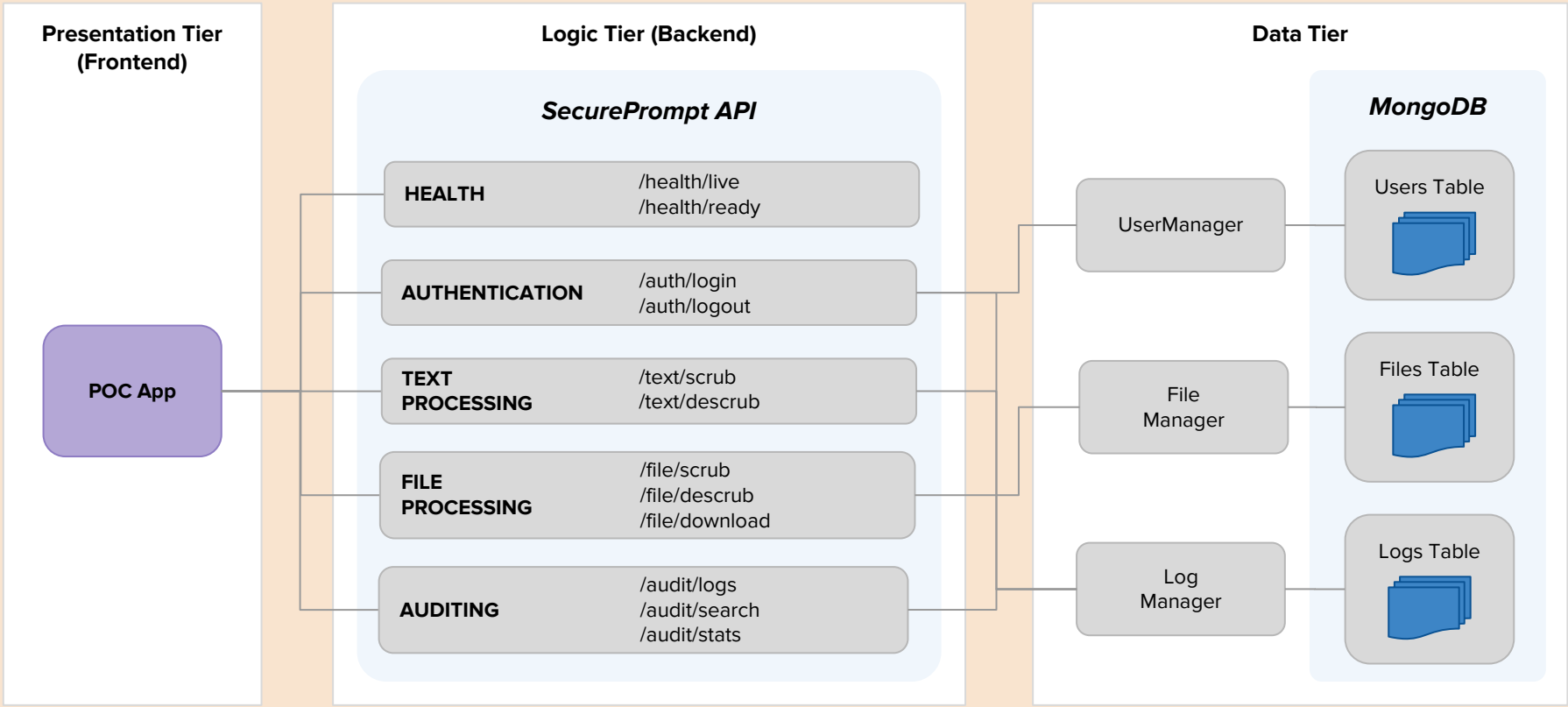
SecurePrompt

TEAM 3

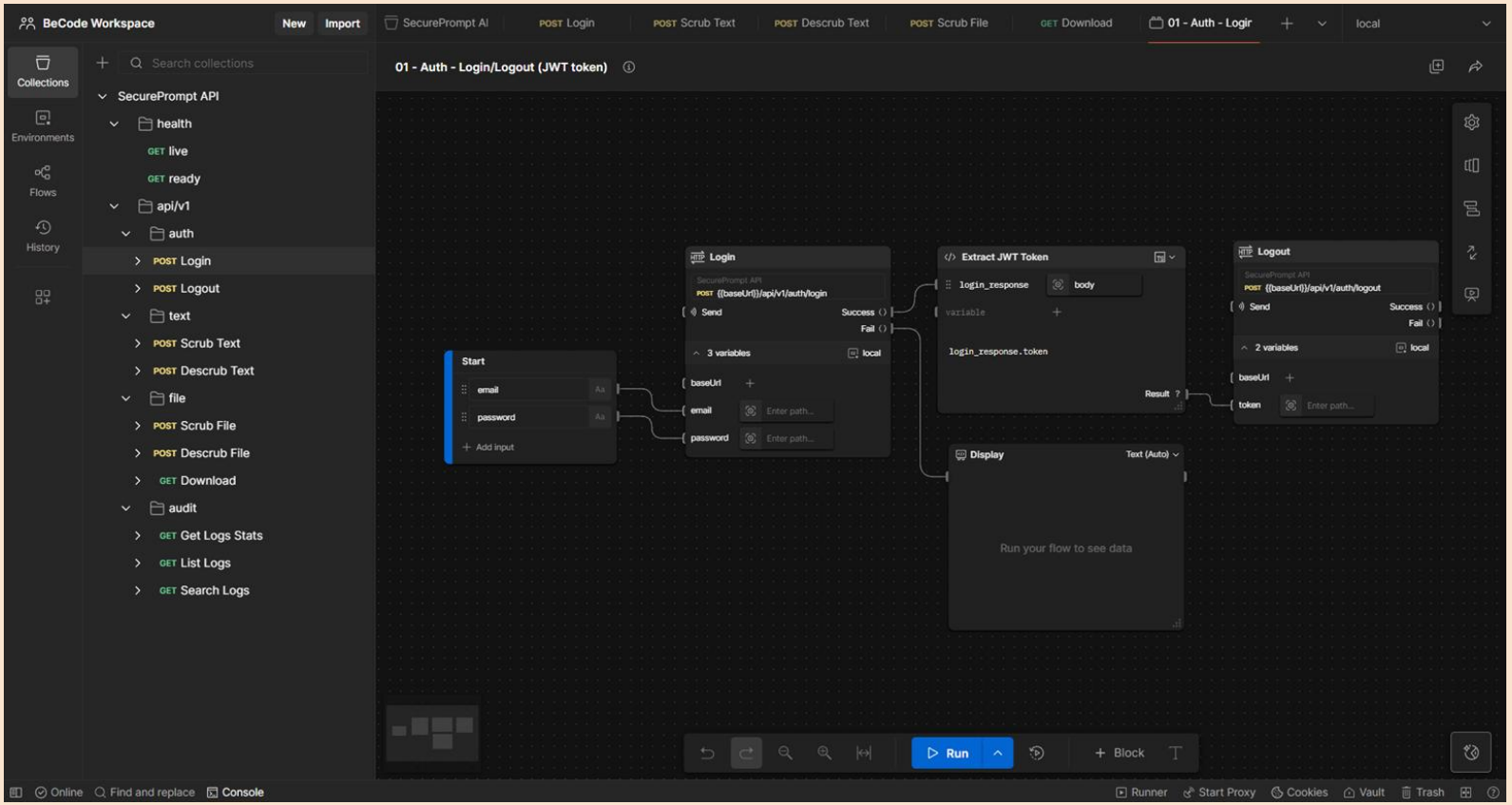
Estefania Sosa
Floriane Haulot
Preeti Duhan
Alberto Pérez

10.10.2025

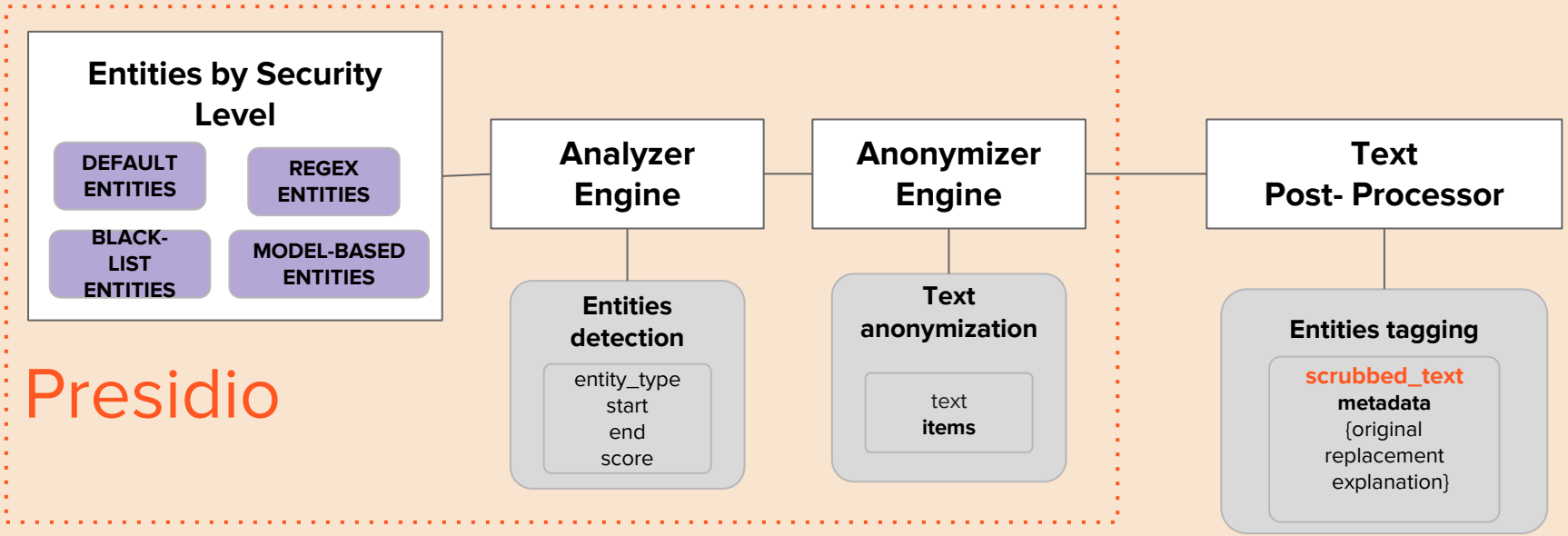
Multi-Tier Architecture



SecurePrompt API - Postman Flows

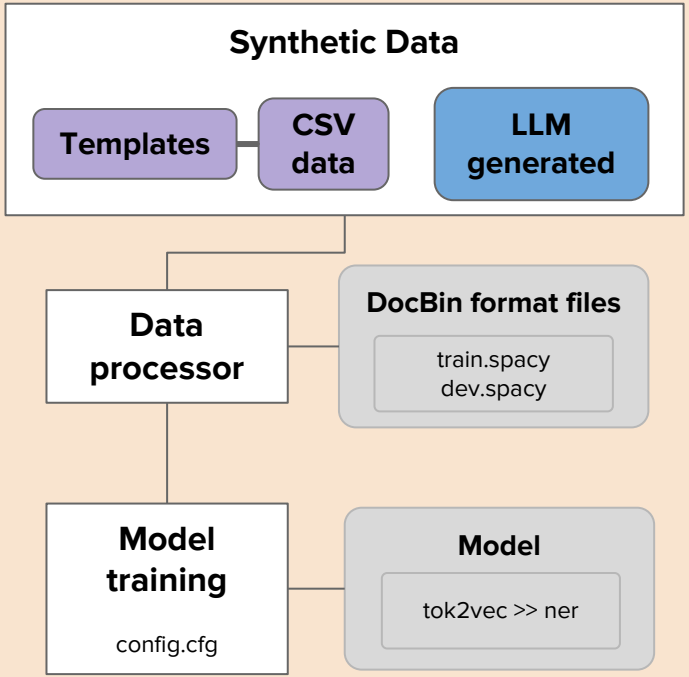


Scrubbing Strategy



NLP Model

Spacy-based models



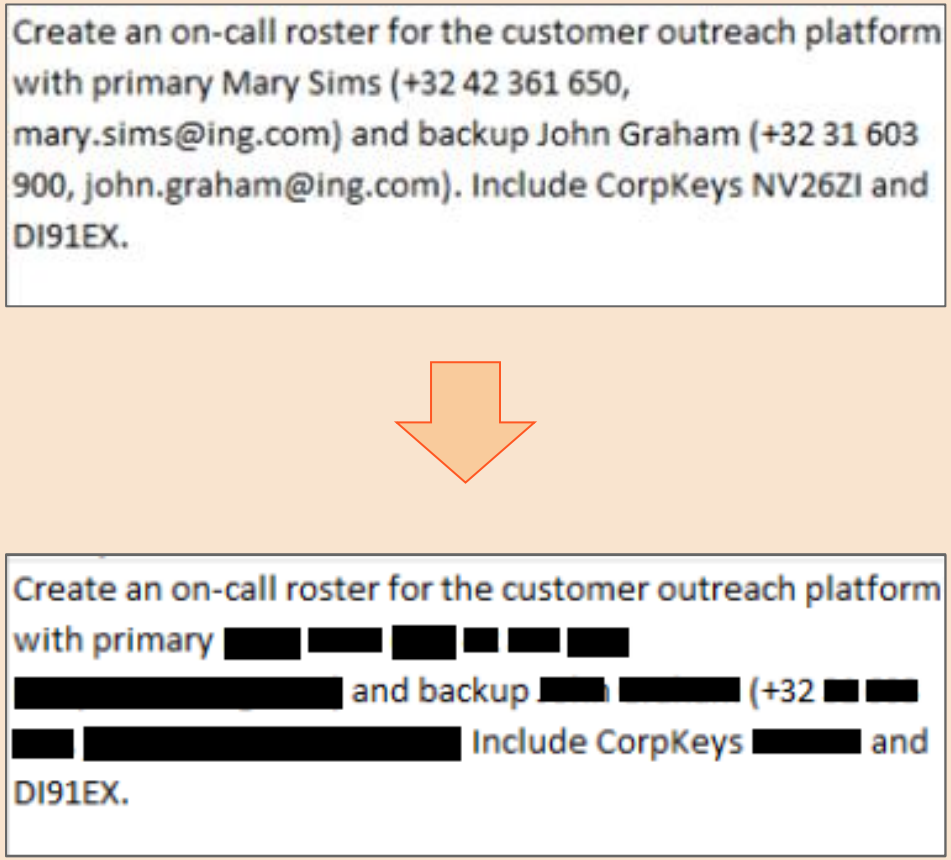

```
secureprompt/backend/scrubbers/nlp on  anonymizer-presidio [⇡] via 🐍 v3.10.9 (venv)
❯ python -m spacy train config.cfg --paths.train ./datafiles/spacy_data_c4/train_v6.spacy
--paths.dev ./datafiles/spacy_data_c4/dev_v6.spacy --output ./models_c4/model_c4_v6
✔ Created output directory: models_c4/model_c4_v6
ℹ Saving to output directory: models_c4/model_c4_v6
ℹ Using CPU

=========================== Initializing pipeline ===========================
✔ Initialized pipeline

============================= Training pipeline =============================
ℹ Pipeline: ['tok2vec', 'ner']
ℹ Initial learn rate: 0.001
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00      8.32    0.00    0.00    0.00    0.00
  0     200         12.14    468.44   96.68   99.03   94.44    0.97
  2     400        144.31    103.10   99.07   99.07   99.07    0.99
  3     600         33.92     83.49  100.00  100.00  100.00    1.00
  5     800          3.36     21.60   99.53  100.00   99.07    1.00
  7    1000         67.11    157.34  100.00  100.00  100.00    1.00
✔ Saved pipeline to output directory
models_c4/model_c4_v6/model-last
```

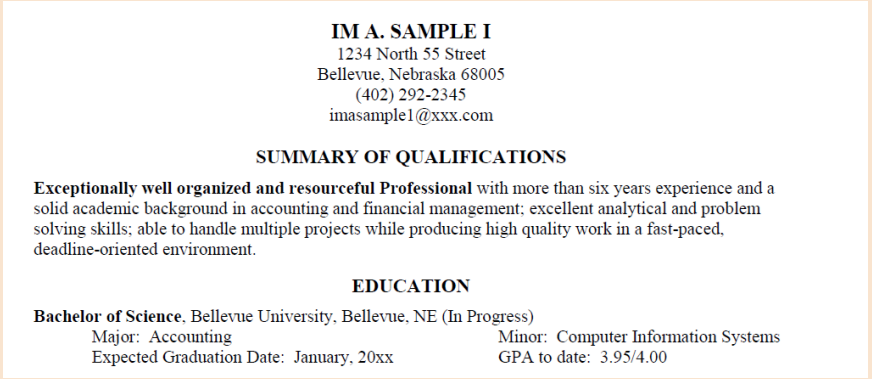
Screenshot Scrub

- OpenCV
- Pytesseract
- Presidio



PDF Scrubbers

- **`pdf2image`**: Converts PDF pages into images.
- **`pytesseract`**: Python wrapper for Tesseract OCR to extract text from PDF images.
- **`anonympy.pdf.pdfAnonymizer`**: Core library that handles PDF anonymization (blurring/redacting text).



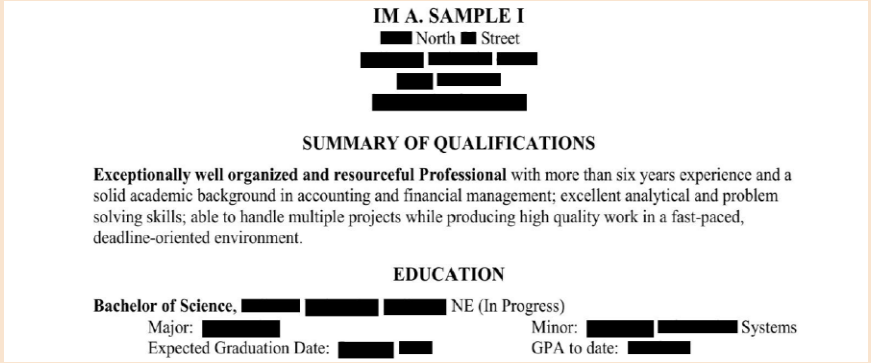
IM A. SAMPLE I
1234 North 55 Street
Bellevue, Nebraska 68005
(402) 292-2345
imasample1@xxx.com

SUMMARY OF QUALIFICATIONS

Exceptionally well organized and resourceful Professional with more than six years experience and a solid academic background in accounting and financial management; excellent analytical and problem solving skills; able to handle multiple projects while producing high quality work in a fast-paced, deadline-oriented environment.

EDUCATION

Bachelor of Science, Bellevue University, Bellevue, NE (In Progress)
Major: Accounting — Minor: Computer Information Systems
Expected Graduation Date: January, 20xx — GPA to date: 3.95/4.00



IM A. SAMPLE I
[redacted] North [redacted] Street
[redacted]

SUMMARY OF QUALIFICATIONS

Exceptionally well organized and resourceful Professional with more than six years experience and a solid academic background in accounting and financial management; excellent analytical and problem solving skills; able to handle multiple projects while producing high quality work in a fast-paced, deadline-oriented environment.

EDUCATION

Bachelor of Science, [redacted] NE (In Progress)
Major: [redacted] — Minor: [redacted] Systems
Expected Graduation Date: [redacted] — GPA to date: [redacted]

Tests results

Testing context :

- 906 new prompts
- Security focused
- Labelling issues
- Structural comparison
- Managing over-detection : better safe than sorry!

```
METRICS SUMMARY
============================================================
SecurePrompt Security Performance
Analysis Date: October 09, 2025

KEY METRICS:
   Security Score: 73.2%
   Structural Accuracy: 35.0%
   Better Safe Than Sorry: 26.7% cautious over-detection
   Custom Models: Integrated successfully

SECURITY LEVEL PERFORMANCE:
   GOOD: C1 (Public Data): 36.3%
   GOOD: C2 (Internal Operations): 66.0%
   BEST: C3 (Customer Data): 99.7%
   GOOD: C4 (Sensitive Data): 90.7%

SECURITY PRINCIPLE: Over-detection = 90% score
   • 26.7% of cases show cautious behavior
   • Only 24.2% cases missed data (needs improvement)

Security-aware analysis complete!
73.2%
```

LLM implementation : RAG and MongoDB


```
{
    "question": "Identity verification: Linda Jones correctly provided PIN
    "top_k": 2,
    "retrieved": [
        {
            "chunk_index": 8,
            "distance": 281.82427978515625,
            "source": "17_c4_customers_auth_data",
            "security": "c4"
        },
        {
            "chunk_index": 19,
            "distance": 321.084228515625,
            "source": "18_c4_customers_s_data",
            "security": "c4"
        }
    ],
    "answer": "Based on the given test data, Linda Jones has provided her
        CVV of 464 does not match with any of the credit card details in t
        proceeding with the transaction."
}
```


1. Load Data to MongoDB
2. Add security level tag
3. API Route : used model

   - Gemini 2.5

   - Ollama Dolphin Mistral 7B

   heavy but permissive model

Next steps

- ➔ Connect LLM to login and security access
- ➔ Improve NLP models
- ➔ Improve prompts testing
- ➔ Expand covered entities
- ➔ Expand file scrubbing to cover more formats
- ➔ Improve audit logging
- ➔ Implement file de-scrubbing
- ➔ Implement metrics dashboard

Thank you