WIZER PFIZER

Connecting People To Knowledge Faster


Anne Castille Buisson


Hayden Ratliff

- **Faculty Advisor:** Yu Ma
- **Company Advisors:** Kim Adler, Abby Freeman

Project Overview

Problem Statement

At Pfizer, it is always a priority to **reduce the time required to get drugs to market.**

A key opportunity to improve time to market is "**Knowledge Transfer**," where relevant R&D and acquired documents are transitioned to the manufacturing department.

9 months per molecule to classify documents

Not possible for Q&A with document contents

Data

33.6k	7	10	20	633	1.6k
Documents	Molecules	File Formats	Q&A Pairs	Nodes	Relationships

Product 1: Identifying Relevant Documents

Methods

Documents	Text Extraction	Embedding	Classification	Relevance
• Confidential data • Unbalanced labels • Clinical, batches, investigations	• Computer Vision w/ Optical Character Recognition • Limit first 2 pages	• Open-source LLM: BioMed RoBERTa • Transfer Learning • Semantic meaning	• HistXGBoost • Weighted loss • Custom metric: recall & AUC	• Classification probability • Threshold = 0.5

Results

Performance

83.5%	0.845	71%
Recall	AUC	Filtering

→ Exceeds 80% target

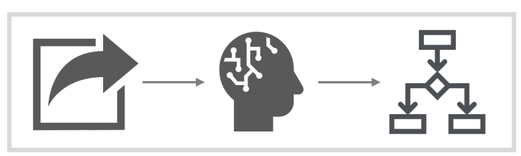
Processing Time per Molecule (>1k Docs)

9 Months Manual Approach → **10 Minutes** Automatic Approach

Implementation

Document Classification Workflow

AWS DB → Cloud Compute (aws) → Graph DB (neo4j) → Wizer UI

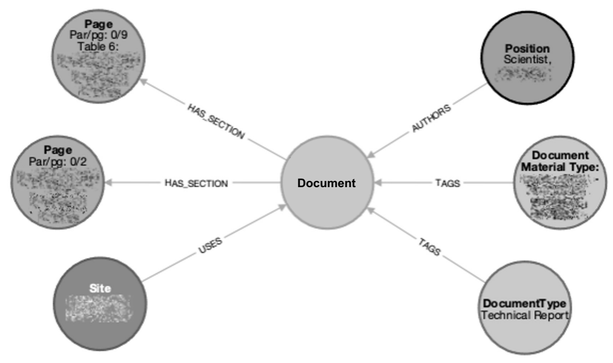


Document Classification Pipeline

Performance and Specifications

- Processes each document in <1 second
- Workflow runs in parallel with 8 AWS EC2 nodes
- Results written to Knowledge Graph with ETL script

Knowledge Graph Database: Neo4j



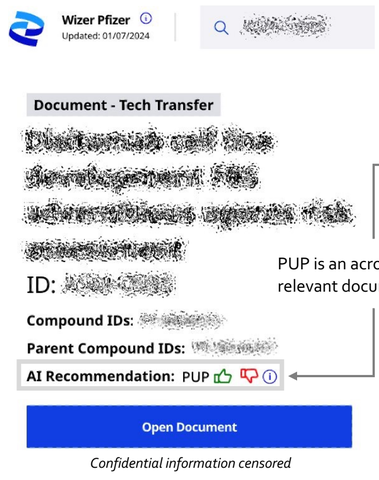
Node properties

Document [redacted]

author	[redacted]
compound_ids	[redacted]
document_id	[redacted]
document_title	[redacted]
pup_probability	0.640241417228176
pup_probability_updated_on	"2024-06-24T20:52:25.285000000Z"

Confidential information censored

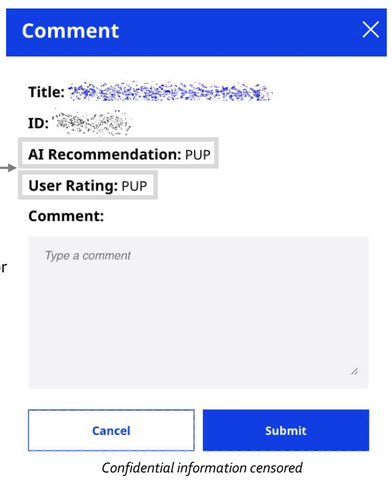
Wizer Pfizer: User Interface



PUP is an acronym for relevant documents

Confidential information censored

Human-In -The-Loop



Confidential information censored

Product Adoption

Q2 2024	Q4 2024	Q1 2025
3 Molecules	10 Molecules	All New Molecules (20+/yr.)

Product 2: Answering User Questions

Methods

Question	Query Generation	Context Retrieval	Answer Extraction	Answer
• About data in Knowledge Graph • Documents, sites, people, teams	• Translation LLM: TinyLLaMA • Question to Cypher • Few-shot learning w/ rules, schema	• Run query on Knowledge Graph • Retrieve data required for answer	• Answer LLM: RoBERTa-Squad2 • Retrieval Augmented Generation (RAG)	• Informed answer if context found • "I don't know" if no context found

Results

Sample Conversation

"How many people are on team X?"

"111"

Performance

- Infers answer from database
- **15-20 seconds** to answer on laptop
- No hallucination during testing
- Concept proven

Business Impact

Identifying Documents

1,200	$5.3 M	↓90%	1,000
Hours Saved	Savings Per Year	Manual Work	Users

Current: Q2 2024 → Projected: Q1 2025

Retrieving Information

14	150	420+	300
Mins Saved / Search	Users	Hours Saved / Year	Users

Current: Q2 2024 → Projected: Q1 2025