperitoneal route with either znBMZ or B. abortus S19 vaccine, followed by assessment of splenic brucella colonization and T cell immunity.

Llama 3 8B direct usage for paper PMID 38257995

Incorrect or missing information

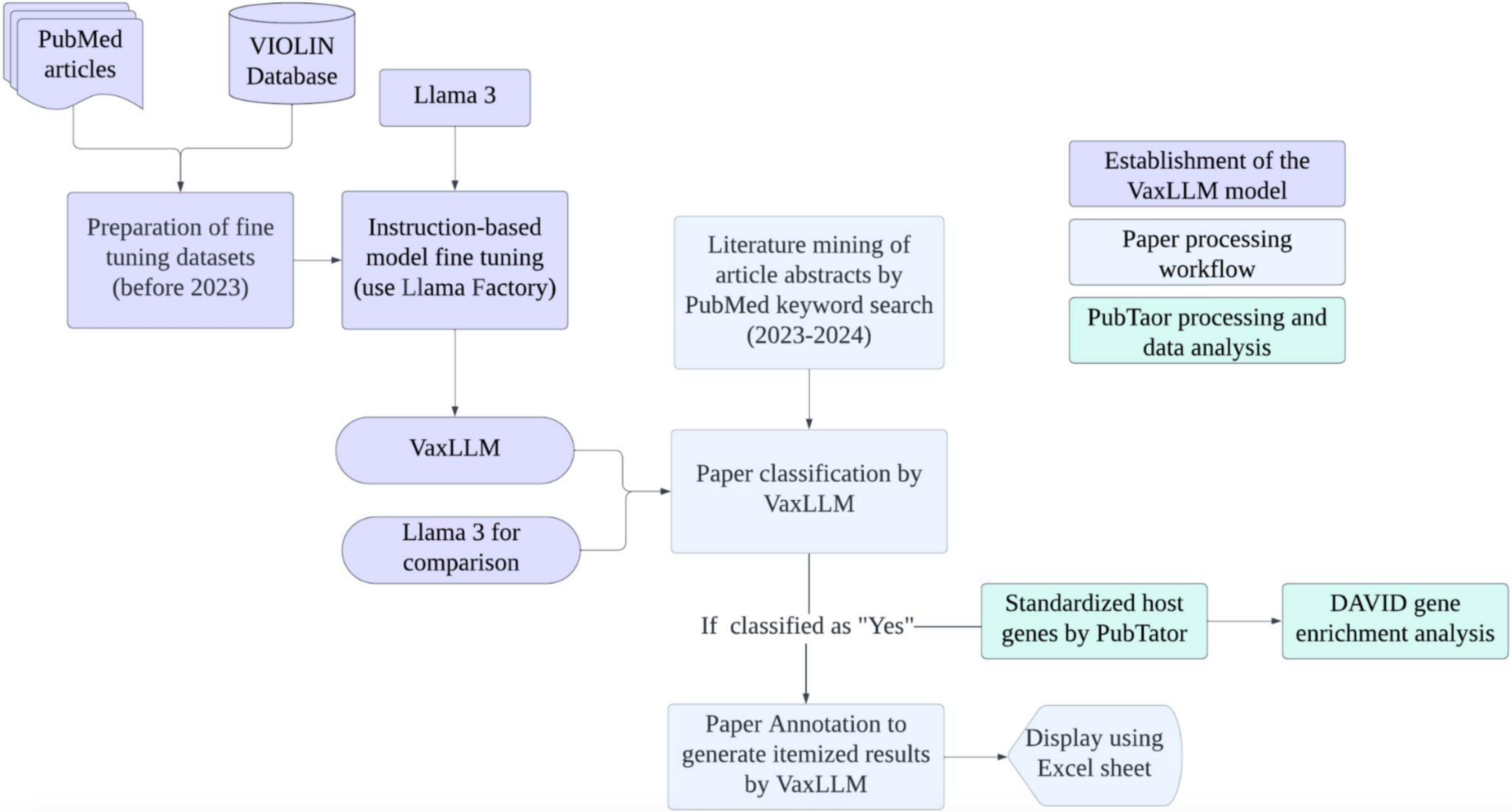
VaxLLM

- A fine-tuned Large Language Model derived from Llama3-8B

The whole project is to develop a pipeline to:

perform **classification** and **standardized annotation** of *Brucella* vaccine articles using VaxLLM

Classification Focus:
filter the relevant articles containing specific vaccine formulation


PubMed articles
VIOLIN Database
Llama 3
Preparation of fine tuning datasets (before 2023)
Instruction-based model fine tuning (use Llama Factory)
Literature mining of article abstracts by PubMed keyword search (2023-2024)
Establishment of the VaxLLM model
Paper processing workflow
PubTaor processing and data analysis
VaxLLM
Llama 3 for comparison
Paper classification by VaxLLM
If classified as "Yes"
Standardized host genes by PubTator
DAVID gene enrichment analysis
Paper Annotation to generate itemized results by VaxLLM
Display using Excel sheet

Development of VaxLLM - Model fine tuning

Let the model know what is correct annotation → VIOLIN !

Classification training data:

- Positive Examples: 50 scientific articles abstracts (From VIOLIN reference)
- Negative Examples: 100 articles abstracts (From PubMed before 2023)

Annotation training data:

- The standardized annotation from VIOLIN

Training data format

Alpaca Format:

- **Instruction**: our instruction (e.g., "Is this article about a *Brucella* vaccine?..." or "Extract the following details...").
- **Input**: The abstract of the article.
- **Output**: The classification result (Yes/No) or detailed vaccine annotation

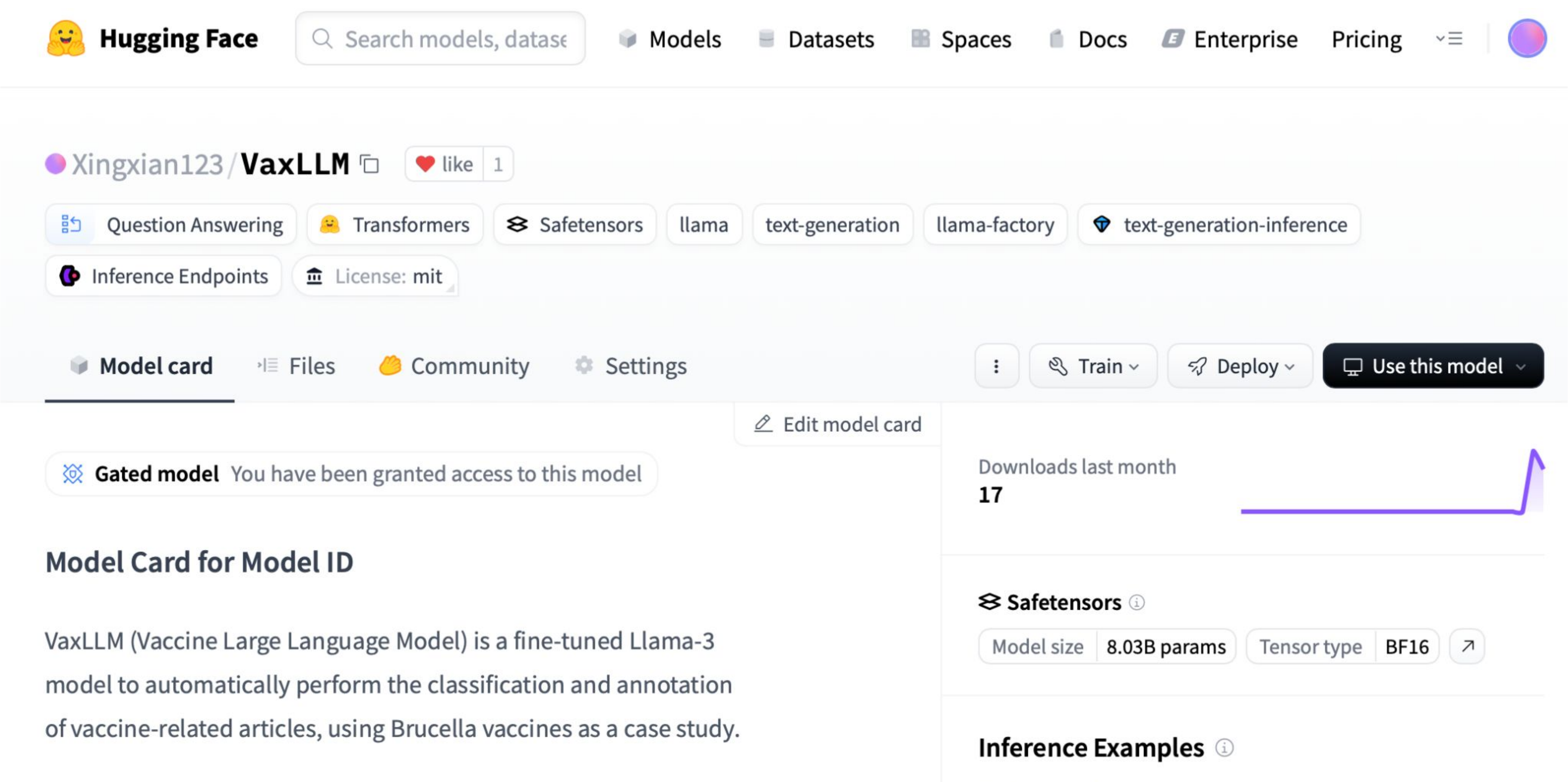
Fine-tuning process: Llama Factory

https://github.com/hiyouga/LLaMA-Factory

A package for fine-tuning of hundreds of LLMs

adopted LoRA, 4 bits for VaxLLM, T4 GPU

https://huggingface.co/Xingxian123/VaxLLM



Gated model You have been granted access to this model

Model Card for Model ID

VaxLLM (Vaccine Large Language Model) is a fine-tuned Llama-3 model to automatically perform the classification and annotation of vaccine-related articles, using Brucella vaccines as a case study.

Downloads last month 17

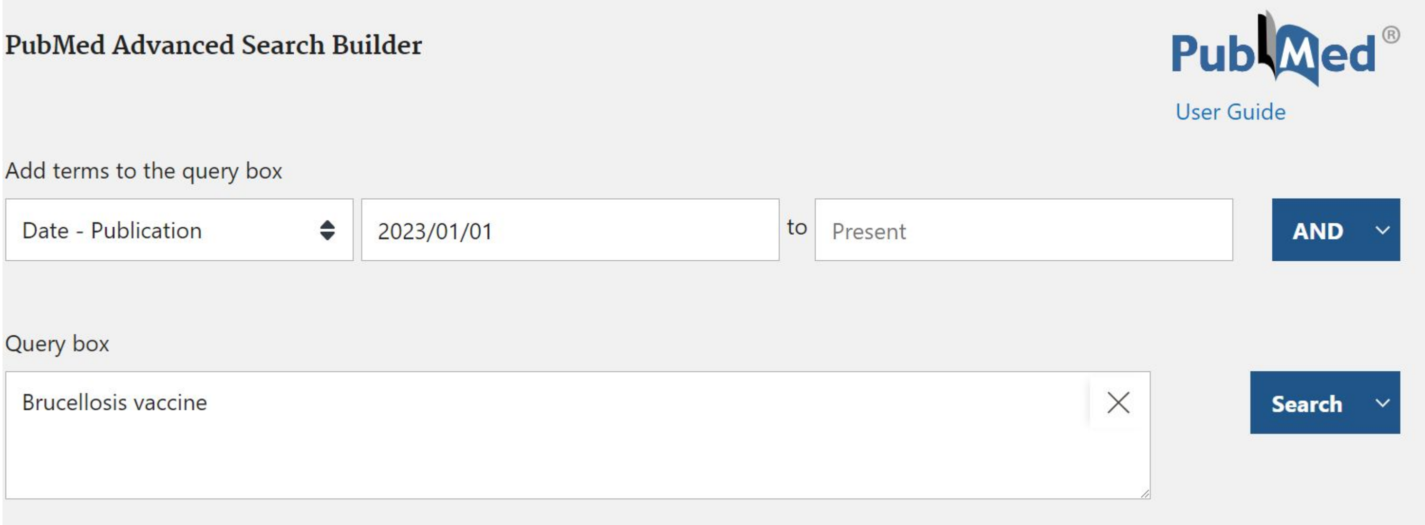
Safetensors

Model size 8.03B params

Tensor type BF16

Inference Examples

PubMed literature Mining



PubMed Advanced Search Builder

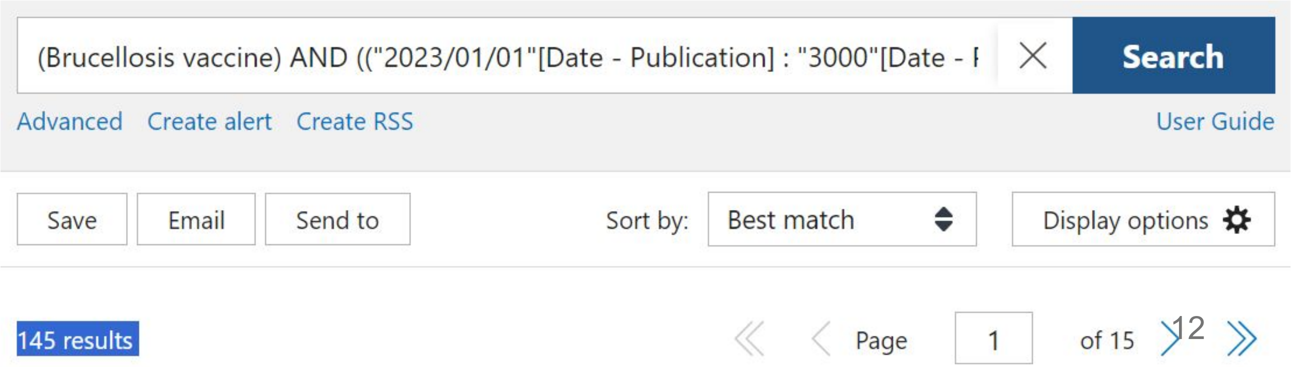
Add terms to the query box

Date - Publication: 2023/01/01 to Present

Query box

Brucellosis vaccine

- from 2023 to 2024
- "*Brucella* vaccine" term keyword search
- identify 148 papers



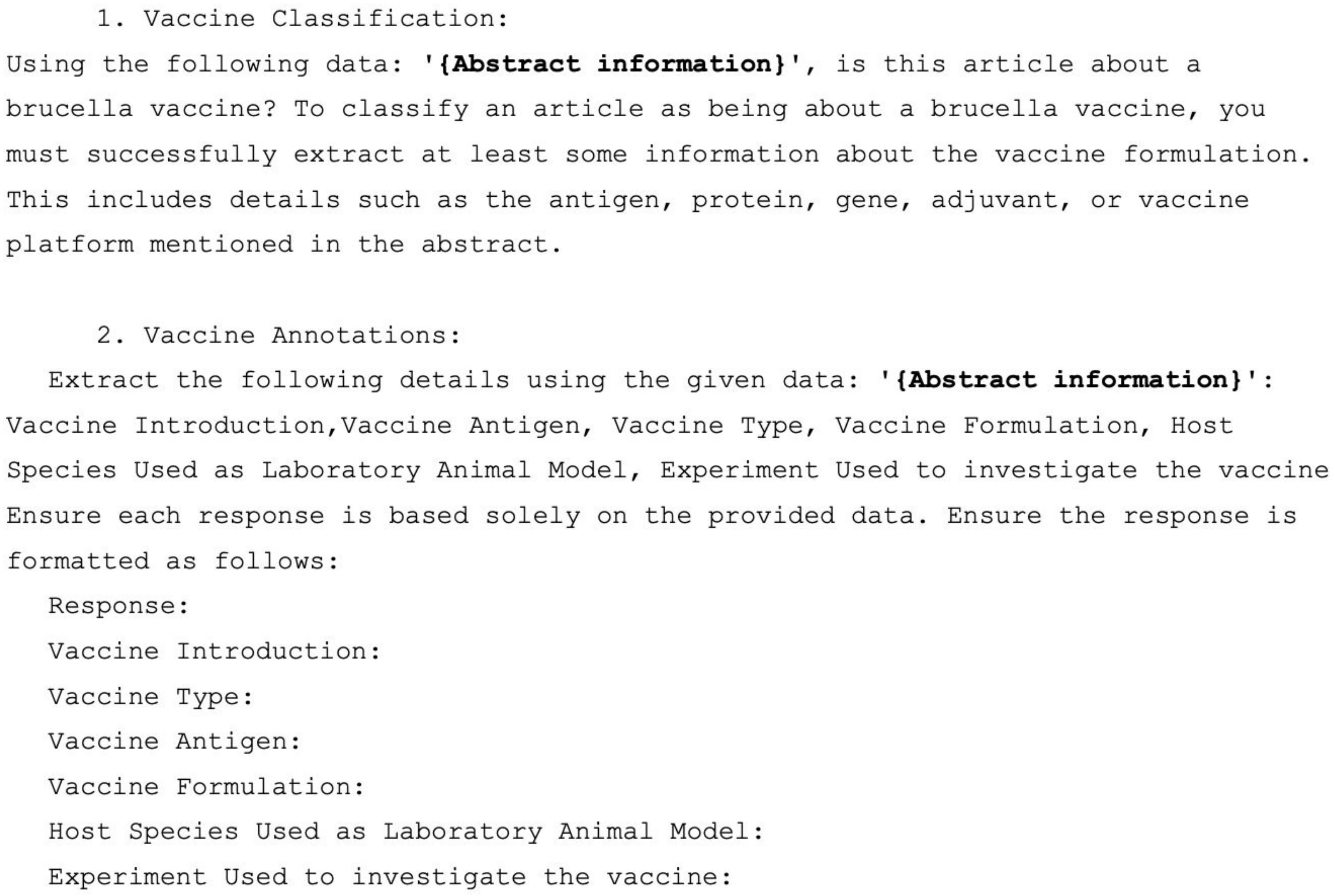
(Brucellosis vaccine) AND (("2023/01/01"[Date - Publication] : "3000"[Date - P

Sort by: Best match

145 results

Page 1 of 15

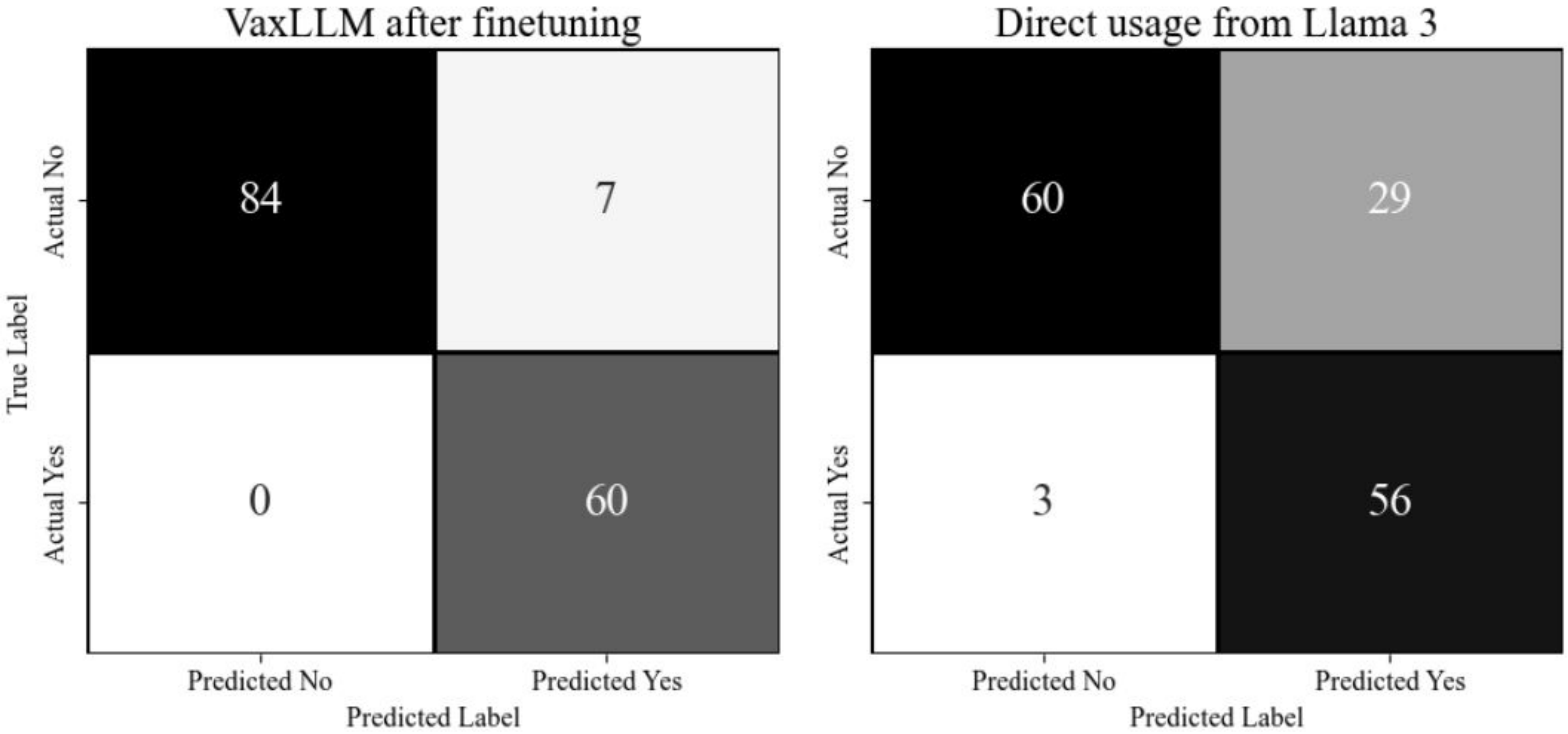
Prompt Engineering



```
     1. Vaccine Classification:
Using the following data: '{Abstract information}', is this article about a
brucella vaccine? To classify an article as being about a brucella vaccine, you
must successfully extract at least some information about the vaccine formulation.
This includes details such as the antigen, protein, gene, adjuvant, or vaccine
platform mentioned in the abstract.

     2. Vaccine Annotations:
  Extract the following details using the given data: '{Abstract information}':
Vaccine Introduction,Vaccine Antigen, Vaccine Type, Vaccine Formulation, Host
Species Used as Laboratory Animal Model, Experiment Used to investigate the vaccine
Ensure each response is based solely on the provided data. Ensure the response is
formatted as follows:
  Response:
  Vaccine Introduction:
  Vaccine Type:
  Vaccine Antigen:
  Vaccine Formulation:
  Host Species Used as Laboratory Animal Model:
  Experiment Used to investigate the vaccine:
```

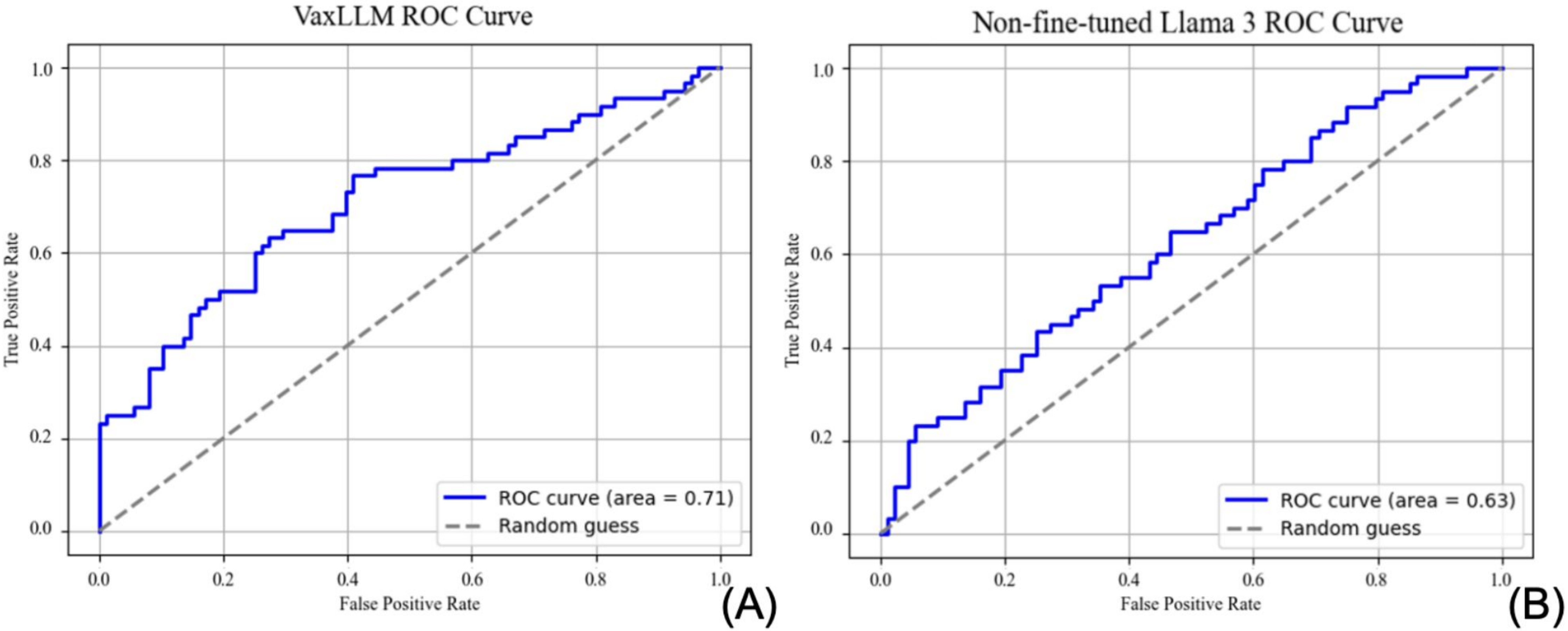
Classification results



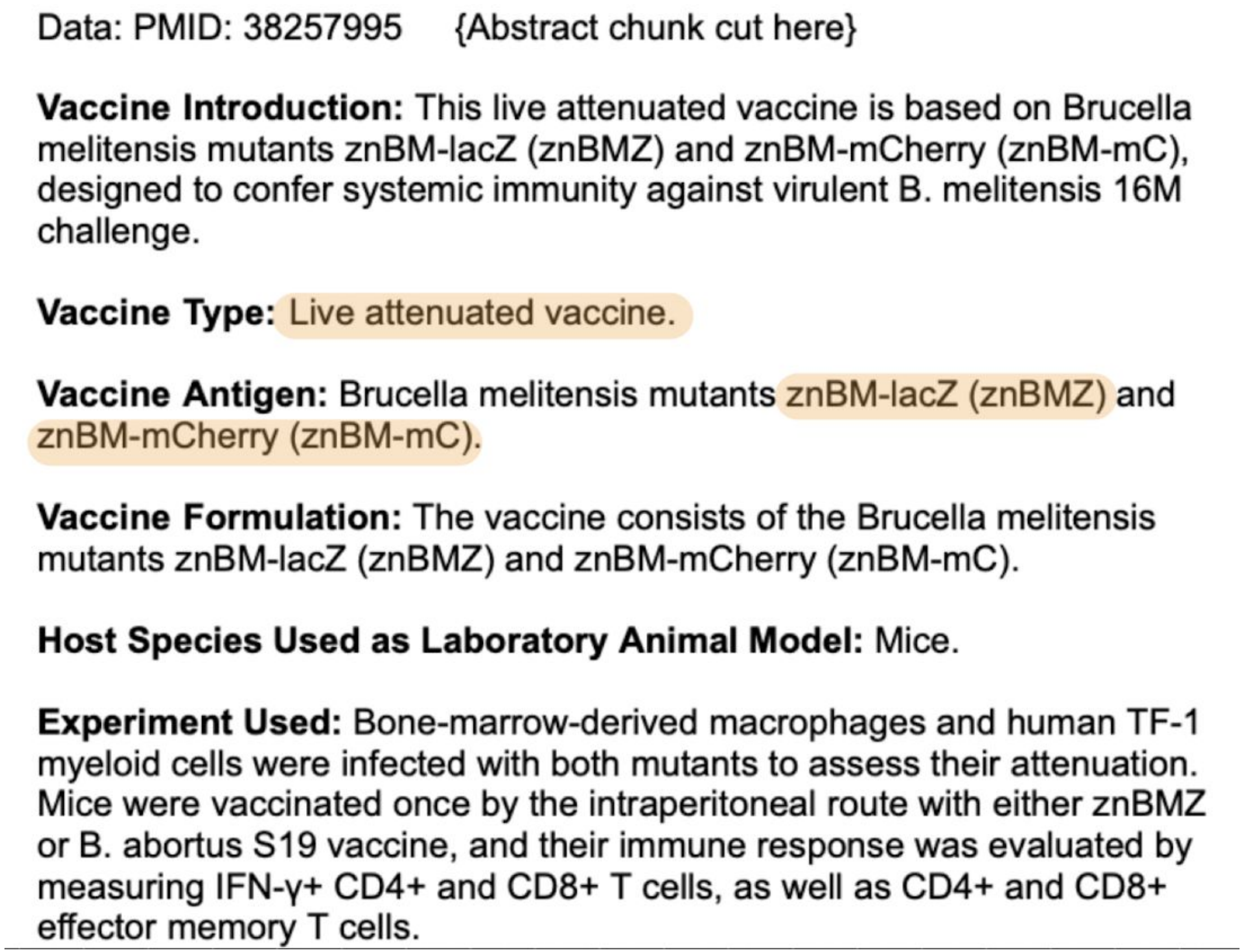
VaxLLM after finetuning

Direct usage from Llama 3

Classification results



VaxLLM annotation output



Data: PMID: 38257995 {Abstract chunk cut here}

Vaccine Introduction: This live attenuated vaccine is based on Brucella melitensis mutants znBM-lacZ (znBMZ) and znBM-mCherry (znBM-mC), designed to confer systemic immunity against virulent B. melitensis 16M challenge.

Vaccine Type: Live attenuated vaccine.

Vaccine Antigen: Brucella melitensis mutants znBM-lacZ (znBMZ) and znBM-mCherry (znBM-mC).

Vaccine Formulation: The vaccine consists of the Brucella melitensis mutants znBM-lacZ (znBMZ) and znBM-mCherry (znBM-mC).

Host Species Used as Laboratory Animal Model: Mice.

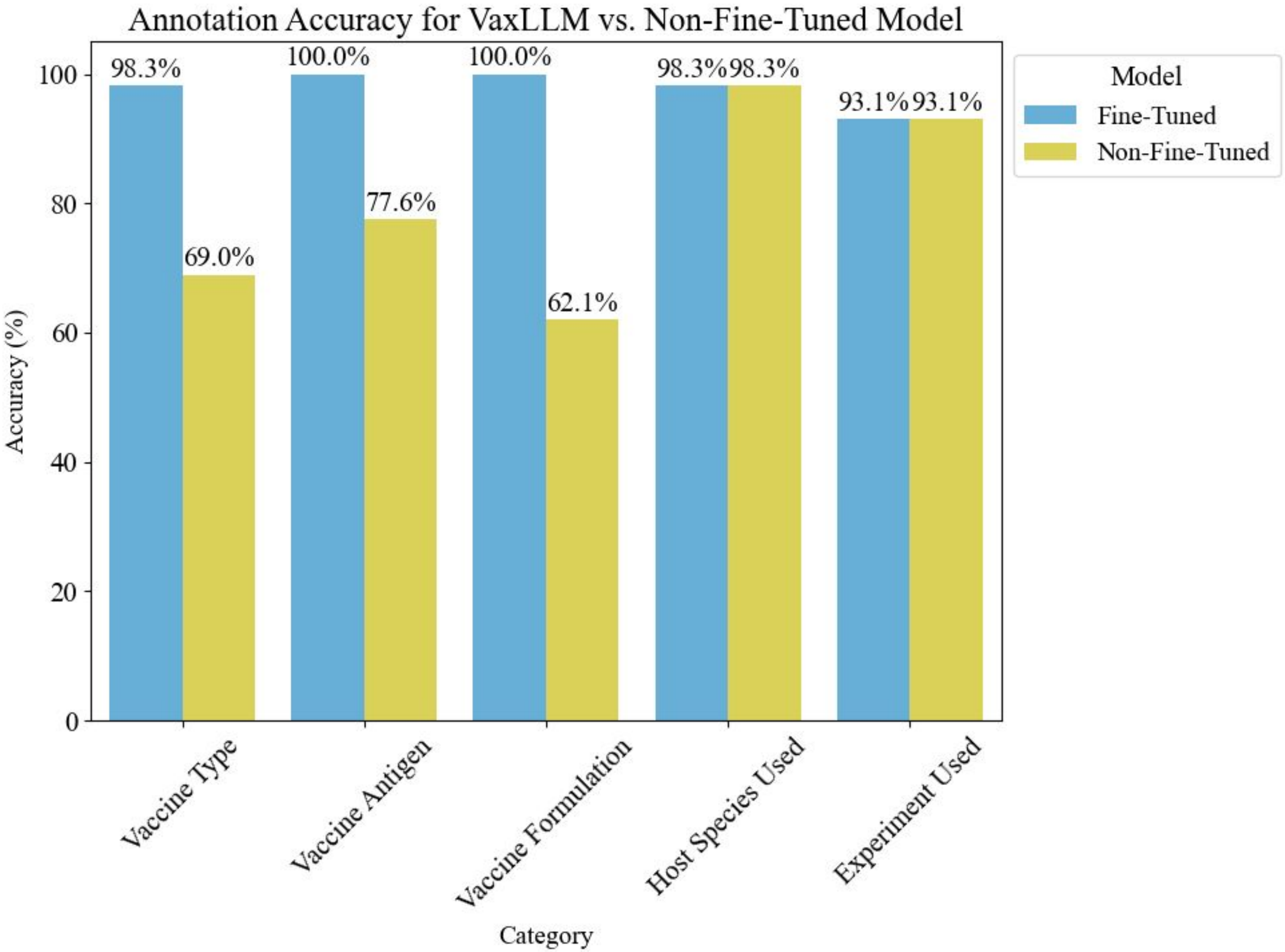
Experiment Used: Bone-marrow-derived macrophages and human TF-1 myeloid cells were infected with both mutants to assess their attenuation. Mice were vaccinated once by the intraperitoneal route with either znBMZ or B. abortus S19 vaccine, and their immune response was evaluated by measuring IFN-γ+ CD4+ and CD8+ T cells, as well as CD4+ and CD8+ effector memory T cells.

VaxLLM usage for paper PMID 38257995

correct information

VaxLLM annotation accuracy

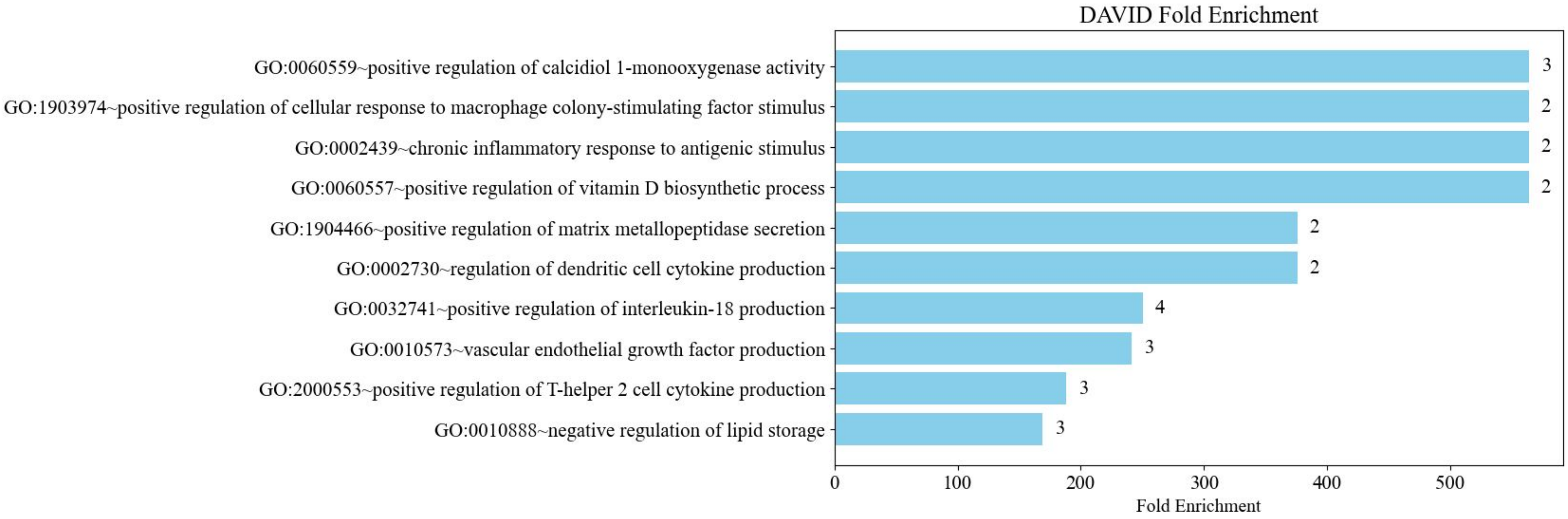
Overall accuracy: 97.9%



Potential downstream use case

Brucella vaccines data-analysis

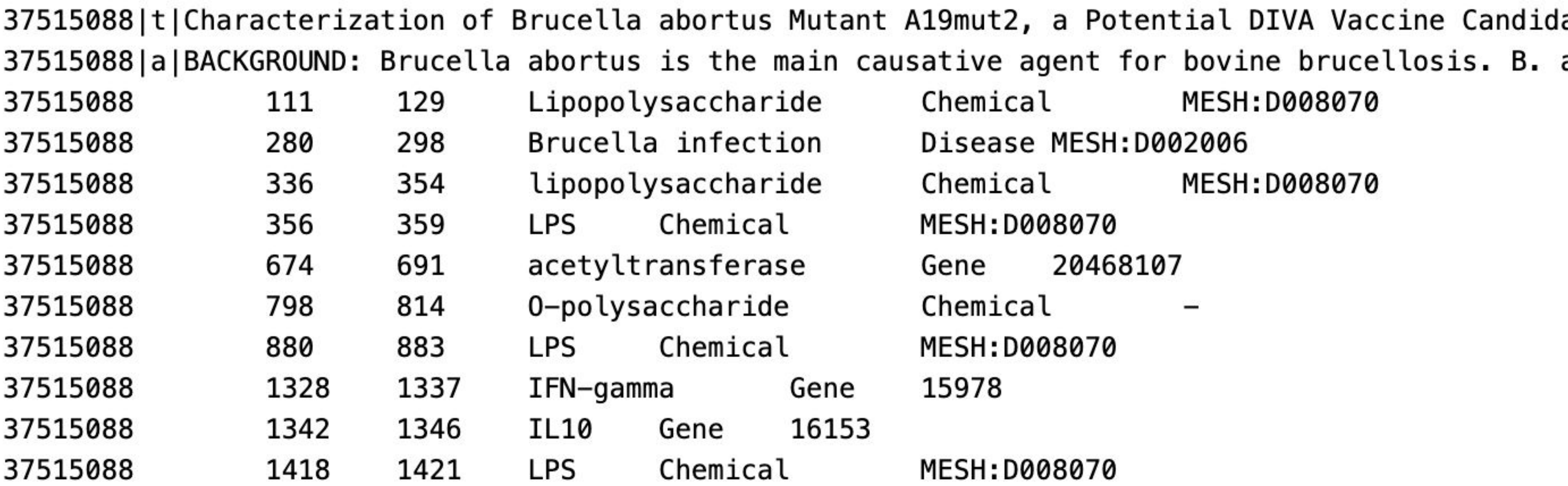
- demonstrate here as gene enrichment analysis



PubTator usage

- **PubTator — uses Name Entity Recognition (NER) to extract the following information:**
    - Genes (both antigen and host genes)
    - Protein
    - Chemicals, if relevant
    - Species mentioned

Pubtator result example



```
37515088|t|Characterization of Brucella abortus Mutant A19mut2, a Potential DIVA Vaccine Candida
37515088|a|BACKGROUND: Brucella abortus is the main causative agent for bovine brucellosis. B. a
37515088        111     129     Lipopolysaccharide      Chemical        MESH:D008070
37515088        280     298     Brucella infection      Disease MESH:D002006
37515088        336     354     lipopolysaccharide      Chemical        MESH:D008070
37515088        356     359     LPS     Chemical        MESH:D008070
37515088        674     691     acetyltransferase       Gene    20468107
37515088        798     814     O-polysaccharide        Chemical        -
37515088        880     883     LPS     Chemical        MESH:D008070
37515088        1328    1337    IFN-gamma       Gene    15978
37515088        1342    1346    IL10    Gene    16153
37515088        1418    1421    LPS     Chemical        MESH:D008070
```

Gene enrichment analysis

- VaxLLM classify 60 articles as yes
- Run the Pubtator of all these articles
- Identify and extract the gene list of 37 standardized for DAVID gene enrichment analysis.

Meaning of VaxLLM

- Automates the annotation of *Brucella* vaccine literature.
- High accuracy and standardized outputs support database integration.
- May realize the integration of thousands of vaccine information in VIOLIN

Future direction and limitation

- Use *Brucella* vaccines as demo to text methodology, may implement more vaccines
- Full text paper Vs. abstract
- may provide standardization (ex. map to ontology)

Data Availability

VaxLLM data and sample code:
https://github.com/xingxianli/VaxLLM

Model availability:

https://huggingface.co/Xingxian123/VaxLLM

Q & A

MICHIGAN

MICHIGAN MEDICINE
UNIVERSITY OF MICHIGAN

He Group