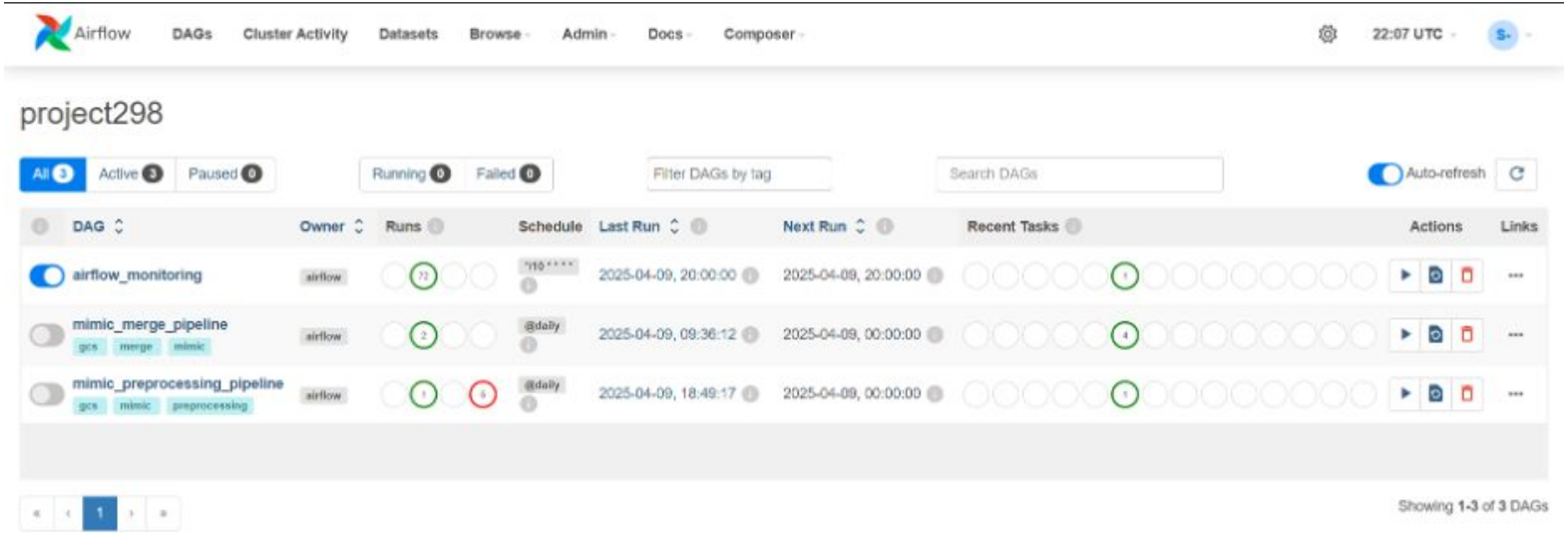
Generative AI Healthcare Agents Pipeline

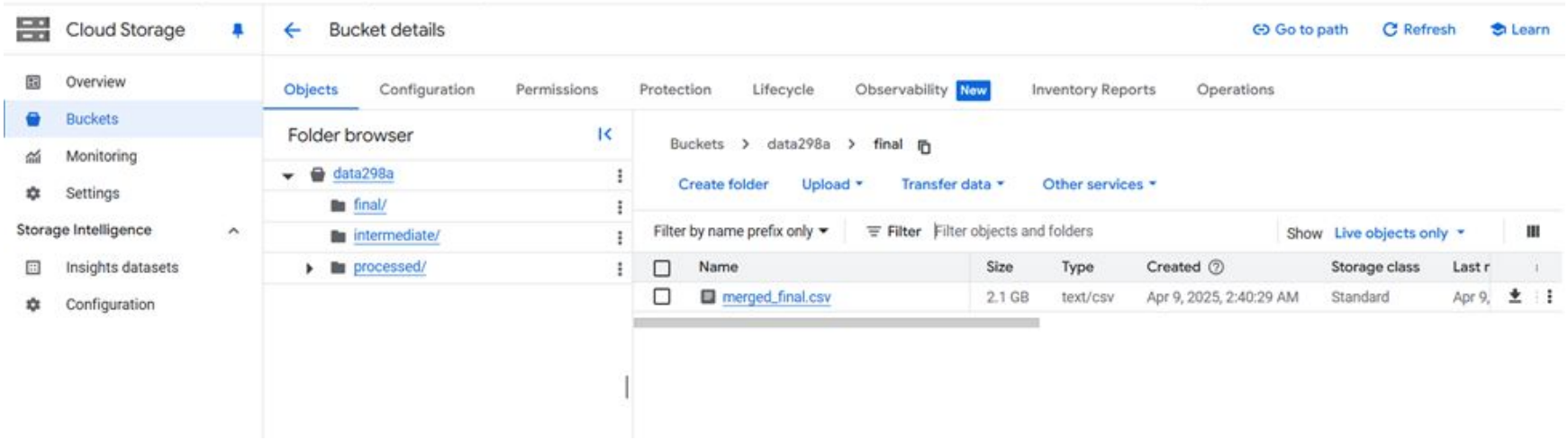
Team 3

- Aparna Bharathi Suresh
- Pavan Srivatsav Devarakonda
- Sai Naga Sanjana Chippada
- Shravani Dattaram Gawade
- Sujata Deepraj Joshi

Data Processing through DAGs



Data Transformation



Transformed MIMIC_III dataset

instruction	input	output
does the patient have a current copd exacerbation	Chief Complaint: copd exacerbation/shortness	chief complaint: copd exacerbation/shortness of breath
does the patient have a history of shortness of breath	Chief Complaint: copd exacerbation/shortness	chief complaint: copd exacerbation/shortness of breath
has been the patient ever been considered for copd exacerbation	Chief Complaint: copd exacerbation/shortness	chief complaint: copd exacerbation/shortness of breath
does the patient have a prior history of shortness of breath	Chief Complaint: copd exacerbation/shortness	chief complaint: copd exacerbation/shortness of breath
what is the patient 's copd exacerbation status	Chief Complaint: copd exacerbation/shortness	chief complaint: copd exacerbation/shortness of breath
does the patient have any copd exacerbation history	Chief Complaint: copd exacerbation/shortness	chief complaint: copd exacerbation/shortness of breath
has the patient had previous shortness of breath	Chief Complaint: copd exacerbation/shortness	chief complaint: copd exacerbation/shortness of breath
has been the patient ever been considered for acute worsening in dyspnea	Chief Complaint: copd exacerbation/shortness	this morning patient developed an acute worsening in dyspnea, and called ems.
what is the patient 's history of acute worsening in dyspnea	Chief Complaint: copd exacerbation/shortness	this morning patient developed an acute worsening in dyspnea, and called ems.
has this patient ever had acute worsening in dyspnea	Chief Complaint: copd exacerbation/shortness	this morning patient developed an acute worsening in dyspnea, and called ems.
does the patient have any history of previous acute worsening in dyspnea	Chief Complaint: copd exacerbation/shortness	this morning patient developed an acute worsening in dyspnea, and called ems.
has tachypnea been considered in the past	Chief Complaint: copd exacerbation/shortness	ems found patient tachypnic at saturating 90% on 5l.
is the patient tachypnic	Chief Complaint: copd exacerbation/shortness	ems found patient tachypnic at saturating 90% on 5l.
does the patient have any history of previous tachypnic	Chief Complaint: copd exacerbation/shortness	ems found patient tachypnic at saturating 90% on 5l.

Transformed MIMIC_IV dataset

instruction	input	output
	Gender: F Chief Complaint: Abdominal distention Diagnoses: OTHER ASCITES, UNSPECIFIED VIRAL HEPATITIS C WITHOUT HEPATIC COMA,	
What was the patient's condition upon discharge?	CIRRHOSIS OF LIVER NOS,	The patient was mentally clear and coherent, ambulatory, and independent.
What precautions does the patient need to take after discha	Gender: F	To follow a low sodium diet and fluid restriction, and contact their healthcare providers if they experience abdominal pain, fever, confusion, or other concernin
Who is responsible for coordinating the patient's care?	Gender: F	The patient's insurance company is handling the coordination of care.
Are there any further procedures planned for the patient?	Gender: F	Yes, the patient requires regular paracentesis sessions and possibly a surgery consultation for variceal screening.
Does the patient require long term monitoring?	Gender: F	Yes, the patient requires close monitoring for potential recurrence of ascites and related conditions such as hepatic encephalopathy, hyponatremia, and cirrh
What is the patient's current status regarding HIV managem	Gender: F	The patient continues taking Truvada and Isentress while awaiting an appointment with an infectious diseases specialist.
How often should the patient receive paracentesis?	Gender: F	Frequency of paracentesis is unclear based on the provided information.
What is the patient's social situation?	Gender: F	The patient lives alone and has limited support system, having lost touch with many relatives except for one brother who resides far away.
What medications did the patient take before being admitte	Gender: F	Before being admitted, the patient took Lactulose 15 mL PO TID, Tiotropium Bromide 1 CAP IH DAILY, Raltegravir 40 mg PO BID, Emtricitabine-Tenofovir (Truva
What medications did the patient take before being admitte	Gender: F	Before being admitted, the patient took Lactulose 15 mL PO TID, Tiotropium Bromide 1 CAP IH DAILY, Raltegravir 40 mg PO BID, Emtricitabine-Tenofovir (Truva
Why was the patient admitted to the hospital?	Gender: F	the patient was admitted to the hospital due to altered mental status caused by hepatic encephalopathy related to her HCV cirrhosis and complications.
Why was the patient admitted to the hospital?	Gender: F	the patient was admitted to the hospital due to altered mental status caused by hepatic encephalopathy related to her HCV cirrhosis and complications.

Transformed MedQuAD dataset code

```python
import pandas as pd

# Load MedQuAD
df = pd.read_csv("medquad.csv")

# Clean whitespace and drop NaNs
df = df.dropna(subset=["question", "focus_area", "answer"])
df["question"] = df["question"].str.strip()
df["focus_area"] = df["focus_area"].str.strip()
df["answer"] = df["answer"].str.strip()

# Remove very short or empty responses
df = df[df["answer"].str.len() > 20]

# Remove duplicate Q-A pairs
df = df.drop_duplicates(subset=["question", "answer"])

# Rename to match BioMistral format
df = df.rename(columns={
    "question": "instruction",
    "focus_area": "input",
    "answer": "output"
})

# Optional: Normalize line breaks and spacing
for col in ["instruction", "input", "output"]:
    df[col] = df[col].str.replace(r"\s+", " ", regex=True)

# Save preprocessed version
df.to_csv("medquad_preprocessed.csv", index=False)
print("✅ Saved: medquad_preprocessed.csv with", len(df), "rows")
```

```
✅ Saved: medquad_preprocessed.csv with 16344 rows
```

Transformed MedQuAD dataset

instruction	output	input
What is (are) Glaucoma ?	Glaucoma is a group of diseases that can damage the eye's optic nerve and result in visior	Glaucoma
What causes Glaucoma ?	Nearly 2.7 million people have glaucoma, a leading cause of blindness in the United State:	Glaucoma
What are the symptoms of Glaucoma ?	Symptoms of Glaucoma Glaucoma can develop in one or both eyes. The most common ty	Glaucoma
What are the treatments for Glaucoma ?	Although open-angle glaucoma cannot be cured, it can usually be controlled. While treatr	Glaucoma
What is (are) Glaucoma ?	Glaucoma is a group of diseases that can damage the eye's optic nerve and result in visior	Glaucoma
What is (are) Glaucoma ?	The optic nerve is a bundle of more than 1 million nerve fibers. It connects the retina to the	Glaucoma
What is (are) Glaucoma ?	Open-angle glaucoma is the most common form of glaucoma. In the normal eye, the clear	Glaucoma
Who is at risk for Glaucoma? ?	Anyone can develop glaucoma. Some people are at higher risk than others. They include -	Glaucoma
How to prevent Glaucoma ?	At this time, we do not know how to prevent glaucoma. However, studies have shown that	Glaucoma
What are the symptoms of Glaucoma ?	At first, open-angle glaucoma has no symptoms. It causes no pain. Vision seems normal. '	Glaucoma
What are the treatments for Glaucoma ?	Yes. Immediate treatment for early stage, open-angle glaucoma can delay progression of	Glaucoma
what research (or clinical trials) is being done for Glaucoma ?	Through studies in the laboratory and with patients, the National Eye Institute is seeking be	Glaucoma
Who is at risk for Glaucoma? ?	Encourage them to have a comprehensive dilated eye exam at least once every two years.	Glaucoma
What is (are) Glaucoma ?	National Eye Institute National Institutes of Health 2020 Vision Place Bethesda, MD 20892	Glaucoma
What is (are) High Blood Pressure ?	High blood pressure is a common disease in which blood flows through blood vessels (art	High Blood Pressure
What causes High Blood Pressure ?	Changes in Body Functions Researchers continue to study how various changes in normal	High Blood Pressure
Who is at risk for High Blood Pressure? ?	Not a Normal Part of Aging Nearly 1 in 3 American adults have high blood pressure. Many p	High Blood Pressure
How to prevent High Blood Pressure ?	Steps You Can Take You can take steps to prevent high blood pressure by adopting these h	High Blood Pressure

Transformed iCliniQ dataset code

```python
import pandas as pd

# Load the dataset
df = pd.read_csv("icliniq_medical_qa_cleaned.csv")

# Drop rows with missing fields
df = df.dropna(subset=["Title", "Question", "Answer"])

# Rename to standard format
df = df.rename(columns={
    "Question": "instruction",
    "Title": "input",
    "Answer": "output"
})

# Clean whitespace
for col in ["instruction", "input", "output"]:
    df[col] = df[col].astype(str).str.replace(r"\s+", " ", regex=True).str.strip()

# Filter: remove very short answers (<20 characters)
df = df[df["output"].str.len() > 20]

# Remove duplicates
df = df.drop_duplicates(subset=["instruction", "output"])

# Save final processed version
df.to_csv("icliniq_preprocessed.csv", index=False)
print(f"✅ Saved: icliniq_preprocessed.csv with {len(df)} rows")
```

```
✅ Saved: icliniq_preprocessed.csv with 39409 rows
```

Transformed ICliniq dataset

Title	Question	Answer
How can a nasal spray efficiently relieve sinusitis and nasal polyps?	My son has sinusitis and also nasal polyps, we got to know this recently. He suffers from a nose block and breathing issues. So when we visited a PCP primary care physician, he suggested using nasal spray. So my doubt is, how can nasal spray efficiently relieve	one of which is to filter the air that we breathe. The other functions such as humidification, a pathway for olfaction smell, etc. The exposure to dust or any irritative substance that your body is hyper-sensitive to, will initially initiate
What is menopause and how to minimize its symptoms?	As I approach middle age, I have been experiencing symptoms that align with menopause, but I have also come across information on late-onset hypogonadism. Can you help me understand the distinctions between these conditions, and how they may manifest in middle-aged women? Additionally, what are the potential treatments or management strategies for addressing	life of a woman because this is the time when the reproductive system has reached its peak and the decline will occur but if this decline is supported by certain lifestyle changes it will lead to happy menopause leading you to have least effects of menopausal symptoms like hot blushes. So, I suggest you follow a few lifestyle
What are the COVID-19 booster doses?	Hello doctor, I read in the news that coronavirus is agair	authorized for emergency use by regulatory agencies like the FDA The United States Food and Drug Administration and EMA The European Medicines Agency are highly effective at preventing severe disease and hospitalization from COVID-19. However, it is important to remember that no vaccine is 100 percent effective, and breakthrough infections may still occur in some vaccinated individuals.

merge csv file with processed medquad and icliniq datasets.

```python
import pandas as pd

# Load preprocessed files
mimic = pd.read_csv("mimic_qa_combined.csv")
medquad = pd.read_csv("medquad_preprocessed.csv")
icliniq = pd.read_csv("icliniq_preprocessed.csv")

# Ensure column names are standardized
for df in [mimic, medquad, icliniq]:
    df.columns = [col.strip().lower() for col in df.columns]

# Ensure all required columns are present
columns = ["instruction", "input", "output"]
for df in [mimic, medquad, icliniq]:
    if not all(col in df.columns for col in columns):
        raise ValueError("Missing one of 'instruction', 'input', or 'output' in a dataset.")

# Combine
final_df = pd.concat(
    [mimic[columns], medquad[columns], icliniq[columns]],
    ignore_index=True
)

# Optional: deduplicate
final_df.drop_duplicates(subset=["instruction", "input", "output"], inplace=True)

# Save merged dataset
final_df.to_csv("bio_mistral_qa_combined.csv", index=False)
print(f"✅ Saved: bio_mistral_qa_combined.csv with {len(final_df)} rows")
```

```
✅ Saved: bio_mistral_qa_combined.csv with 112292 rows
```

Final Merged CSV for all datasets

instruction	input	output
What was the patient's condition upon discharge?	Gender: F Chief Complaint: Abdominal distention Diagnoses: OTHER ASCITES, UNSPECIFIED VIRAL	The patient was mentally clear and coherent, ambulatory, and independent.
What precautions does the patient need to take after discharge?	Gender: F Chief Complaint: Abdominal distention Diagnoses: OTHER ASCITES, UNSPECIFIED VIRAL HEPATITIS C WITHOUT HEPATIC COMA, CIRRHOSIS OF LIVER NOS, ASYMPTOMATIC HIV INFECTION Vitals: Temp 97.9, HR 81.0, RR 17.5, O2Sat 94.0, SBP	To follow a low sodium diet and fluid restriction, and contact their healthcare providers if th
Who is responsible for coordinating the patient's care?	Gender: F Chief Complaint: Abdominal distention Diagnoses: OTHER ASCITES, UNSPECIFIED VIRAL HEPATITIS C WITHOUT HEPATIC COMA, CIRRHOSIS OF LIVER NOS, ASYMPTOMATIC HIV INFECTION Vitals: Temp 97.9, HR 81.0, RR 17.5, O2Sat 94.0, SBP 95.5, DBP 60.5	The patient's insurance company is handling the coordination of care.

Model Selection:

Selected open-source LLMs optimized for chronic disease management, ensuring accuracy, scalability, and patient engagement.

Selected Models:

- **MedLlama3**: General-purpose Q&A and patient education
- **BioMistral-7B**: Biomedical knowledge retrieval and clinical research
- **BioGPT**: Generative medical text and conversational Q&A
- **Meditron**: Advanced medical reasoning and differential diagnosis

We have currently fine-tuned the BioMistral-7B model on healthcare-specific Q/A datasets, to enhance its performance in biomedical knowledge retrieval and clinical research queries.

Base-Model code

for BioMistral-7B

Loading Biomistral model

```
# Block 4: Load BioMistral 7B Model
def load_model_and_tokenizer(model_name="BioMistral/BioMistral-7B"):
    """Load BioMistral 7B model and tokenizer with quantization."""
    tokenizer = AutoTokenizer.from_pretrained(model_name)
    if tokenizer.pad_token is None:
        tokenizer.pad_token = tokenizer.eos_token
    bnb_config = BitsAndBytesConfig(
        load_in_4bit=True,
        bnb_4bit_quant_type="nf4",
        bnb_4bit_compute_dtype=torch.float16,
        bnb_4bit_use_double_quant=True
    )
    model = AutoModelForCausalLM.from_pretrained(
        model_name,
        quantization_config=bnb_config,
        device_map="auto",
        torch_dtype=torch.float16,
        trust_remote_code=True
    )
    model.config.pad_token_id = tokenizer.eos_token_id
    model.eval()
    return model, tokenizer

model, tokenizer = load_model_and_tokenizer()
print(f"BioMistral-7B model and tokenizer loaded successfully!")
```

```
/opt/conda/lib/python3.10/site-packages/torch/_utils.py:831: UserWarning: Typ
e will be the only storage class. This should only matter to you if you are u
ped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
BioMistral-7B model and tokenizer loaded successfully!
```

Fine-Tuned Model

code for

BioMistral-7B

[12]:
```python
# Load fine-tuned model and tokenizer
def load_fine_tuned_model(model_name="BioMistral/BioMistral-7B", checkpoint_dir="biomistral_finetuned"):
    """Load the fine-tuned QLoRA model and tokenizer."""
    try:
        tokenizer = AutoTokenizer.from_pretrained(model_name)
        if tokenizer.pad_token is None:
            tokenizer.pad_token = tokenizer.eos_token

        bnb_config = BitsAndBytesConfig(
            load_in_4bit=True,
            bnb_4bit_quant_type="nf4",
            bnb_4bit_compute_dtype=torch.float16,
            bnb_4bit_use_double_quant=False,
            llm_int8_enable_fp32_cpu_offload=True,
            llm_int8_skip_modules=["lm_head"]
        )

        base_model = AutoModelForCausalLM.from_pretrained(
            model_name,
            quantization_config=bnb_config,
            device_map="auto",
            torch_dtype=torch.float16,
            trust_remote_code=True,
            offload_folder="offload",
            low_cpu_mem_usage=True,
            offload_state_dict=True
        )
        base_model.config.pad_token_id = tokenizer.eos_token_id

        if os.path.exists(checkpoint_dir):
            if os.path.exists(os.path.join(checkpoint_dir, "adapter_model.bin")):
                print(f"Loading fine-tuned model from {checkpoint_dir}")
                model = PeftModel.from_pretrained(base_model, checkpoint_dir, is_trainable=False)
            else:
                checkpoints = [d for d in os.listdir(checkpoint_dir) if d.startswith("checkpoint-")]
                if checkpoints:
                    latest_checkpoint = max(checkpoints, key=lambda x: int(x.split("-")[1]))
                    checkpoint_path = os.path.join(checkpoint_dir, latest_checkpoint)
                    print(f"Loading fine-tuned model from checkpoint: {checkpoint_path}")
                    model = PeftModel.from_pretrained(base_model, checkpoint_path, is_trainable=False)
                else:
                    raise ValueError(f"No checkpoints or final model found in {checkpoint_dir}")
        else:
            raise ValueError(f"Checkpoint directory {checkpoint_dir} does not exist")

        return model, tokenizer
    except Exception as e:
        print(f"Error loading fine-tuned model: {e}")
        return None, None

model, tokenizer = load_fine_tuned_model()
if model is None or tokenizer is None:
    raise ValueError("Failed to load fine-tuned model or tokenizer")
```

```
/opt/conda/lib/python3.10/site-packages/torch/_utils.py:831: UserWarning: TypedStorage is deprecated. It will be removed in the future a
nd UntypedStorage will be the only storage class. This should only matter to you if you are using storages directly.  To access UntypedS
torage directly, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
Loading fine-tuned model from checkpoint: biomistral_finetuned/checkpoint-33573
```

Evaluation Metrics Comparison Table

Metric	Baseline (pre-trained)	Fine-Tuned (Healthcare-Specific)
BERTScore Precision	0.8	0.8
BERTScore Recall	0.86	0.89
BERTScore F1	0.83	0.84
ROUGE-L	0.08	0.08
Entity F1	0.2	0.1
FCS (Factual Consistency Score)	0.3859	0.53
MCR (Medical Concept Recall)	0.2	0.2
BLEU	0.01	0.02
Hybrid BERT-BLEU	0.58	0.59
LLM Judge Score	0.52	0.84
GEval Score	0.44	0.75
METEOR	0.16	0.24

THANK YOU