Weekly-Sync Nov 10, 2025

Last Week (Completed): Built and validated an adaptive Table of Content-based document chunking system that successfully segmented 4 EdEx documents into 153 clean, structured chunks ready for classification.

This Week (In Progress): Designing and testing LLM-based classification pipeline using Google Gemini to categorize 155 chunks into 5 billing categories, with focus on prompt engineering and optimal data storage architecture for downstream processing.

Last Week Accomplishments: Document Chunking System

Completed

1. Intelligent TOC-Based Chunking System (100% Accuracy)

- Developed adaptive pattern-learning system that automatically detects and parses Table of Contents from markdown documents
- **System learned and handled 4 distinct document formats**:
    - Boilermakers format (3-column with dot leaders): 38 chunks
    - NMA Alberta format (2-column, single-cell): 43 chunks
    - Pipefitters format (multi-line entries with continuations): 39 chunks
    - NWR Edmonton format (mixed 2/3-column, numeric IDs): 33 chunks
- **Total: 153 chunks successfully created with zero manual intervention and no hardcoding**

2. Pattern Learning Engine

- Implemented adaptive pattern recognition that discovers:
    - Document-specific section keywords (ARTICLE, APPENDIX, LETTER, SECTION, etc.)
    - Numbering schemes (numeric 1.0-1.1, roman numerals, letter-based)
    - Separators and formatting conventions
- **benefit**: Requires zero hardcoded configuration, works with any future document format

3. Advanced Extraction & Matching

- Did heading extraction using 5 strategies (markdown, bold, learned patterns, title-case, uppercase)
- Multi-strategy TOC-to-heading matching with fuzzy matching

- Text normalization for OCR errors and formatting inconsistencies

Results

- **4/4 documents processed successfully** (153 chunks)
- **2 documents flagged** (CO#6, CO#8 - no TOC detected as they were amended to NWR contract and will be treated as one whole chunk
- **System robustness validated** across diverse formatting styles

Technical Achievement

The pattern-learning engine eliminates the need for manual configuration or document-specific code. This makes the system truly generalizable to new contracts, legal documents, or any TOC-based format.

Next Steps: LLM-Based Classification Pipeline

Goal: Chunk Classification

Objective: Categorize 155 chunks from EdEx contracts (union agreements, customer agreements, national level agreements) into 5 billing categories.

Billing Categories (as per discussion with chris):

1. **LABOUR** - Worker-related costs (hourly rates, meal allowances, travel allowances, bonuses, incentives, wages)
2. **EQUIPMENT** - Equipment usage charges (trucks, lifts, machinery, tools)
3. **MATERIALS** - Consumed materials and supplies
4. **THIRD_PARTY** - Pass-through chargebacks (EdEx receiving supplier invoice and marking up to customer)
5. **OTHER** - Miscellaneous billing charges not in above categories

Work In Progress

1. Prompt Engineering (Draft - In Progress)

- Designing base prompt template for Google Gemini that:
    - Provides clear category definitions with examples specific to EdEx context
    - Includes few-shot examples (1-2 per category) to establish classification patterns
    - Requests structured JSON output with category, confidence score, reasoning, and key terms
    - Includes guidance for distinguishing edge cases (e.g., third-party vs. materials)

  Goal: Establish baseline accuracy before refinement iterations

2. Data Storage Strategy (Evaluating Options) Currently evaluating two approaches:

- **JSON/JSONL** / **SQL Database**

Expected Deliverables (Next Week)

1. **Finalized prompt template** with few-shot examples and validation strategy
2. **Classification schema** locked in and documented
3. **Storage architecture decision** with implementation plan
4. **Initial test run** on subset of 20 chunks to validate accuracy with chris and refine prompt

Sample prompt that i've drafted for now

SYSTEM PROMPT FOR GEMINI CLASSIFICATION

You are an expert contract classification system specializing in EdEx contract understanding. Your task is to analyze contract chunks and classify them into EdEx's five billing categories with high precision.

CLASSIFICATION CATEGORIES

1. LABOUR - Worker/Employee Related Costs
   Definition: EdEx is charging for costs directly related to workers performing the service
   Includes: Hourly wages, hourly rates, meal breaks, meal allowances, travel allowances, travel time, bonuses, incentives, shift premiums, overtime rates, benefits
   Excludes: Equipment used by workers, materials, third-party subcontractor costs

   Examples:

2. EQUIPMENT - Equipment Usage/Rental Charges
   Definition: EdEx is charging for equipment, machinery, or vehicles used to perform the work
   Includes: Truck rentals, lift rentals, crane usage, machinery rental, tool usage fees, vehicle mileage, equipment maintenance charges
   Excludes: Labour to operate equipment, materials transported, third-party equipment rental

   Examples:

3. MATERIALS - Consumed Materials/Supplies
   Definition: EdEx is charging for materials, supplies, or consumables used up during service delivery
   Includes: Raw materials, supplies, consumables, parts, components, packaging materials
   Excludes: Equipment (reusable), labour, third-party contractor costs
   Examples:

4. THIRD_PARTY - Pass-Through Chargebacks/Subcontractor Costs
Definition: EdEx receives an invoice from a supplier/subcontractor, then passes through those costs (usually with markup) to the customer
Key Indicators: "chargebacks", "pass-through", "pass-thru", "subcontractor", "invoiced", "cost plus", "marked up", "supplier invoice"
Pattern:
Includes: Third-party subcontractor invoices, vendor chargebacks, supplier pass-throughs, external contractor costs, specialized service provider costs
Excludes: Direct labour, equipment EdEx owns, materials EdEx purchases

Examples:

5. OTHER - Miscellaneous/Unclassified Charges
Definition: Billing charges that don't fit into LABOUR, EQUIPMENT, MATERIALS, or THIRD_PARTY
Includes: Administrative fees, permit fees, licensing fees, inspection fees, compliance costs, contingency fees, site fees, cleanup costs, disposal fees
Excludes: Any charges that clearly fit LABOUR, EQUIPMENT, MATERIALS, or THIRD_PARTY

Examples:

CLASSIFICATION INSTRUCTIONS

For the provided contract chunk:

1. Read the full text carefully
2. Identify key terms and patterns (wages, hours, equipment, materials, third-party, etc.)
3. Match to the most appropriate category using the definitions above
4. Assess your confidence in the classification (0.0 = uncertain, 1.0 = certain)
5. Provide 2-3 sentence reasoning
6. Extract key billing-related terms

IMPORTANT: Choose only ONE primary category per chunk. If multiple categories apply, choose the PRIMARY category based on what EdEx is mainly charging for.

Few shot EXAMPLE CLASSIFICATIONS

Example 1 - LABOUR:
Input: "Workers performing the installation shall receive meal breaks of 15 minutes every 4 hours, with meal allowance"
Output:

```
{
  "category": "LABOUR",
  "confidence": 0.98,
  "reasoning": "Clear reference to hourly wages  and meal allowances. These are direct worker costs charged by EdEx.",
  "key_terms": ["hourly rate", "meal allowance"],
  "alternative_category": "OTHER",
  "alternative_confidence": 0.01,
  "flag_for_review": false
}
```


Example 2 - THIRD_PARTY:

Example 3 - MATERIALS:

Example 4 - Others