

Automated Scraping of PDF School Transcripts for Surveys

April 2025

Emily Hadley
Stuart Allen
John Bollenbacher
Dale Holstein
Andy Kawabata
Michael Long
Ethan Ritchie
Michael Wenger





RTI TranscriptGenie™

Agenda

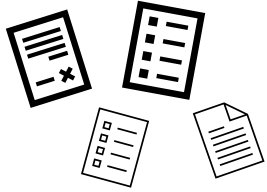
- Introduction to the challenge
- Overview of RTI TranscriptGenie
- Graph Database
- GenAI and Secure Environment
- Questions

School Transcripts in Surveys and Research

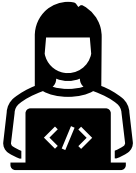
East High School Official Transcript					
123 Mountain Road, Dead River, Maine 04000				Student Personal Information Date of Birth: _____ Parent/Guardian: _____	
Phone: _____ Website: _____				Date of Enrollment: _____ Date of Graduation: _____ Address: _____ Contact Info: _____	
Learning Experience	Proficiency Level	Duration	Type		
2009-2010					
English 9	3.5	Year	Honors		
History 9	3.0	Year	Course		
Geometry	3.0	Year	Course		
Spanish I	3.5	Year	Course		
Earth Science	3.0	Year	Course		
Art I	3.0	Year	Course		
2010-2011					
English 10	3.5	Year	Course		
History 10	4.0	Year	Honors		
Algebra II	4.0	Year	Course		
Spanish II	3.5	Year	Course		
Chemistry	4.0	Year	Honors		
Drama	4.0	Semester	Course		
2011-2012					
English Language and Composition (AP)	3.5	Year	Honors		
US History	3.0	Year	Independent		
Calculus	3.5	Year	Course		
Spanish III	3.5	Year	Course		
Physics	3.5	Year	Course		
Software Development, Inc.	3.5	Year	Internship		
2012-2013					
English 101	4.0	Semester	Dual-Enrollment		
Mountain Biotech Industries (STEM)	3.5	Year	Internship		
Physical Education	4.0	Semester	Course		
Statistics (AP)	3.8	Year	Course		
Health	4.0	Semester	Course		
Psychology (AP)	4.0	Semester	Honors		
Graduation Standards: Cross-Curricular Skills		Proficiency Level			
Maine Learning Results	Clear and effective communicator	3.0			
Guiding Principles	Self-directed and lifelong learner	3.25			
	Creative and practical problem solver	3.50			
	Responsible and involved citizen	3.75			
	Integrative and informed thinker	4.0			
Academic Summary					
GPA: 3.75					
Magna Cum Laude					
SAT/ACT Scores					
Academic Awards					
Academic Achievements					
Title of Capstone Project					
Grading System					
3.6-4.0: Exceeds Proficiency					
3.0-3.5: Proficient					
2.0-2.9: Partially Proficient					
1.0-1.9: Insufficient Evidence					
Graduation Requirements					
Demonstrated Proficiency					
Maine Guiding Principles					
English Language Arts					
Mathematics					
Science and Technology					
Social Studies					
Health Education and Physical Education					
World Languages					
Visual and Performing Arts					
Career Education and Development					
Additional Graduation Requirements					
Capstone Project					
Service Learning Requirement					

- Part of federal education survey data collection since the 1980s
- Both secondary and postsecondary transcripts
- Considered to be more accurate than student self-reporting
- Can be used to study:
 - Course-taking patterns
 - Student momentum and attrition
 - Credit transfer
 - Connection among courses, occupations, and wages

The Challenge



Wide variety of school transcript layouts and data types



30 – 45+ minutes to manually key data with legacy software



Complicated data structure makes for difficult validation

\$\$\$ and

But what if we could...

- Remove need for manual keying through automated PDF text extraction
- Improve database architecture
- Support enhanced data validation processes
- Still include humans in the loop

And cut costs!

Application Scope

- Frontend
 - Login and view transcripts
 - Annotate transcripts
 - Review automated annotations
- Backend
 - Split PDFs
 - Scrape data
 - Create annotation suggestions
- Database
 - Store data

Architecture

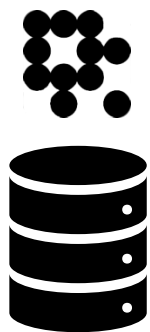
Frontend



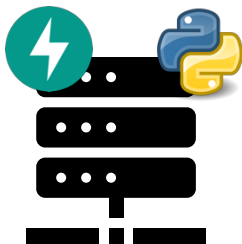
Transcript
Dashboard



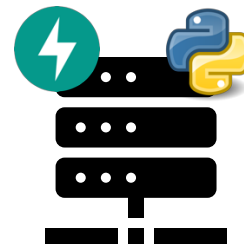
Vector Database



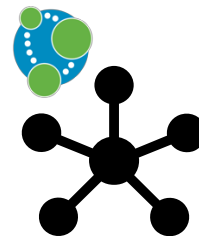
Backend



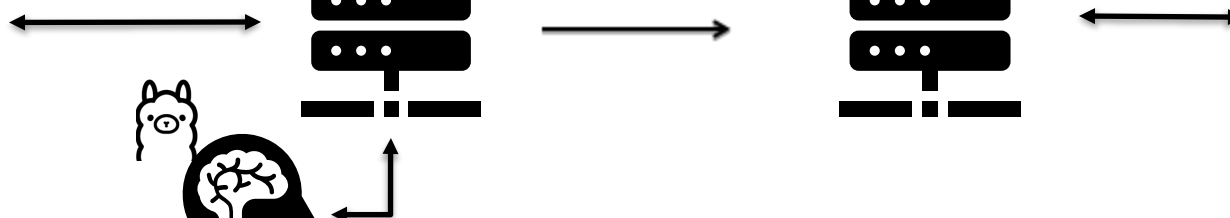
Database API



Graph Database



LLM



PDF Scrapping

Smokey Bear
UNM ID: 123-45-6789
DATE OF BIRTH: 21-DEC-1989

THE UNIVERSITY OF NEW MEXICO
OFFICE OF THE REGISTRAR
ALBUQUERQUE, NEW MEXICO 87131-0001

PAGE: 1
DATE ISSUED: 10-OCT-2012

Course Level: Undergraduate

Current Program
BS in Natural Resources
Program :BS Natural Resources
College , University College Campus ,
Albuquerque/Main
Major :Natural Resources

SUBJ NO.	COURSE TITLE	CRED GRD	PTS R
----------	--------------	----------	-------

INSTITUTION CREDIT:

Term: Fall 2011
University College
BIOL 237 Hum Anat & Physic I Hlth Sci 3.00 B+ 9.99
BIOL 247L Anatomy & Physiology Lab 1.00 A 4.00
MUS 172 Jazz History 3.00 B+ 9.99
NURS 238 Pharmin Nurs & Hlth Prof 3.00 C+ 6.99
soc 101 Introductio Sociology 3.00 B 9.00
Term: Ehrs: 13.00 GPA-Hrs: 13.00 QPts: 39.97 GPA: 3.07

Term: Spring 2012
University College
ANTH 130 Cultures of the World 3.00 A- 11.01
BIOL 238 Hum Anat & Physic II Hlth Sci 3.00 A- 11.01
BIOL 248L Anatomy & Physiology Lab II 1.00 A- 3.67
CLST 333 T: Big Screen Rome 3.00 A+ 12.99
PSY 231 Psychology of Human Sexuality 3.00 B 9.00
Term: Ehrs: 13.00 GPA-Hrs: 13.00 QPts: 47.68 GPA: 3.66

Dean's List

*****TRANSCRIPT TOTALS*****				
	Earned Hrs	GPA Hrs	Points	GPA
TOTAL INSTITUTION	83.00	83.00	258.65	3.11
TOTAL TRANSFER	0.00	0.00	0.00	0.00
OVERALL	83.00	83.00	258.65	3.11

*****END OF TRANSCRIPT*****

SUBJ NO.	COURSE TITLE	CRED GRD	PTS R
----------	--------------	----------	-------

pdfquery

Smokey Bear
UNM ID: 123-45-6789
DATE OF BIRTH: 21-DEC-1989

THE UNIVERSITY OF NEW MEXICO
OFFICE OF THE REGISTRAR
ALBUQUERQUE, NEW MEXICO 87131-0001

PAGE: 1
DATE ISSUED: 10-OCT-2012

Course Level: Undergraduate

Current Program
BS in Natural Resources
Program :BS Natural Resources
College , University College Campus ,
Albuquerque/Main
Major :Natural Resources

SUBJ NO.	COURSE TITLE	CRED GRD	PTS R
----------	--------------	----------	-------

INSTITUTION CREDIT:

Term:
University College
BIOL 237 Hum Anat & Physic I Hlth Sci 3.00 B+ 9.99
BIOL 247L Anatomy & Physiology Lab 1.00 A 4.00
MUS 172 Jazz History 3.00 B+ 9.99
NURS 238 Pharmin Nurs & Hlth Prof 3.00 C+ 6.99
soc 101 Introductio Sociology 3.00 B 9.00
Term: Ehrs: 13.00 GPA-Hrs: 13.00 QPts: 39.97 GPA: 3.07

Term: Spring 2012
University College
ANTH 130 Cultures of the World 3.00 A- 11.01
BIOL 238 Hum Anat & Physic II Hlth Sci 3.00 A- 11.01
BIOL 248L Anatomy & Physiology Lab II 1.00 A- 3.67
CLST 333 T: Big Screen Rome 3.00 A+ 12.99
PSY 231 Psychology of Human Sexuality 3.00 B 9.00
Term: Ehrs: 13.00 GPA-Hrs: 13.00 QPts: 47.68 GPA: 3.66

Dean's List

*****TRANSCRIPT TOTALS*****				
	Earned Hrs	GPA Hrs	Points	GPA
TOTAL INSTITUTION	83.00	83.00	258.65	3.11
TOTAL TRANSFER	0.00	0.00	0.00	0.00
OVERALL	83.00	83.00	258.65	3.11

*****END OF TRANSCRIPT*****

SUBJ NO.	COURSE TITLE	CRED GRD	PTS R
----------	--------------	----------	-------

Text Extraction

Original PDF

Smokey Bear
UNM ID: 123-45-6789
DATE OF BIRTH: 21-DEC-1989

THE UNIVERSITY OF NEW MEXICO
OFFICE OF THE REGISTRAR
ALBUQUERQUE, NEW MEXICO 87131-0001

PAGE: 1
DATE ISSUED: 10-OCT-2012

Course Level: Undergraduate

Current Program
BS in Natural Resources
Program : BS Natural Resources
College , University College Campus ,
Albuquerque/Main
Major : Natural Resources

SUBJ NO.	COURSE TITLE	CRED GRD	PTS R
----------	--------------	----------	-------

INSTITUTION CREDIT:

Term: Fall Semester

University College				
BIOL 237	Hum Anat & Physic I	Hlth Sci	3.00 B+	9.99
BIOL 247L	Anatomy & Physiology	Lab	1.00 A	4.00
MUS 172	Jazz History		3.00 B+	9.99
NURS 238	Pharmin Nurs & Hlth Prof		3.00 C+	6.99
soc 101	Introductio Sociology		3.00 B	9.00
Term:	Ehrs: 13.00	GPA-Hrs: 13.00	QPts: 39.97	GPA: 3.07

Term: Spring 2012

University College				
ANTH 130	Cultures of the World		3.00 A-	11.01
BIOL 238	Hum Anat & Physic II	Hlth Sci	3.00 A-	11.01
BIOL 248L	Anatomy & Physiology	Lab II	1.00 A-	3.67
CLST 333	T: Big Screen Rome		3.00 A+	12.99
PSY 231	Psychology of Human Sexuality		3.00 B	9.00
Term:	Ehrs: 13.00	GPA-Hrs: 13.00	QPts: 47.68	GPA: 3.66

Dean's List

TRANSCRIPT TOTALS				
	Earned Hrs	GPA Hrs	Points	GPA
TOTAL INSTITUTION	83.00	83.00	258.65	3.11
TOTAL TRANSFER	0.00	0.00	0.00	0.00
OVERALL	83.00	83.00	258.65	3.11

Extracted

Smokey Bear

UNM ID: 123-45-6789

DATE OF BIRTH: 21-DEC-1989

University College 2012 Spring Semester

ANTH 130 Cultures of the World 3.00 A- 11.01

BIOL 238 Human Anat & Physic II Hlth Sci
3.00 A- 11.01

BIOL 248L Anatomy & Physiology Lab II 1.00
A- 3.67

CLST 333 T: Big Screen Rome 3.00 A+ 12.99

PSY 231 Psychology of Human Sexuality 3.00
B 9.00

LLM Cleanup

UniversityCollege 2012 Spring Semester
ANTH 130 Cultures of the World 3.00 A- 11.01
BIOL 238 Human Anat & Physic II Hlth Sci 3.00 A- 11.01
BIOL 248L Anatomy & Physiology Lab II 1.00 A- 3.67
CLST 333 T: Big Screen Rome 3.00 A+ 12.99
PSY 231 Psychology of Human Sexuality 3.00 B 9.00



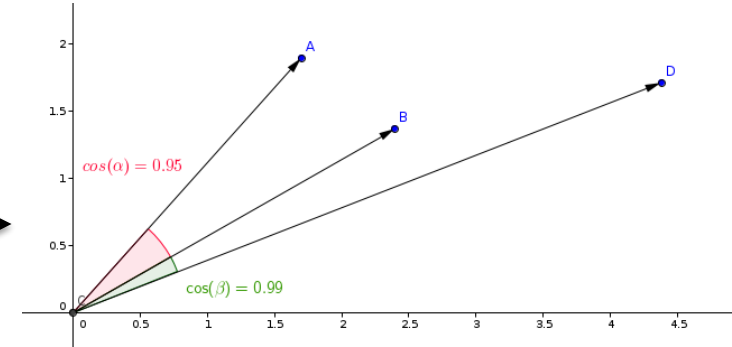
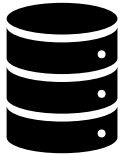
```
{'courses': [{ 'courseCode': 'ANTH 130',  
                'courseTitle': 'ANTH 130 Cultures of  
the World',  
                'earned': 11.01,  
                'grade': 'A-',  
                'hours': 3.0},  
               { 'courseCode': 'BIOL 238',  
                'courseTitle': 'BIOL 238 Human  
Anat & Physic II Hlth Sci',  
                'earned': 11.01,  
                'grade': 'A-',  
                'hours': 3.0},  
               ...],  
 'semester': 'UniversityCollege 2012 Spring  
Semester'}
```

Course Matching

Search query: "History"

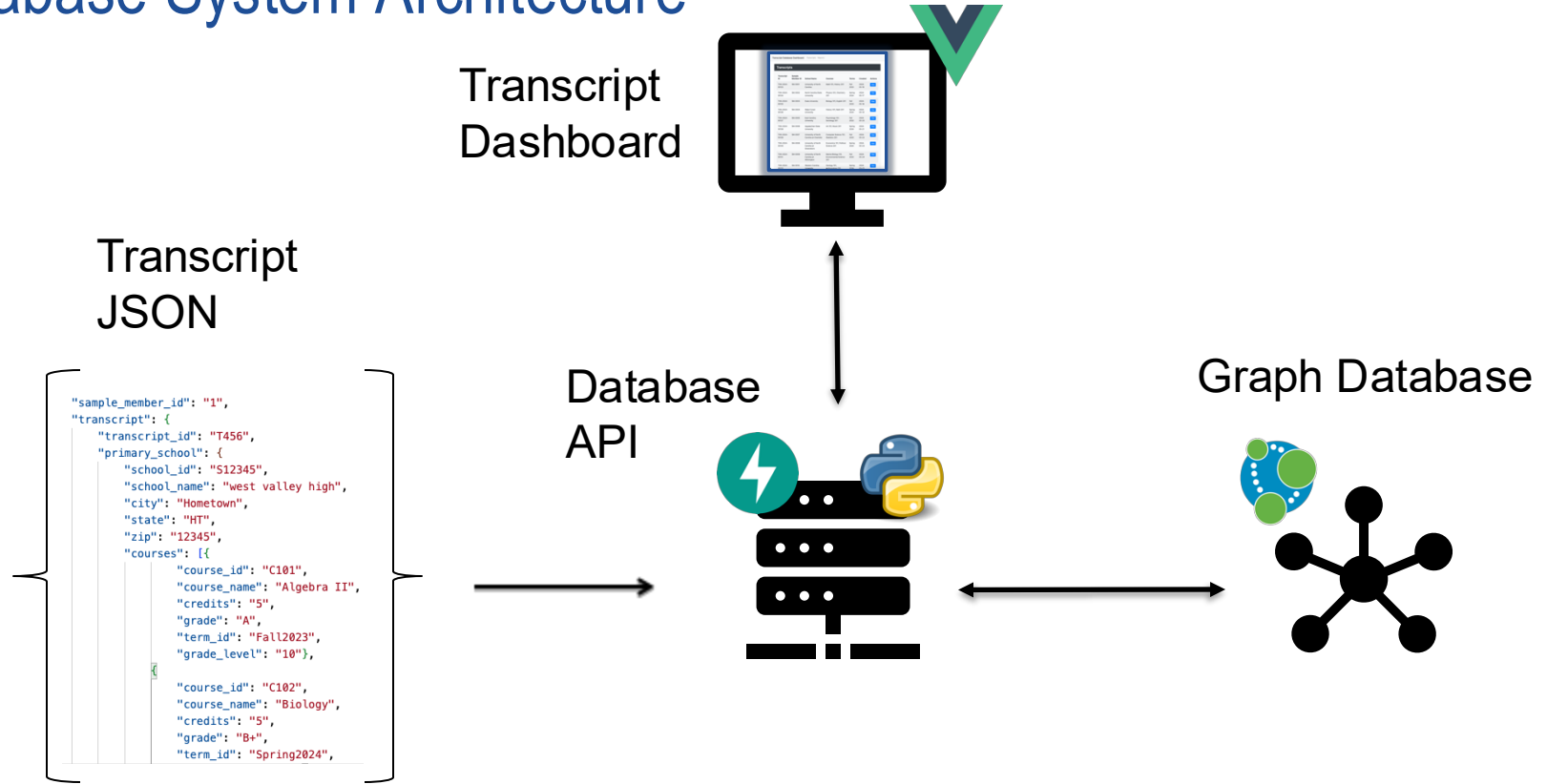
Embeddings Model

$[-0.00813760906, 0.00959756970,$
 $\dots, -0.00207111388, 0.00955703855]$



1. text: World History, score: 0.84732103
2. text: US History, score: 0.7855702
3. text: Economics, score: 0.36231852
4. text: English 1, score: 0.35061562
5. text: English 2, score: 0.34172332

Database System Architecture

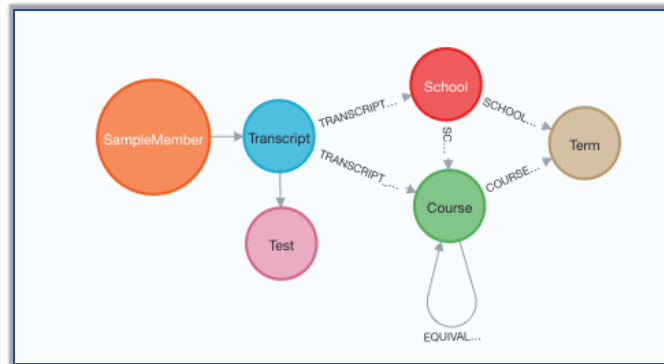


Graph Database Overview

- Single location for all saved transcript data
 - Reduced sprawl and duplication
- Users can access through dashboard
 - List, search, query, and update transcript info
 - Transcripts linked to original pdf
- Why Graph DB?
 - Optimized for interconnected data
 - Intuitive modeling and querying

```
MATCH (t:Transcript)-[HAS_SCHOOL]→(s:School {id: '123'})  
RETURN t
```

Find all transcripts that contain a specific school – Cypher Query



Graph DB schema

```
SELECT t.*  
FROM transcript t  
JOIN transcript_school ts ON t.id = ts.id  
JOIN school s ON ts.id = s.id  
WHERE s.id = '123';
```

Equivalent SQL query

Graph Database Overview

- Dashboard example

Transcript Database Dashboard

Transcripts

Reports

Transcripts

Transcript ID	Sample Member ID	School Name	Courses	Terms	Created	Actions
TXN-2024-00123	SM-0001	University of North Carolina	Math 101, History 201	Fall 2023	2024-05-16	View
TXN-2024-00124	SM-0002	North Carolina State University	Physics 101, Chemistry 201	Spring 2024	2024-05-17	View
TXN-2024-00125	SM-0003	Duke University	Biology 101, English 201	Fall 2023	2024-05-18	View
TXN-2024-00126	SM-0004	Wake Forest University	History 101, Math 201	Spring 2024	2024-05-19	View
TXN-2024-00127	SM-0005	East Carolina University	Psychology 101, Sociology 201	Fall 2023	2024-05-20	View
TXN-2024-00128	SM-0006	Appalachian State University	Art 101, Music 201	Spring 2024	2024-05-21	View
TXN-2024-00129	SM-0007	University of North Carolina at Charlotte	Computer Science 101, Statistics 201	Fall 2023	2024-05-22	View
TXN-2024-00130	SM-0008	University of North Carolina at Greensboro	Economics 101, Political Science 201	Spring 2024	2024-05-23	View
TXN-2024-00131	SM-0009	University of North Carolina at Wilmington	Marine Biology 101, Environmental Science 201	Fall 2023	2024-05-24	View
TXN-2024-00132	SM-0010	Western Carolina University	Geology 101, Anthropology 201	Spring 2024	2024-05-25	View

Transcript Database Dashboard

Transcripts

Reports

Update Transcript: TXN-2024-00456

School

Name: THE UNIVERSITY OF NEW MEXICO

Address: 1 University of New Mexico

City: Albuquerque

State: New Mexico

ZIP: 87131

Update School

Original PDF

Current Program

BS in Natural Resources

Program : BS Natural Resources

College : University College Campus , Albuquerque/Main

Major : Natural Resources

COURSE NO.	COURSE TITLE	CRED	GRD	PTS	R
INSTITUTION CREDIT:					
Term: Fall 2011					
University/College:					
BIO1037	Hum/Env & Physic (H&E) Sci	3.00 B+	9.99		
BIO1247	Anatomy & Physiology Lab	1.00 A	4.00		
MUS 172	Jazz History	3.00 B+	9.99		
NURS 288	Pharmac Nurs & H&H Prof	3.00 C+	6.99		
soc 101	Introduction to Sociology	3.00 B	9.00		
Term: Ehrs: 13.00 GPA-Hrs: 13.00 QPts: 39.97 GPA: 3.07					
Spring 2012					
University/College:					
ANTH 130	Cultures of the World	3.00 A-	11.01		
BIO1238	Hum Anat & Physic II: H&H Sci	3.00 A-	11.01		
BIO1248L	Anatomy & Physiology Lab II	1.00 A-	3.67		
CLST 333	T: Big Screen Rome	3.00 A+	12.99		
PSY 231	Psychology of Human Sexuality	3.00 B	9.00		
Term: Ehrs: 13.00 GPA-Hrs: 13.00 QPts: 47.68 GPA: 3.66					
Dean's List					
*****TRANSCRIPT TOTALS*****					
TOTAL INSTITUTION					
Ehrs: 83.00 GPA-Hrs: 83.00 QPts: 256.65 GPA: 3.11					
TOTAL TRANSFER					
Ehrs: 0.00 GPA-Hrs: 0.00 QPts: 0.00 GPA: 0.00					
OVERALL					
Ehrs: 83.00 GPA-Hrs: 83.00 QPts: 256.65 GPA: 3.11					
*****END OF TRANSCRIPT*****					

Courses

Course 1: Math 101 **A**

Update

Course 2: History 201 **B**

Update

Course 3: Science 301 **C**

Update

Tests

Test 1: SAT **1200**

Update

Test 2: ACT **28**

Update

Test 3: GRE **320**

Update

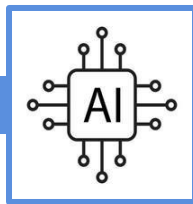
Generative AI Helps Standardize Messy PDF Data

Problem



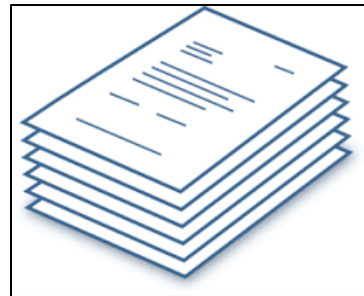
Tables in transcripts
vary in structure

Solution



Secure & Offline
AI Language Model

Result

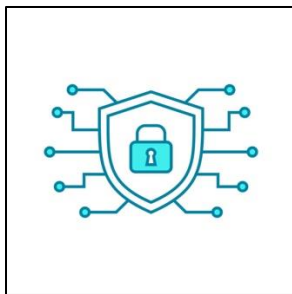


Clean, Standardized
Transcript Tables

Roadmap to Secure Hosting & FERPA Compliance



Create Detailed
Systems Diagram &
Logging Plan



Security &
Compliance
Review



Securely Deploy
App to FIPS Mod
Environment

Tools for Secure Development



Next Steps

- Continue to explore client use cases and needs
- Evaluate other LLMs
- Prepare training and resources



Thank you

Contact: Emily Hadley | email: ehadley@rti.org