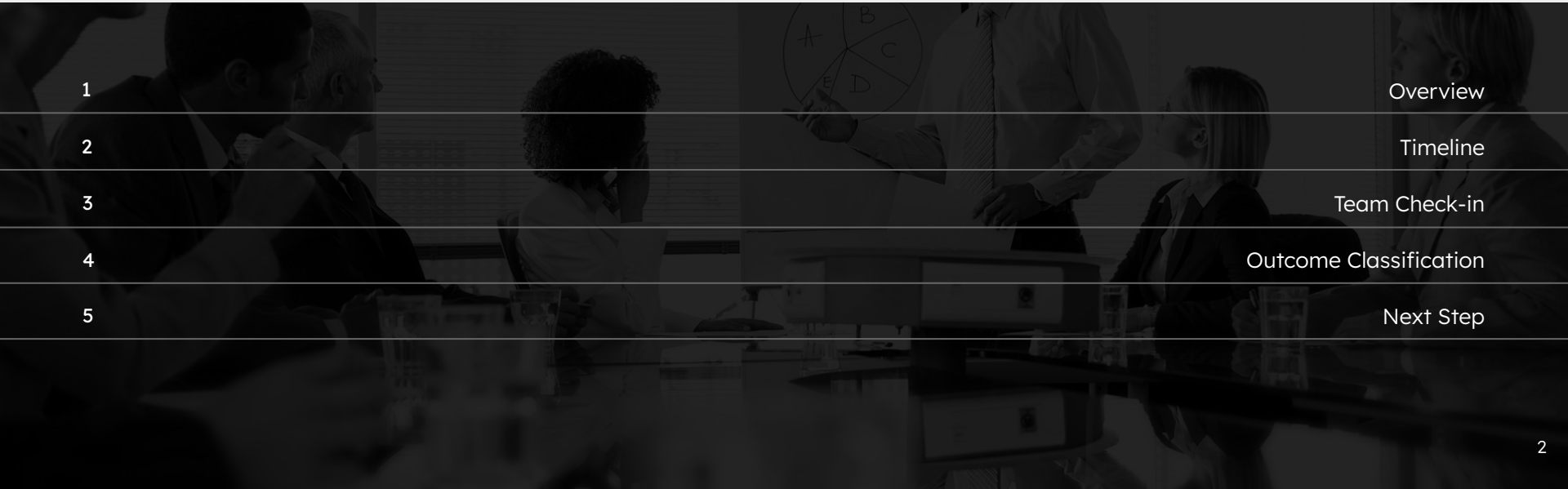
IBM x Columbia Weekly Meeting

IBM x Columbia Core Team

11/06/2025

AGENDA

Why we're here today



1 Overview

2 Timeline

3 Team Check-in

4 Outcome Classification

5 Next Step

Overview

This weekly meeting will be focusing on establishing a clear roadmap and reviewing initial progress for the IBM x Columbia project. Including the timeline, team check-in, and some Q&A at the end.

Timeline

Phase	Start Week	End Week	W1 09/23-09/29	W2 09/30-10/06	W3 10/07-10/13	W4 10/14-10/20	W5 10/21-10/27	W6 10/28-11/03	W7 11/04-11/10	W8 11/11-11/17	W9 11/18-11/24	W10 11/25-12/01	W11 12/02-12/08
Initiation & Planning	1	2											
Objective Clarification													
Literature Review													
Timeline & scope													
Data Exploration & Requirements	3	7											
Data Collecting & Preprocessing													
Data Cleaning & Structuring													
RAG Pipeline Prototyping													
Prototype Development (POC)	3	7											
MCP Workflow Diagrams													
Wireframes													
UI UX Mockups													
GenAI coding													
Prototype Implementation													
First demo refining													
Full Solution Development & Testing	9	11											
Performance Test													
Usability feedback													
Model refining													
Final Presentation & Reporting	9	11											
Project Powerpoint													
Project Final Report Drafting													
Demo Presentation													

Team Check-in

1. Data & Knowledge Engineering Team
2. Generative AI & Model Development Team
3. User Experience & Application Development Team
4. Evaluation & Compliance Team

Data & Knowledge Engineering Team

Embedding Model Upgrade: From MiniLM → BGE-Large-en-v1.5

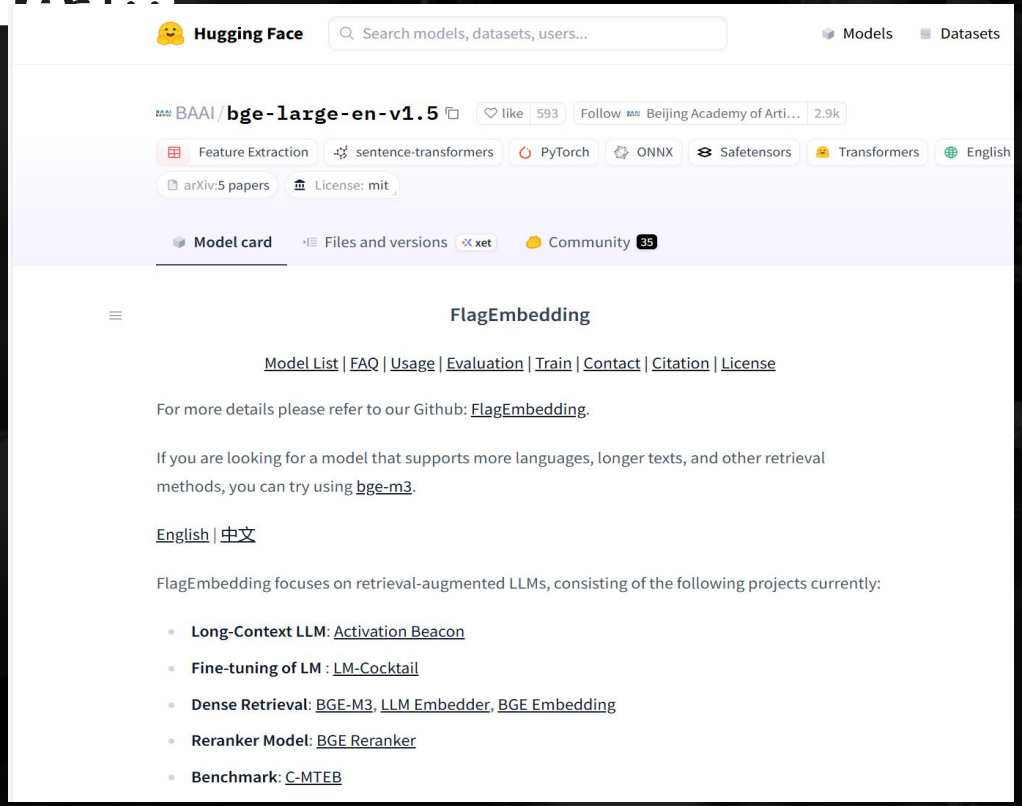
Why We Switched：

- The original model `all-MiniLM-L6-v2` showed several issues in our EPLC document retrieval pipeline:
  - Cosine similarity scores mostly clustered around **0.5–0.6**, making related and unrelated sections hard to separate
  - Retrieval results often matched semantically close but **incorrect sections**

Improvement with BGE-Large-en-v1.5:

- New model trained for **retrieval-augmented tasks (RAG)**
  - Embedding dimension increased **384 → 1024**, enabling finer semantic granularity
  - Similarity distribution now typically ranges **0.7–0.85**, providing clearer differentiation

Data & Knowledge Engineering Team



Data & Knowledge Engineering Team

[24]
```
def ask(query, df, model, top_k=3):
    from numpy import dot
    from numpy.linalg import norm
    q_vec = model.encode(query, normalize_embeddings=True)
    df["similarity"] = df["embedding"].apply(lambda e: dot(e, q_vec)/(norm(e)*norm(q_vec)))
    top = df.sort_values("similarity", ascending=False).head(top_k)
    print(f"\n Query: {query}")
    print(top[["section_number", "title", "similarity"]])
```

[25]
```
ask("Who approves the physical data model?", df, model)
ask("What is the purpose of the Physical Data Model document?", df, model)
```

```
 Query: Who approves the physical data model?
  section_number                                    title  similarity
1              2  General Overview and Design Guidelines/Approach    0.778102
4              A         Appendix A: Physical Data Model Approval    0.739013
0            1.1                                  Purpose    0.736968

 Query: What is the purpose of the Physical Data Model document?
  section_number                                    title  similarity
0            1.1                                  Purpose    0.840698
1              2  General Overview and Design Guidelines/Approach    0.772642
5              B                          Appendix B: References    0.754903
```

[ ]
开始借助 AI 编写或生成代码。

Data & Knowledge Engineering Team

Complete data cleaning for EPLC Framework ✅

Embedding for all CDC models✅(uploaded to Repo)

Word/PDF Templates

↓

Text Extraction + Cleaning (done)

↓

Standardized JSON (done)

↓

Chunk & Flatten (done)

↓

Embedding Model (Hugging Face)

↓

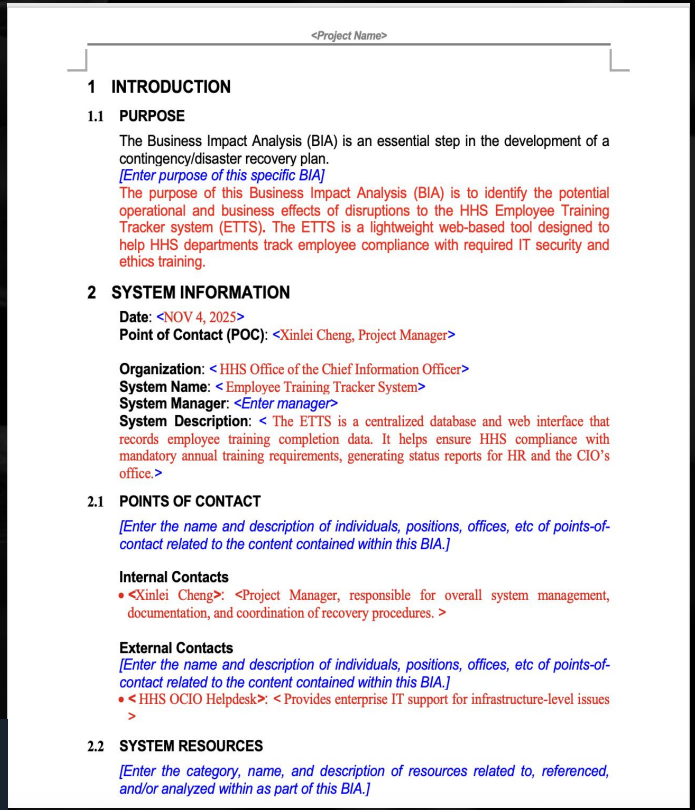
Vector Database (Chroma / FAISS / Pinecone)

↓

RAG Retrieval (GenAI Team)

Data & Knowledge Engineering Team

- Extract the policies from the HHS EPLC website
- Fill out one of the templates from the Implementation Phase



<Project Name>

1 INTRODUCTION

1.1 PURPOSE

The Business Impact Analysis (BIA) is an essential step in the development of a contingency/disaster recovery plan.
[Enter purpose of this specific BIA]
The purpose of this Business Impact Analysis (BIA) is to identify the potential operational and business effects of disruptions to the HHS Employee Training Tracker system (ETTS). The ETTS is a lightweight web-based tool designed to help HHS departments track employee compliance with required IT security and ethics training.

2 SYSTEM INFORMATION

Date: <NOV 4, 2025>
Point of Contact (POC): <Xinlei Cheng, Project Manager>

Organization: < HHS Office of the Chief Information Officer>
System Name: < Employee Training Tracker System>
System Manager: <*Enter manager*>
System Description: < The ETTS is a centralized database and web interface that records employee training completion data. It helps ensure HHS compliance with mandatory annual training requirements, generating status reports for HR and the CIO's office.>

2.1 POINTS OF CONTACT

[Enter the name and description of individuals, positions, offices, etc of points-of-contact related to the content contained within this BIA.]

Internal Contacts
- **<**Xinlei Cheng**>**: <Project Manager, responsible for overall system management, documentation, and coordination of recovery procedures. >

External Contacts
[Enter the name and description of individuals, positions, offices, etc of points-of-contact related to the content contained within this BIA.]
- **<** HHS OCIO Helpdesk**>**: < Provides enterprise IT support for infrastructure-level issues >

2.2 SYSTEM RESOURCES

[Enter the category, name, and description of resources related to, referenced, and/or analyzed within as part of this BIA.]

```
HTTPError: 403 Client Error: Forbidden for url: https://www.hhs.gov/web/governance/digital-strategy/it-policy-archive/policy-for-information-technology-enterprise-performance.html
```

GenAI & Model Development

Uploaded inspect_chroma.py:

Connected local database successfully

```
[inspect] DB path: /Users/violet/Desktop/Q&A/vector_db  exists=True
[inspect] Collections found:
  - name=Implementation_Phase | count=95
  - name=default | count=0

[inspect] Sampling first collection: Implementation_Phase
[inspect] sample ids:       ['Business Analysis Impact_embedded (1).json_0', 'Business Analysis Impact_embedded (1).json_1', 'Business Analysis Impact_embedded (1)
[inspect] sample metadatas: [{'source': 'Business Analysis Impact_embedded (1).json'}, {'source': 'Business Analysis Impact_embedded (1).json'}, {'source': 'Busine
[inspect] doc#1 preview: [Enter purpose of this specific BIA]
[inspect] doc#2 preview: [Enter the name and description of individuals, positions, offices, etc of points-of-contact related to the content cont
[inspect] doc#3 preview: [Enter the category, name, and description of resources related to, referenced, and/or analyzed within as part of this B
```

GenAI & Model Development

Steps	Components	Function
1	SBERT	Understanding Question
2	Chroma	Retrieve similar text
3	GPT	Summarize and explain the answer
4	Output	Return answer with citations

GenAI & Model Development

Embedding

```python
def embed(texts: List[str]) -> List[List[float]]:
    return sbert.encode(texts, normalize_embeddings=True).tolist()

def retrieve(query: str, k: int = TOP_K) -> Tuple[list, list, list]:
    qv = embed([query])
    res = coll.query(
        query_embeddings=qv,
        n_results=k,
        include=["documents", "metadatas", "distances"]
    )
    ids = res.get("ids", [[]])[0]
    docs = res.get("documents", [[]])[0]
    dists = res.get("distances", [[]])[0]
    return ids, docs, dists

def pretty_sim(dist: float) -> float:
    try:
        return 1.0 - float(dist)
    except Exception:
        return float("nan")
```

Answer Generation

```python
SYSTEM_PROMPT = (
    "You are an EPLC assistant. Answer only using the information in the CONTEXT. "
    "If the answer can be inferred from the context, explain it briefly. "
    "If the context provides no relevant information, reply exactly: Not specified in the provided context."
)

def make_prompt(question: str, docs: List[str]) -> str:
    context = "\n\n---\n\n".join(docs)
    return f"CONTEXT:\n{context}\n\nQUESTION:\n{question}\n"

def ask_openai(prompt: str) -> str:
    try:
        resp = oa.responses.create(
            model=CHAT_MODEL,
            input=[
                {"role": "system", "content": SYSTEM_PROMPT},
                {"role": "user", "content": prompt},
            ],
            temperature=0
        )
        return (resp.output_text or "").strip()
    except Exception as e:
        return f"[openai error] {e}"
```

GenAI & Model Development

Main Loop

```python
def main():
    try:
        cnt = coll.count()
        print(f"[startup] Loaded collection '{COLL_NAME}' with {cnt} records.")
    except Exception as e:
        print("[startup] Collection error:", e)
        sys.exit(1)

    print(f"[ready] Using GPT model: {CHAT_MODEL} | top_k={TOP_K}")
    print("Ask any EPLC question. Type 'exit' to quit.")

    while True:
        try:
            q = input("\nQ> ").strip()
        except (EOFError, KeyboardInterrupt):
            print("\nbye.")
            break

        if not q or q.lower() in {"exit", "quit"}:
            print("bye.")
            break

        ids, docs, dists = retrieve_exact(q, TOP_K)
        if not docs:
            ids, docs, dists = retrieve(q, TOP_K)
        if not docs:
            print("A> Not specified in the provided context.")
            continue

        prompt = make_prompt(q, docs)
        answer = ask_openai(prompt)
        print("\nA>", answer if answer else "Not specified in the provided context.")
        print("   citations:", ids)

        # Debug
        print(f"\n[DEBUG] ids={ids}")
        for i, (d, dist) in enumerate(zip(docs, dists), start=1):
            prev = (d or "")[:200].replace("\n", " ")
            print(f"[DEBUG ctx#{i}] dist={dist:.3f} | sim≈{pretty_sim(dist):.3f} | {prev}")
```

GenAI & Model Development

```
Q> what is purpose of service level agreement?

A> The purpose of the Service Level Agreement (SLA) is to define the overall purpose and intent of the agreement between the involved parties.
   citations: ['SLA_MOU.json_embedding.json_1', 'SLA_MOU.json_embedding.json_17', 'SLA_MOU.json_embedding.json_7', 'SLA_MOU.json_embedding.json_0', 'SLA_MOU.json_e

[DEBUG] ids=['SLA_MOU.json_embedding.json_1', 'SLA_MOU.json_embedding.json_17', 'SLA_MOU.json_embedding.json_7', 'SLA_MOU.json_embedding.json_0', 'SLA_MOU.json_emb
[DEBUG ctx#1] dist=0.620 | sim≈0.380 | Describes the scope and boundaries of the agreement, including applicable systems or services.
[DEBUG ctx#2] dist=0.625 | sim≈0.375 | Defines key terminology and acronyms relevant to the Service Level Agreement.
[DEBUG ctx#3] dist=0.630 | sim≈0.370 | Summarizes the specific service requirements to be fulfilled under this agreement.
[DEBUG ctx#4] dist=0.793 | sim≈0.207 | Defines the overall purpose and intent of the agreement between the involved parties.
[DEBUG ctx#5] dist=0.883 | sim≈0.117 | Lists the specific roles involved and their corresponding duties under the agreement.
[DEBUG ctx#6] dist=0.899 | sim≈0.101 | Provides contextual information about the IT system or services covered by the agreement.
Q> Provide the Instructions for Changing the Contents of Drop-Down Menus

A> To change the contents of drop-down menus, follow these instructions:

1. Highlight the cell where you wish to change the content of the drop-down menu.
2. From the file menu, click 'Data' -> 'Validation'.
3. Change the content of the source field as needed.
   citations: ['Lessons_Learned_Log_embedded.json_1', 'SLA_MOU.json_embedding.json_13', 'SLA_MOU.json_embedding.json_16', 'EPLC_Acquisiti

[DEBUG] ids=['Lessons_Learned_Log_embedded.json_1', 'SLA_MOU.json_embedding.json_13', 'SLA_MOU.json_embedding.json_16', 'EPLC_Acquisition
[DEBUG ctx#1] dist=0.870 | sim≈0.130 | Highlight the cell where you wish to change the content of the drop-down menu. From the file menu
[DEBUG ctx#2] dist=1.359 | sim≈-0.359 | Details the procedure for modifying, reviewing, or updating the agreement.
[DEBUG ctx#3] dist=1.504 | sim≈-0.504 | Lists supporting documents, versions, descriptions, and locations referenced in the agreement.
[DEBUG ctx#4] dist=1.514 | sim≈-0.514 | [Insert the name, version number, description, and physical location of any documents referenced
[DEBUG ctx#5] dist=1.591 | sim≈-0.591 | <optional> <required> <required> <required>
[DEBUG ctx#6] dist=1.594 | sim≈-0.594 | [Discuss any Government information such as manuals, drawings, and test data to be provided, and
```

GenAI & Model Development

Q> What is the CONTACTS in SERVICE LEVEL AGREEMENT/MEMORANDUM OF UNDERSTANDING?

A> The context specifies that the Service Level Agreement includes "contact information for all primary stakeholders and parties to the agreement." However, the specific contact details are not provided in the context. Therefore, the answer is:

Not specified in the provided context.
citations: ['SLA_MOU.json_embedding.json_1', 'SLA_MOU.json_embedding.json_17', 'SLA_MOU.json_embedding.json_7', 'SLA_MOU.json_embedding.json_5', 'SLA_MOU.json_embedding.json_6', 'SLA_MOU.json_embedding.json_2']

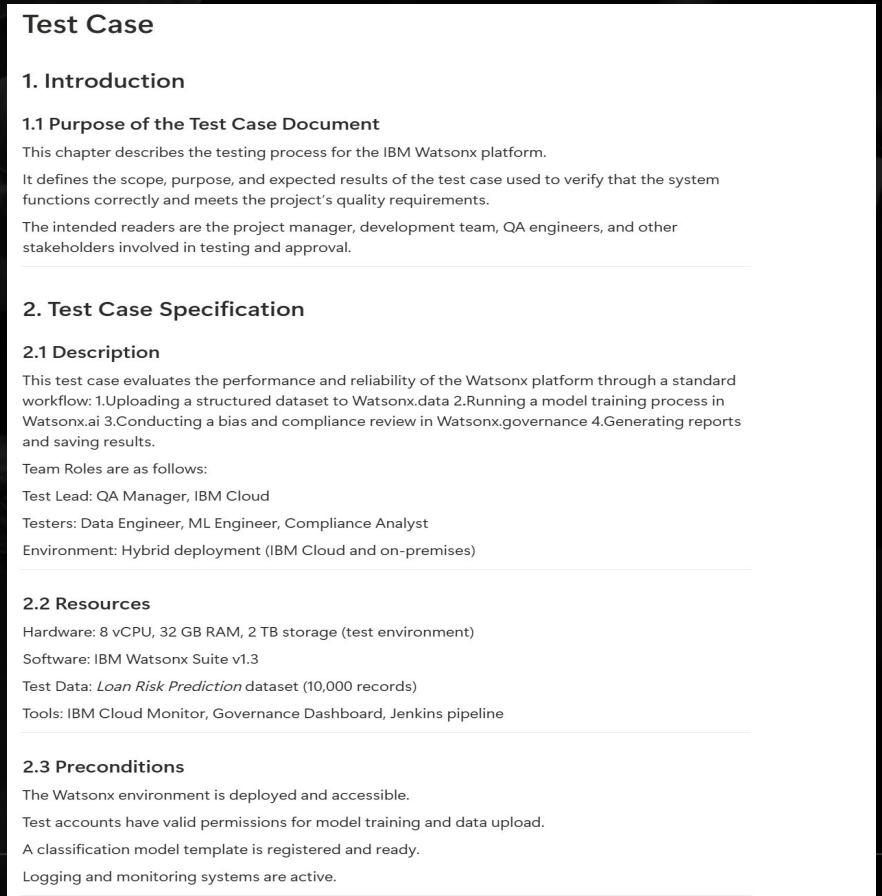
[DEBUG] ids=['SLA_MOU.json_embedding.json_1', 'SLA_MOU.json_embedding.json_17', 'SLA_MOU.json_embedding.json_7', 'SLA_MOU.json_embedding.json_5', 'SLA_MOU.json_embedding.json_6', 'SLA_MOU.json_embedding.json_2']
[DEBUG ctx#1] dist=0.585 | sim≈0.415 | Describes the scope and boundaries of the agreement, including applicable systems or services.
[DEBUG ctx#2] dist=0.593 | sim≈0.407 | Defines key terminology and acronyms relevant to the Service Level Agreement.
[DEBUG ctx#3] dist=0.615 | sim≈0.385 | Summarizes the specific service requirements to be fulfilled under this agreement.
[DEBUG ctx#4] dist=0.734 | sim≈0.266 | Lists the specific roles involved and their corresponding duties under the agreement.
[DEBUG ctx#5] dist=0.764 | sim≈0.236 | Provides contact information for all primary stakeholders and parties to the agreement.
[DEBUG ctx#6] dist=0.779 | sim≈0.221 | Provides contextual information about the IT system or services covered by the agreement.

```
SYSTEM_PROMPT = (
    "You are an EPLC assistant. Answer only using the information in the CONTEXT. "
    "If the answer can be inferred from the context, explain it briefly. "
    "If the context provides no relevant information, reply exactly: Not specified in the provided context."
)
```

User Experience & Application Development Team

11/6 Done

1. Confirmed with the team and finalized the decision for choosing streamlit —>Streamlit (accessibility and builtin python integration)
2. Data collection for use case→Done(image on the right)
3. Finalize design→Done(Added “How to Use” and “Introduction” page)



Test Case

1. Introduction

1.1 Purpose of the Test Case Document

This chapter describes the testing process for the IBM Watsonx platform.

It defines the scope, purpose, and expected results of the test case used to verify that the system functions correctly and meets the project’s quality requirements.

The intended readers are the project manager, development team, QA engineers, and other stakeholders involved in testing and approval.

2. Test Case Specification

2.1 Description

This test case evaluates the performance and reliability of the Watsonx platform through a standard workflow: 1.Uploading a structured dataset to Watsonx.data 2.Running a model training process in Watsonx.ai 3.Conducting a bias and compliance review in Watsonx.governance 4.Generating reports and saving results.

Team Roles are as follows:

Test Lead: QA Manager, IBM Cloud

Testers: Data Engineer, ML Engineer, Compliance Analyst

Environment: Hybrid deployment (IBM Cloud and on-premises)

2.2 Resources

Hardware: 8 vCPU, 32 GB RAM, 2 TB storage (test environment)

Software: IBM Watsonx Suite v1.3

Test Data: *Loan Risk Prediction* dataset (10,000 records)

Tools: IBM Cloud Monitor, Governance Dashboard, Jenkins pipeline

2.3 Preconditions

The Watsonx environment is deployed and accessible.

Test accounts have valid permissions for model training and data upload.

A classification model template is registered and ready.

Logging and monitoring systems are active.

EPLC Assistant

Empowering IT Project Managers with smarter, faster documentation.

Managing EPLC documents can be complex and time-consuming. EPLC Assistant helps automate this process — using generative AI to create, review, and refine key deliverables such as SLA/MOU, Training Plans, and O&M Manuals. Designed for government and enterprise projects, it combines IBM's trusted AI technology with a human-centered UX to ensure accuracy, compliance, and efficiency.

Start a Project

How to Use EPLC Assistant

Learn how to interact with your AI assistant to generate and understand EPLC materials effectively.

Try the Chatbot →

Step 1 — Ask Your Question

Type your EPLC-related question about an executive order in the input box.

Step 2 — Get Responses

The chatbot searches policy libraries and provides accurate, summarized answers.

Step 3 — Review and Save

Edit or export the response for your project.

Try Asking...

What is the EPLC Initial Phase?

Show me a CDC UP template for planning.

Explain the difference between initiation and planning phases.

💡 Tips for Best Results

- Be specific — mention the EPLC phase or document type you're referring to.
- Try rephrasing your question if the chatbot doesn't understand.
- You can always find official templates and policies linked below.

Project

New Project

Name your project

start

Project

New Project

IBM Watsonx

<uploaded document>

Training Plan

(O&M) Manual

Service Level Agreement / MOU

Test Case

Export all

IBM Watsonx

Training Plan

20%

Training Plan

20%

Training Plan

20%

Training Plan

20%

Let me know more about your project!

+

Project

New Project

IBM Watsonx

<uploaded document>

Training Plan

(O&M) Manual

Service Level Agreement / MOU

Test Case

Export all

IBM Watsonx > Training Plan

Session	Status
Session 2 Test Case Specification	×
2.1 Description	×

Preview Download

2.1 Description

xxxxxxxxxxxxxxxxxx.

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxx

Accept Edit Regenerate

\+

Describe the test case and the individuals involved in the testing. Include diagrams depicting the interaction between individuals and the different elements being tested.

Project

New Project

IBM Watsonx

<uploaded document>

Training Plan

(O&M) Manual

Service Level Agreement / MOU

Test Case

Export all

IBM Watsonx > Training Plan

Preview

CDC

<PROJECT NAME>

TEST PLAN

Version *<1.0>*

<mm/dd/yyyy>

Project

New Project

IBM Watsonx

<uploaded document>

Training Plan

(O&M) Manual

Service Level Agreement / MOU

Test Case

Export all

IBM Watsonx > Training Plan

Session	Status
Session 3 service level agreement/ MOU	×
3.1 Introduction	×
3.1.1 Purpose of Service legal agrreement/Memoradum	
3.1.2 Scope	
3.1.3 Background	
3.1.4 Audience	
3.1.5 Assumptions	
3.1.6 Roles and Responsibilities	
3.1.7 Contacts	
3.2 Service Details	
3.2.1 Requirements	
3.2.2 Service Level Expectations	
3.2.3 Escalation Actions	
3.2.4 Service Provider/ Services Recipient	
3.2.5 Service Hours for Problem Solution	
3.2.6 Performance Guarantee	
3.2.7 Agreement Change Process	

Preview Download

3.1 introduction

xxxxxxxxxxxxxxxxxx.

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxx

Accept Edit Regenerate

\+

Describe the test case and the individuals involved in the testing. Include diagrams depicting the interaction between individuals and the different elements being tested.

User Experience & Application Development Team

Now in progress:

1. Front-end Development(ddl:next Monday deliver MVP website to the data team, prepare for algorithm integration with the genAI team)
2. Finalize design (ddl:next Monday, keep improving the core features and user feedback within the team)
3. Keep other teams updated on feasibility (pdf fill in)

Step	Action	Expected Result
1	Upload dataset to Watsonx.data	File passes validation and uploads successfully
2	Start AutoML training	Training runs without error and progress is shown
3	Run bias and compliance check	System produces a report with bias metrics
4	Export model artifact	Export completes, model ID is generated
5	Archive reports	Report marked “Compliant” and stored for audit

Evaluation & Compliance Team

https://docs.google.com/document/d/1BRKGJcGlFmDYa5HiAxOdrogoXzCqA24CLPXS1THT5TI/edit?usp=sharing

Keep refining the survey question lists for the final product.

Questions

NEXT STEPS

Keep up with the timeline, communicate with the sub-team for better support. Diving into the topics we've touched on today to deepen our understanding on the project, and starting the data preparation procedure.

THANK YOU!

Columbia Practicum Core Team