Advanced Product Mapping

Yihao Fang, Ph.D.

December 16th, 2024

Context and Objective

- Context:
  - Stakeholder manages convenience-store-like markets, receiving weekly supplier shipments.
  - Two datasets involved:
    - Internal Product List (stakeholder's)
    - External Product List (suppliers')
- Objective:
  - Replace slow, manual mapping of product lists with an intelligent, automated system.
  - Ensure exact matches based on:
    - Manufacturer
    - Name
    - Size
  - Integrate prompt engineering into the solution.

Examples and Challenge

- Examples:
  - Correct Match:
    - External: DIET LIPTON GREEN TEA W/ CITRUS 20 OZ
    - Internal: Lipton Diet Green Tea with Citrus (20oz)
  - Wrong Match:
    - External: Hersheys Almond Milk Choco 1.6 oz
    - Internal: Hersheys Milk Chocolate with Almonds (1.85oz)

- Key Challenge:
  - Designing a robust solution to minimize manual effort while maintaining high accuracy in product mapping.

Comparison of BM25-Llama 3.2 vs. ChatGPT o1-mini Approaches

BM25 Retriever + Llama 3.2 (Open Source)

- Strengths:
  - Cost-efficient solution with no usage-based fees.
  - Combines advanced NLP techniques:
    - BM25 retrieval for candidate selection.
    - Llama 3.2 LLM with few-shot prompting and chain-of-thought reasoning for accuracy and interpretability.
  - Employs self-consistency voting for reliable results.
  - Output aggregated using a map-reduce-style framework for structured summaries.
- Weaknesses:
  - Higher setup complexity requiring custom pipelines.

ChatGPT o1-mini (Proprietary)

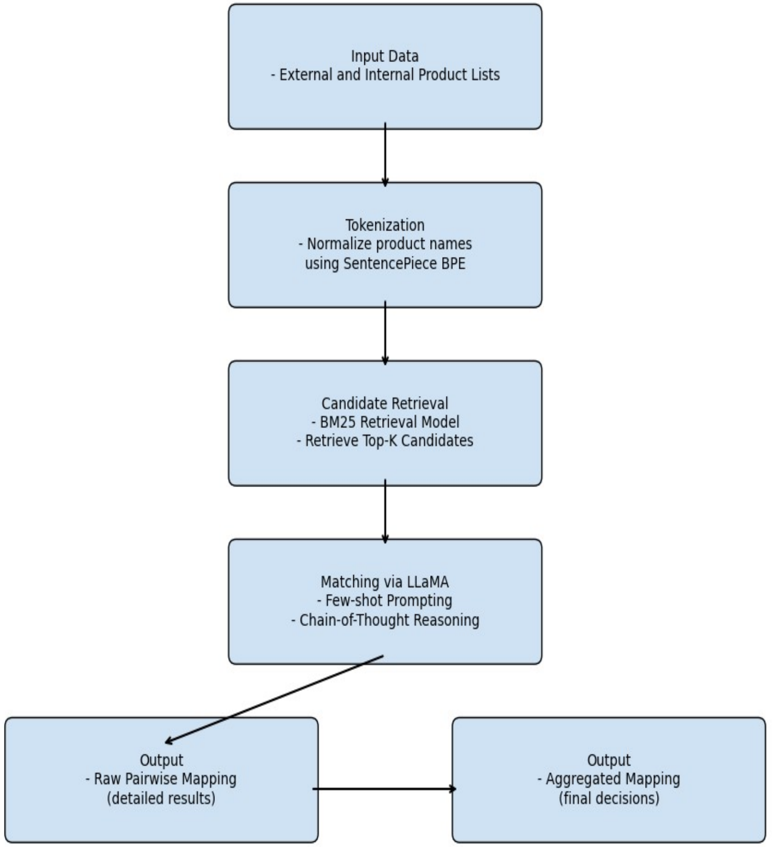
- Strengths:
  - Simplifies complex tasks with intuitive workflows.
  - Leverages ChatGPT's reasoning and language capabilities:
    - Iterative refinement for accurate and scalable results.
    - Prompt engineering reduces reliance on custom algorithms.
  - Faster deployment with reduced implementation complexity.
- Weaknesses:
  - Higher cost compared to open-source solutions.
  - Limited customizability.

Leveraging SentencePiece, BM25, and LLaMA for Accurate Product Matching

- Techniques Used:
  - SentencePiece Subword Encoding
  - BM25 Retrieval Model
  - LLaMA 3.2 Large Language Model
  - Few-shot Prompting
  - Chain-of-Thought Reasoning
  - Self-Consistency with Majority Voting
- Objective:
  - To accurately map external product names to internal product names using a robust, explainable, and efficient pipeline.

Product Mapping Pipeline: Key Components

- Diagram:
  - Input Data
    - External and Internal Product Lists
  - Tokenization
    - Normalize product names using SentencePiece BPE subword encoding.
  - Candidate Retrieval
    - Use BM25 Retrieval Model to rank and retrieve the top-K internal product candidates for each external product.
  - Matching via LLaMA
    - Employ the LLaMA 3.2 model with few-shot prompting and chain-of-thought reasoning to evaluate matches.
- Output:
  - Raw Pairwise Mapping (detailed results)
  - Aggregated Mapping (final decisions)



Enhancing Accuracy Through Self-Consistency

- Key Points:
  - LLM Probabilistic Nature:
    - Single model generations may produce inconsistent results.
  - Self-Consistency with Majority Voting:
    - Generate multiple responses for each input using LLaMA 3.2.
    - Compare responses and select the most frequent (mode) result.
- Aggregation Process:
  - Map-Reduce Workflow:
    - Raw results recorded in a comprehensive mapping matrix.
    - Aggregated results summarize external-to-internal mappings with final decisions.
- Outcome:
  - Improved accuracy and reliability at the cost of higher compute.

Results

- Outputs:
  - Detailed Pairwise Mapping (mapping_raw.csv)
  - Aggregated Final Results (mapping_aggr.csv)

External_Product	Internal_Product
CELSIUS PEACH VIBE 12 OZ	Celsius Sparkling Peach Vibe (12oz)
DOVE BAR DARK CHOC 1.44 OZ	Dove Dark Chocolate Bar (1.44oz)
FAIRLIFE 2% STRAWBERRY MILK 14 OZ	Fairlife 2% Ultra Filtered Strawberry Milk (14oz)
HY Hersheys Milk Chocolate w Almonds 1.45oz Each	Hersheys Milk Chocolate with Almonds (1.45oz)
...	...

Simplifying Real-World Tasks with ChatGPT o1 Family

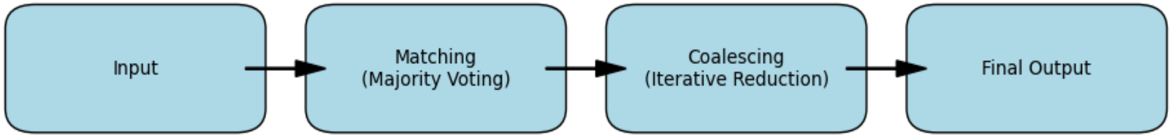
- Leveraging the reasoning capabilities of ChatGPT o1 models.
- Tackling complex workflows with simplicity and scalability.
- Achieving robust results through iterative and parallel processing.

Breaking Down the Problem

- Objective:
  - Match external and internal product lists and refine results into a coalesced final table.
- Challenges:
  - Large datasets with potential inconsistencies.
  - Need for accuracy and scalability.
- Solution:
  - Divide tasks into smaller chunks.
  - Use ChatGPT o1-mini for reasoning and decision-making via prompt engineering.

Simplified Workflow with ChatGPT

- Step 1: Matching
  - Generate multiple outputs for each chunk using ChatGPT.
  - Use majority voting to determine the most consistent results.
- Step 2: Coalescing
  - Combine intermediate results iteratively.
  - ChatGPT merges results by prioritizing non-NULL values and outputs the final table.
- Parallel Processing:
  - Both steps executed efficiently with multithreading to handle large datasets.



Why ChatGPT o1 Family?

- Simplifies Complexity:
  - Replaces intricate logic with straightforward AI-driven decisions.
  - Clear prompts guide reasoning and output consistency.
- Highly Scalable:
  - Multithreading efficiently processes large datasets.
- Robust and Accurate:
  - Majority voting ensures self-consistency.
  - Coalescing merges results systematically, prioritizing data integrity.
- Flexible and Intuitive:
  - Adaptable to different datasets and tasks with minor adjustments.

Takeaways

- BM25-Llama 3.2: Best for cost-sensitive, customizable applications requiring advanced interpretability.
- ChatGPT o1 family: Ideal for quick, scalable deployment with minimal development effort.