Stanford CS224v Course
Conversational Virtual Assistants with Deep Learning

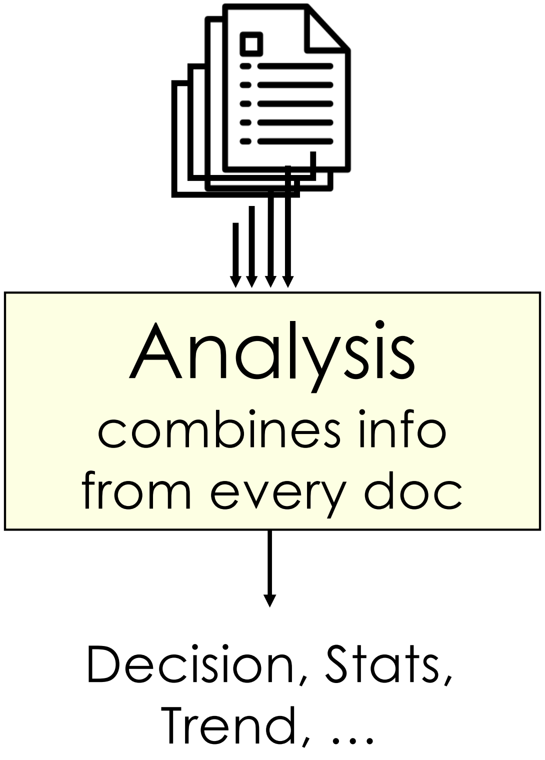
Lecture 8

Question Answering on Sets of Long Documents

SLIDERS: the Scalable Long-Document Integration through Decomposed Extraction and Reconciliation System

Monica Lam Harshit Joshi

Knowledge in Large Sets of Long Documents



Analysis

combines info from every doc

Decision, Stats, Trend, …

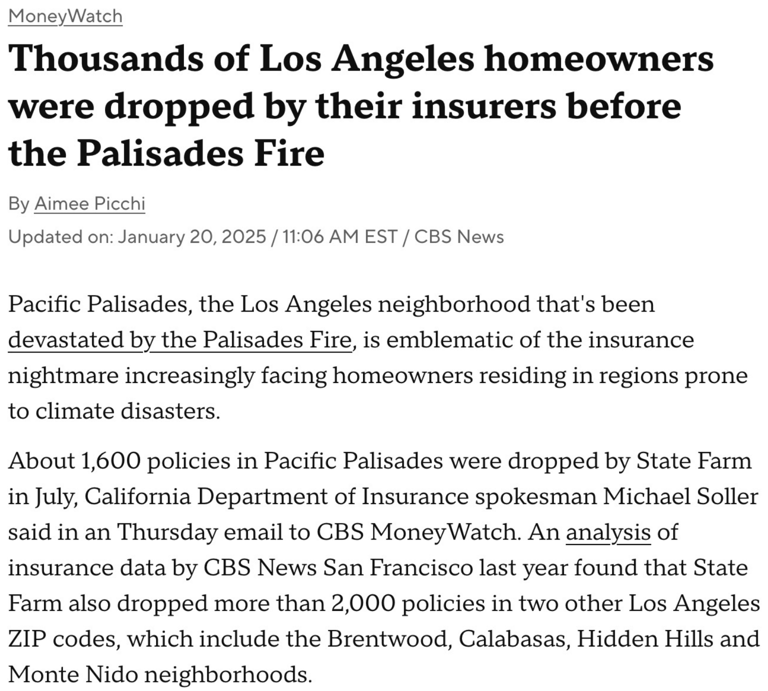
Future lecture

Resumes Paper Reviews Proposals Financial Reports	Pubmed Articles Human Genome Drug/Protein/Disease Event News	News Social Media Interviews Medical Rec.	Requirements Regulations Clinical Trials Invoice Match
Ranking (Top n) of ALL documents	Accumulation	Stats/trends	Satisfiability
Finalists	Knowledge Bases	Discover Knowledge	Compliance

Examples from Class Projects

- U.S. Securities and Exchange Commission (SEC) data – investment decisions
- Insurance filings – analyzing insurance policies
- Hazard mitigation plans – for finding funding opportunities

Insurance



MoneyWatch

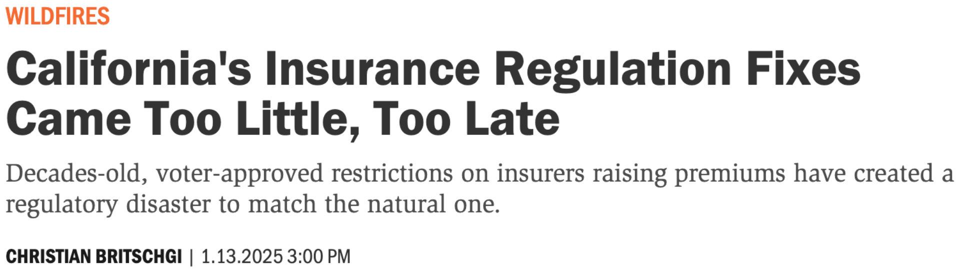
Thousands of Los Angeles homeowners were dropped by their insurers before the Palisades Fire

By Aimee Picchi

Updated on: January 20, 2025 / 11:06 AM EST / CBS News

Pacific Palisades, the Los Angeles neighborhood that's been devastated by the Palisades Fire, is emblematic of the insurance nightmare increasingly facing homeowners residing in regions prone to climate disasters.

About 1,600 policies in Pacific Palisades were dropped by State Farm in July, California Department of Insurance spokesman Michael Soller said in an Thursday email to CBS MoneyWatch. An analysis of insurance data by CBS News San Francisco last year found that State Farm also dropped more than 2,000 policies in two other Los Angeles ZIP codes, which include the Brentwood, Calabasas, Hidden Hills and Monte Nido neighborhoods.



WILDFIRES

California's Insurance Regulation Fixes Came Too Little, Too Late

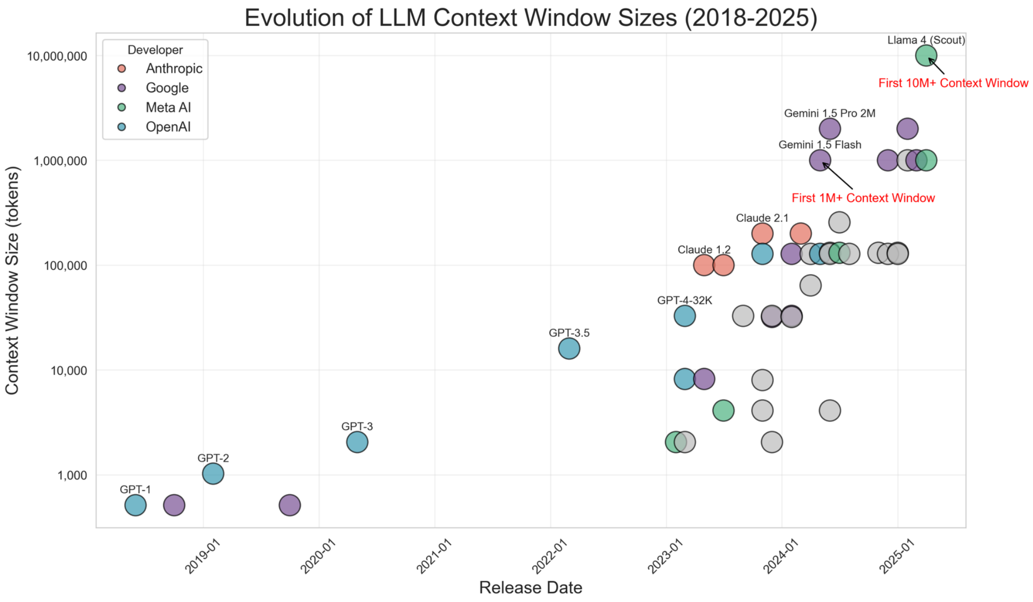
Decades-old, voter-approved restrictions on insurers raising premiums have created a regulatory disaster to match the natural one.

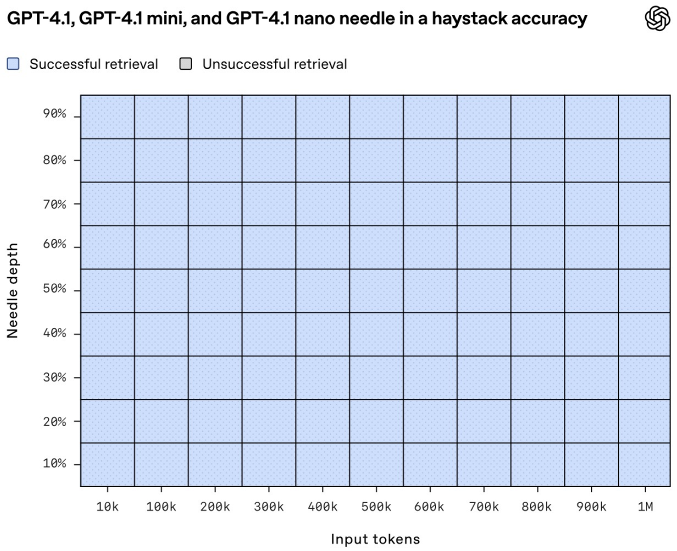
CHRISTIAN BRITSCHGI | 1.13.2025 3:00 PM

- Climate changes dramatically change natural disaster risks
- Insurance is regulated by the government
- Insurance companies have to file "rate filings" to get approval
- Filings today are tentatively approved today -- lacking manpower to process documents

LLMs Long-Context Capability

- Ever increasing context window of LLMs
- Near perfect needle in a haystack capabilities





In our internal needle in a haystack eval, GPT-4.1, GPT-4.1 mini, and GPT 4.1 nano are all able to retrieve the needle at all positions in the context up to 1M.

GPT 4.1 blogpost

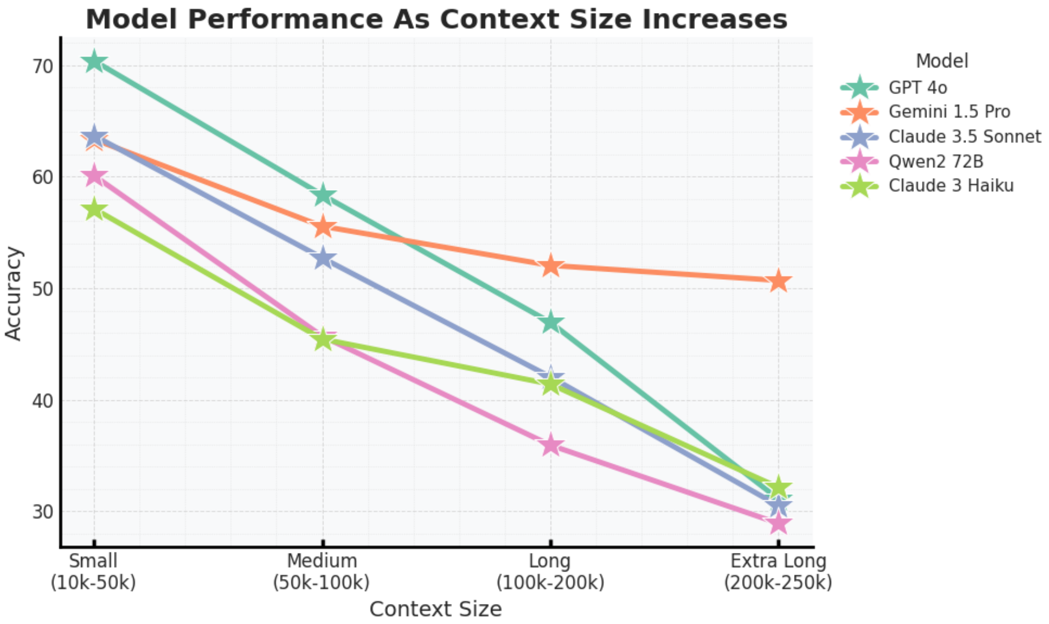
LLM Context Windows are Never Long Enough

Many long documents – and many documents in a set!

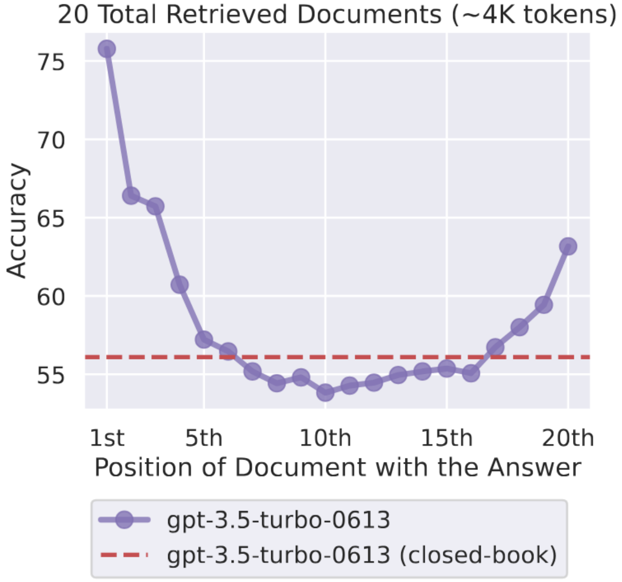
- Pubmed articles (as of 23 May 2023)
  - 24.6M have abstracts, 26.8M link to full text, 10.9M free full text
  - 1M new records added each year (2010-2019)
  - 35M citations
- 600k-700k annual SEC filings.
  - Each document has ~100k tokens
- Thousands of Insurance Filing documents for a single state

LLMs Lack Precision for Long Contexts

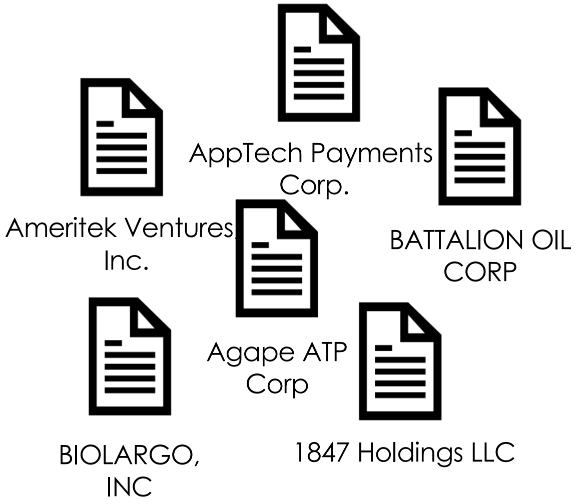
- Context length increases → performance for the same task decreases
- Attention is not uniform across the entire context window.



"Loong Benchmark" Wang et. al 2024



"Lost in the middle" Liu et. al 2023

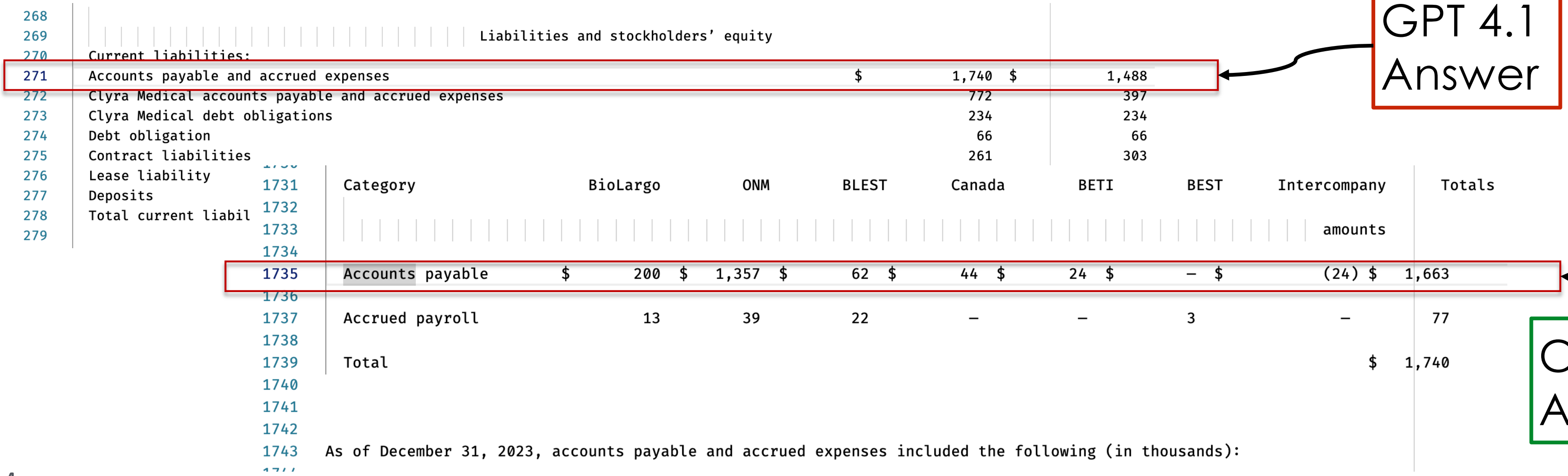


Examples of Imprecision

What is the accounts payable of BIOLARGO, INC.?

GPT 4.1:
The accounts payable of BIOLARGO, INC. as of March 31, 2024, is $1,740,000.

Line 271: accounts payable and accrued expenses
Line 1735: accounts payable,
separate from accrued expenses (line 1737)



```
268
269                                  Liabilities and stockholders' equity
270    Current liabilities:
271    Accounts payable and accrued expenses                        $     1,740  $      1,488
272    Clyra Medical accounts payable and accrued expenses                  772           397
273    Clyra Medical debt obligations                                       234           234
274    Debt obligation                                                       66            66
275    Contract liabilities                                                 261           303
276    Lease liability
277    Deposits
278    Total current liabil
279
```

```
1731    Category              BioLargo     ONM     BLEST    Canada    BETI    BEST    Intercompany    Totals
1732
1733                                                                                       amounts
1734
1735    Accounts payable    $      200  $ 1,357  $    62  $     44  $    24  $     –  $        (24) $   1,663
1736
1737    Accrued payroll             13      39        22        –        –       3              –         77
1738
1739    Total                                                                                      $   1,740
1740
1741
1742
1743   As of December 31, 2023, accounts payable and accrued expenses included the following (in thousands):
```

GPT 4.1 Answer

Correct Answer

Examples of Imprecision

Given 4 academic papers

We hope you will carefully study the provided papers and determine the citation relationships between them.

1. Reference: references are about what the given paper is using.

2. Citation: citations are about who is using the given paper.

The paper you need to analyze:
Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems

GPT

```
{
"Reference": [],
 'Citation’:
['# Unbridled Icarus: A Survey of the Potential Perils of Image Inputs in Multimodal Large Language Model Security ‘]
}
```

Gold Answer

```
{
'Reference’: ['# Evil Geniuses: Delving into the Safety of LLM-based Agents ‘,
                '# Towards Optimal Statistical Watermarking ‘],
 'Citation’:
['# Unbridled Icarus: A Survey of the Potential Perils of Image Inputs in Multimodal Large Language Model Security ‘]
}
```

Misses several references

Motivation and Background:
Precise QA for Sets of Long Documents

- A huge need to analyze across many long documents
- LLM contexts are never long enough
- Even if the context is long enough,
  the precision degrades with increasing length
  - Finding a needle in a haystack is considered easier,
  - The precision depends on the position
  - When the answer is wrong,
    - It is not interpretable
    - We don’t have a way to improve it

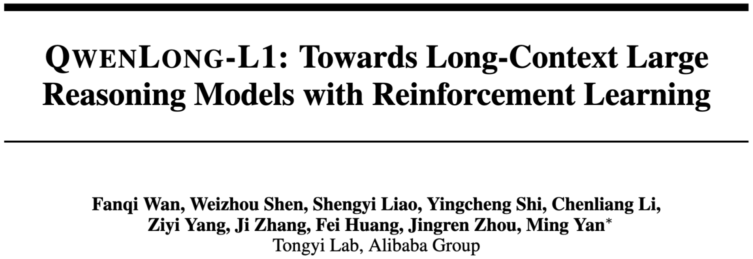
Outline

- **High-level approaches: Training vs. Chunking**
- Introduction to SLIDERS
- Design
- Preliminary Evaluation

Approaches

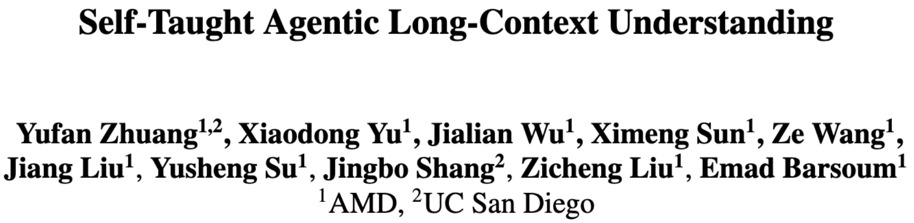
- Training based: to improve precision
- Chunking based: to improve precision and scaling
  - Representations for chunks
    - Natural language
    - Structured representation

TRAINING BASED



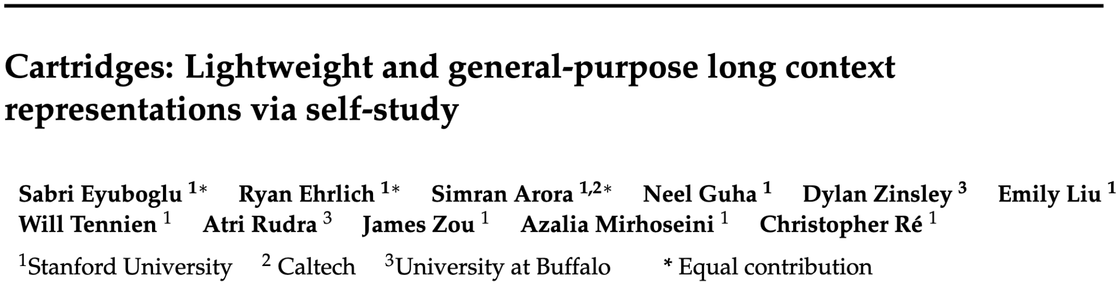
QWENLONG-L1: Towards Long-Context Large Reasoning Models with Reinforcement Learning

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li,
Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, Ming Yan*
Tongyi Lab, Alibaba Group



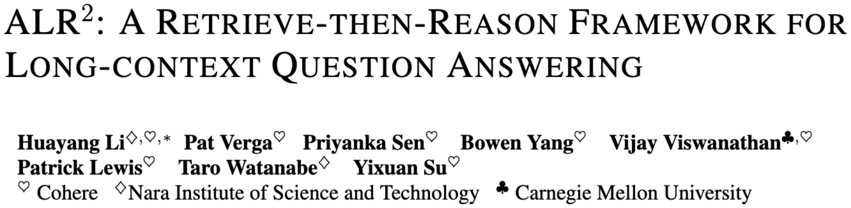
Self-Taught Agentic Long-Context Understanding

Yufan Zhuang[1,2], Xiaodong Yu[1], Jialian Wu[1], Ximeng Sun[1], Ze Wang[1],
Jiang Liu[1], Yusheng Su[1], Jingbo Shang[2], Zicheng Liu[1], Emad Barsoum[1]
[1]AMD, [2]UC San Diego



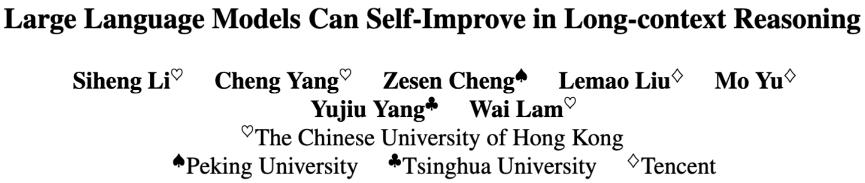
Cartridges: Lightweight and general-purpose long context representations via self-study

Sabri Eyuboglu [1*] **Ryan Ehrlich [1*]** **Simran Arora [1,2*]** **Neel Guha [1]** **Dylan Zinsley [3]** **Emily Liu [1]**
Will Tennien [1] **Atri Rudra [3]** **James Zou [1]** **Azalia Mirhoseini [1]** **Christopher Ré [1]**
[1]Stanford University [2] Caltech [3]University at Buffalo * Equal contribution



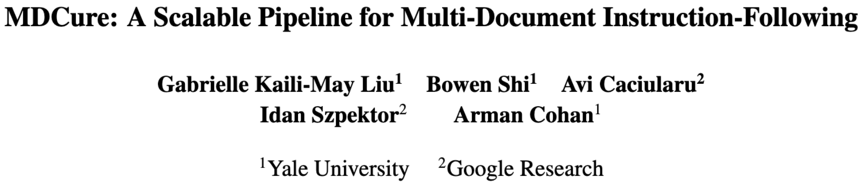
ALR[2]: A RETRIEVE-THEN-REASON FRAMEWORK FOR LONG-CONTEXT QUESTION ANSWERING

Huayang Li[◇,♡,*] **Pat Verga[♡]** **Priyanka Sen[♡]** **Bowen Yang[♡]** **Vijay Viswanathan[♣,♡]**
Patrick Lewis[♡] **Taro Watanabe[◇]** **Yixuan Su[♡]**
[♡] Cohere [◇]Nara Institute of Science and Technology [♣] Carnegie Mellon University



Large Language Models Can Self-Improve in Long-context Reasoning

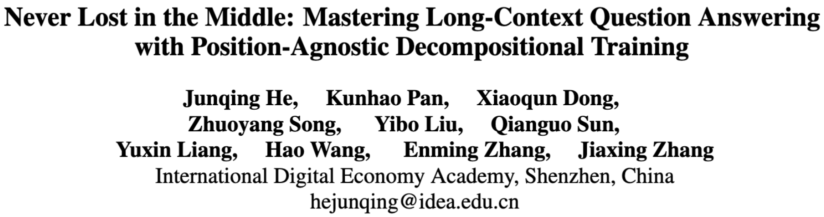
Siheng Li[♡] **Cheng Yang[♡]** **Zesen Cheng[♠]** **Lemao Liu[◇]** **Mo Yu[◇]**
Yujiu Yang[♣] **Wai Lam[♡]**
[♡]The Chinese University of Hong Kong
[♠]Peking University [♣]Tsinghua University [◇]Tencent



MDCure: A Scalable Pipeline for Multi-Document Instruction-Following

Gabrielle Kaili-May Liu[1] **Bowen Shi[1]** **Avi Caciularu[2]**
Idan Szpektor[2] **Arman Cohan[1]**

[1]Yale University [2]Google Research



Never Lost in the Middle: Mastering Long-Context Question Answering with Position-Agnostic Decompositional Training

Junqing He, Kunhao Pan, Xiaoqun Dong,
Zhuoyang Song, Yibo Liu, Qianguo Sun,
Yuxin Liang, Hao Wang, Enming Zhang, Jiaxing Zhang
International Digital Economy Academy, Shenzhen, China
hejunqing@idea.edu.cn

Training-Based Methods

QwenLong-L1:

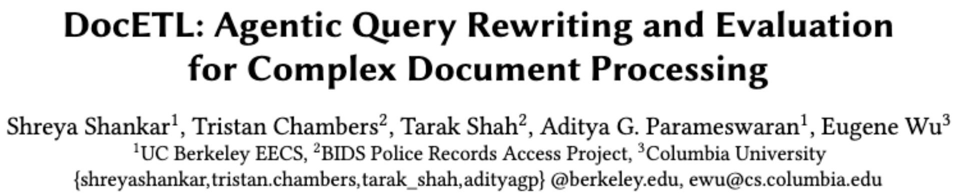
- Curate a long-context dataset using existing benchmarks,
- Progressively increase the context size with RL
  using rule-based verification and LLM-as-a-judge
- Improve over baseline models,
  but not SOTA with proprietary/open-source models

Cartridges

- Train KV cache on a document set by generating & answering questions
- Reuse cache to answer new questions across the trained document set
- Lose to complete in-context documents

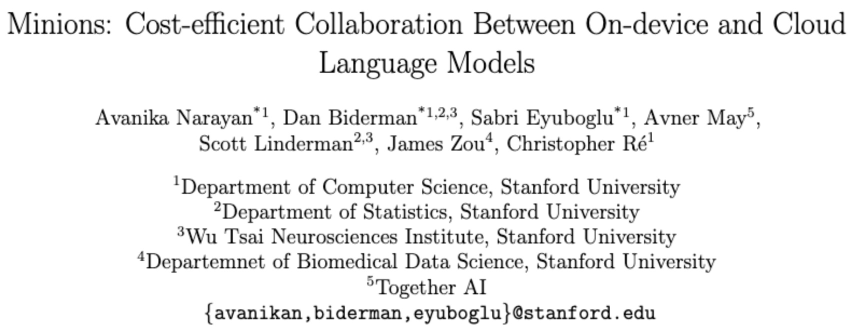
“QWENLONG-L1: Towards Long-Context Large Reasoning Models with Reinforcement Learning” Wan et. al 2025
“Cartridges: Lightweight and general-purpose long context representations via self-study”, Eyuboglu et. al 2025

CHUNKING-BASED



DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing

Shreya Shankar[1], Tristan Chambers[2], Tarak Shah[2], Aditya G. Parameswaran[1], Eugene Wu[3]
[1]UC Berkeley EECS, [2]BIDS Police Records Access Project, [3]Columbia University
{shreyashankar,tristan.chambers,tarak_shah,adityagp} @berkeley.edu, ewu@cs.columbia.edu



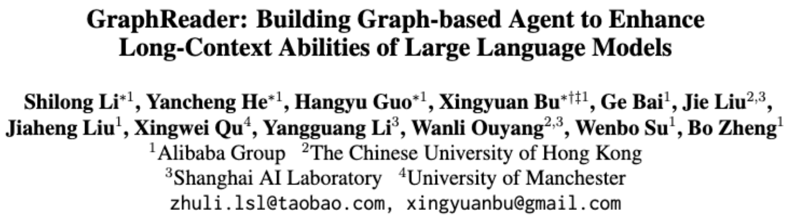
Minions: Cost-efficient Collaboration Between On-device and Cloud Language Models

Avanika Narayan[*1], Dan Biderman[*1,2,3], Sabri Eyuboglu[*1], Avner May[5],
Scott Linderman[2,3], James Zou[4], Christopher Ré[1]

[1]Department of Computer Science, Stanford University
[2]Department of Statistics, Stanford University
[3]Wu Tsai Neurosciences Institute, Stanford University
[4]Departemnet of Biomedical Data Science, Stanford University
[5]Together AI
{avanikan,biderman,eyuboglu}@stanford.edu

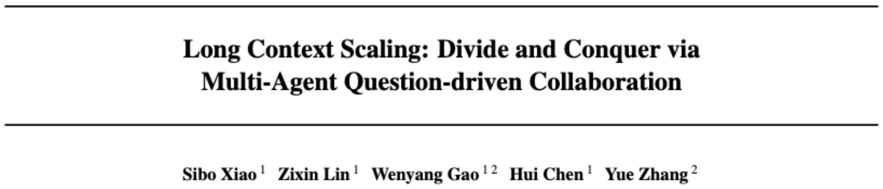
Chain of Agents: Large Language Models Collaborating on Long-Context Tasks

Yusen Zhang[♣*], Ruoxi Sun[◇], Yanfei Chen[◇], Tomas Pfister[◇], Rui Zhang[♣†], Sercan Ö. Arik[◇†]
♣ Penn State University, ◇ Google Cloud AI Research
{yfz5488, rmz5227}@psu.edu, {ruoxis, yanfeichen, tpfister, soarik}@google.com



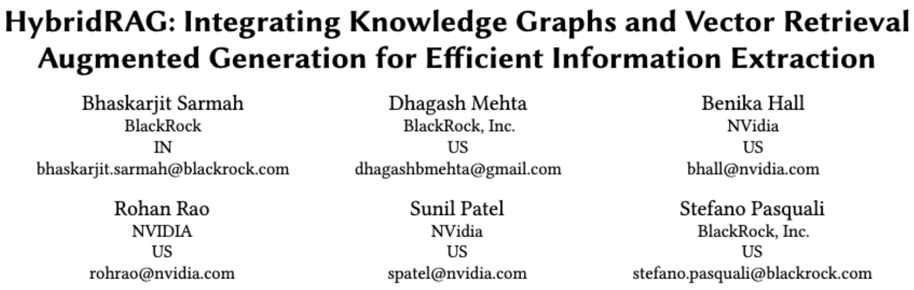
GraphReader: Building Graph-based Agent to Enhance Long-Context Abilities of Large Language Models

Shilong Li[*1], Yancheng He[*1], Hangyu Guo[*1], Xingyuan Bu[*†‡1], Ge Bai[1], Jie Liu[2,3],
Jiaheng Liu[1], Xingwei Qu[4], Yangguang Li[3], Wanli Ouyang[2,3], Wenbo Su[1], Bo Zheng[1]
[1]Alibaba Group [2]The Chinese University of Hong Kong
[3]Shanghai AI Laboratory [4]University of Manchester
zhuli.lsl@taobao.com, xingyuanbu@gmail.com



Long Context Scaling: Divide and Conquer via Multi-Agent Question-driven Collaboration

Sibo Xiao[1] Zixin Lin[1] Wenyang Gao[1 2] Hui Chen[1] Yue Zhang[2]



HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction

Bhaskarjit Sarmah
BlackRock
IN
bhaskarjit.sarmah@blackrock.com

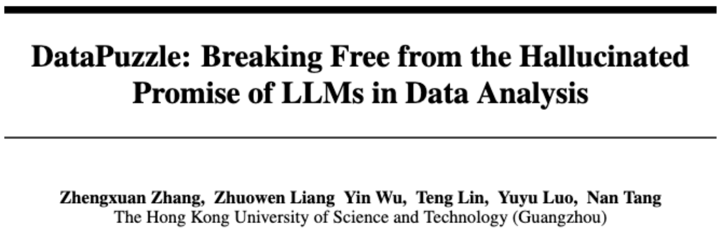
Dhagash Mehta
BlackRock, Inc.
US
dhagashbmehta@gmail.com

Benika Hall
NVidia
US
bhall@nvidia.com

Rohan Rao
NVIDIA
US
rohrao@nvidia.com

Sunil Patel
NVidia
US
spatel@nvidia.com

Stefano Pasquali
BlackRock, Inc.
US
stefano.pasquali@blackrock.com

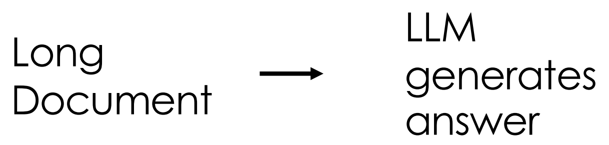


DataPuzzle: Breaking Free from the Hallucinated Promise of LLMs in Data Analysis

Zhengxuan Zhang, Zhuowen Liang Yin Wu, Teng Lin, Yuyu Luo, Nan Tang
The Hong Kong University of Science and Technology (Guangzhou)

Chunking-Based Methods

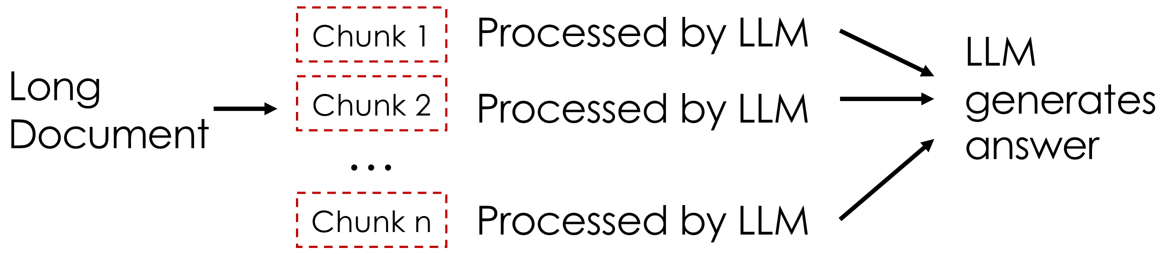
- Divide the context into multiple chunks.
- Answer question based on individual chunk
- Combine them together and get the final answer



Long Document

LLM generates answer

Single step approach



Long Document

Chunk 1 Processed by LLM

Chunk 2 Processed by LLM

…

Chunk n Processed by LLM

LLM generates answer

Decomposing the problem

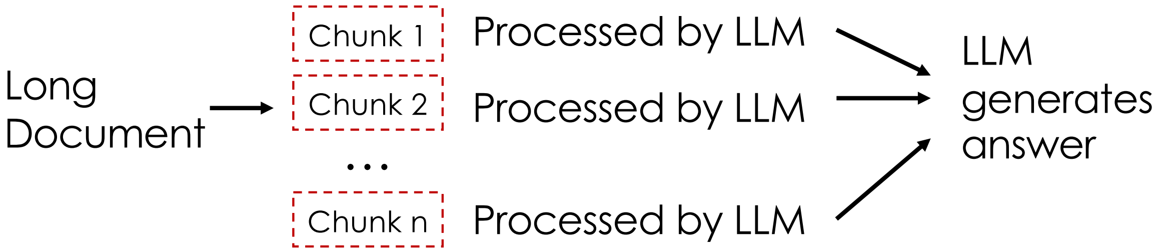
Chunking-Based Methods

- Chain of Agents:
  - An LLM processes a chunk and passes on summary w.r.t. the task to the next LLM call along with the next chunk.
  - Final agent synthesizes the answer based on last provided summary.
- DocETL:
  - User provides a schema for the representation
  - Each chunk is processed by an LLM that outputs structured data.
  - The user prompts the LLM to reduce/resolve the output from each chunk to answer the question.

"Chain of Agents: Large Language Models Collaborating on Long-Context Tasks" Zhang et. al 2024
"DocETL: Agentic Query Rewriting and Evaluation for Complex Document Processing" Shankar et. al 2024

Advantages of Chunking-Based Methods

- **Precision:**
  Processing each chunk individually increases attention to local details
- **Scalability:** Enables natural scalability to large document collections far beyond model limits



Long Document

Chunk 1 — Processed by LLM

Chunk 2 — Processed by LLM

…

Chunk n — Processed by LLM

LLM generates answer

Decomposing the problem

Existing Chunking-Based Methods

Use LLMs to synthesize a final answer from all chunk outputs

- **Effective for small inputs:**
  Works well with a few hundred tokens
- **Breaks down at scale:**
  Thousands of chunk outputs overwhelm the LLM
- **Scalability issue:**
  As chunks/documents grow,
  the synthesis context becomes too large and unreliable

Challenges of Chunking

1. **How do we represent the information in each chunk?**

2. **Problem answering separated into two steps**
   a. **Correct extraction of information chunk by chunk?**
      - Chunk boundaries: the chunks may not be self-contained

      - Lacking global context: interpretation may be incomplete

   b. **Compile correct answer from the set of extracted information**
      - Given the independently extracted information,
        what issues arise when assembling the answers.

      - What technique can we use to perform the assembly?

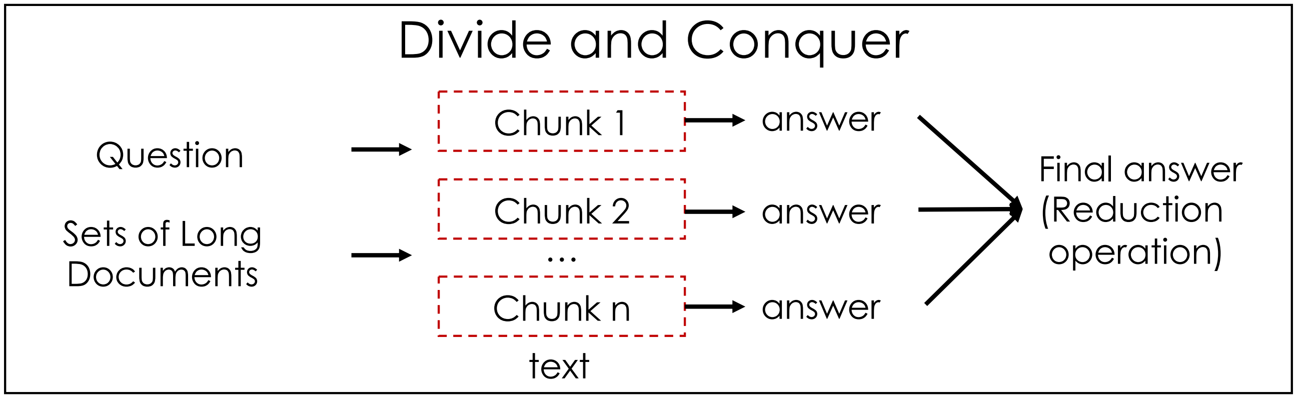
Our Solution: SLIDERS

1. **How do we represent the information in each chunk?**
   Represent chunks as rows in a table with an automatically-induced schema.
2. **Problem answering separated into two steps**
   a. **Correct extraction of information chunk by chunk?**
      - Chunk boundaries: the chunks may not be self-contained
        Semantics-driven chunking
      - Lacking global context: interpretation may be incomplete
        Contextualized information extraction

   b. **Compile correct answer from the set of extracted information**
      - Given the independently extracted information,
        what issues arise when assembling the answers
        We discovered: duplication, incomplete information, …
      - What technique can we use to perform the assembly?
        SQL

Outline

- High-level approaches: Training vs. Chunking
- Introduction to SLIDERS
- **Design**
- Preliminary Evaluation



Divide and Conquer

Question

Sets of Long Documents

Chunk 1 → answer

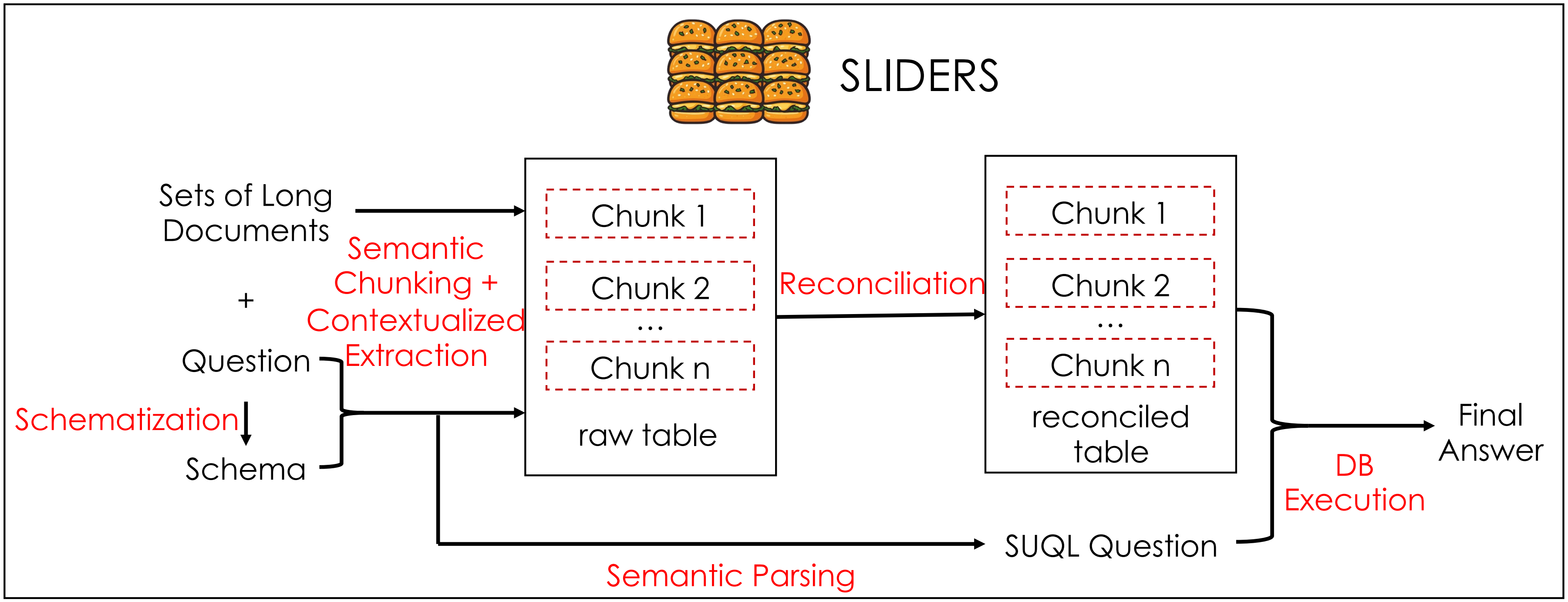
Chunk 2 → answer

…

Chunk n → answer

text

Final answer (Reduction operation)



SLIDERS

Sets of Long Documents

+

Question

Schematization

Schema

Semantic Chunking + Contextualized Extraction

Chunk 1

Chunk 2

…

Chunk n

raw table

Reconciliation

Chunk 1

Chunk 2

…

Chunk n

reconciled table

Semantic Parsing

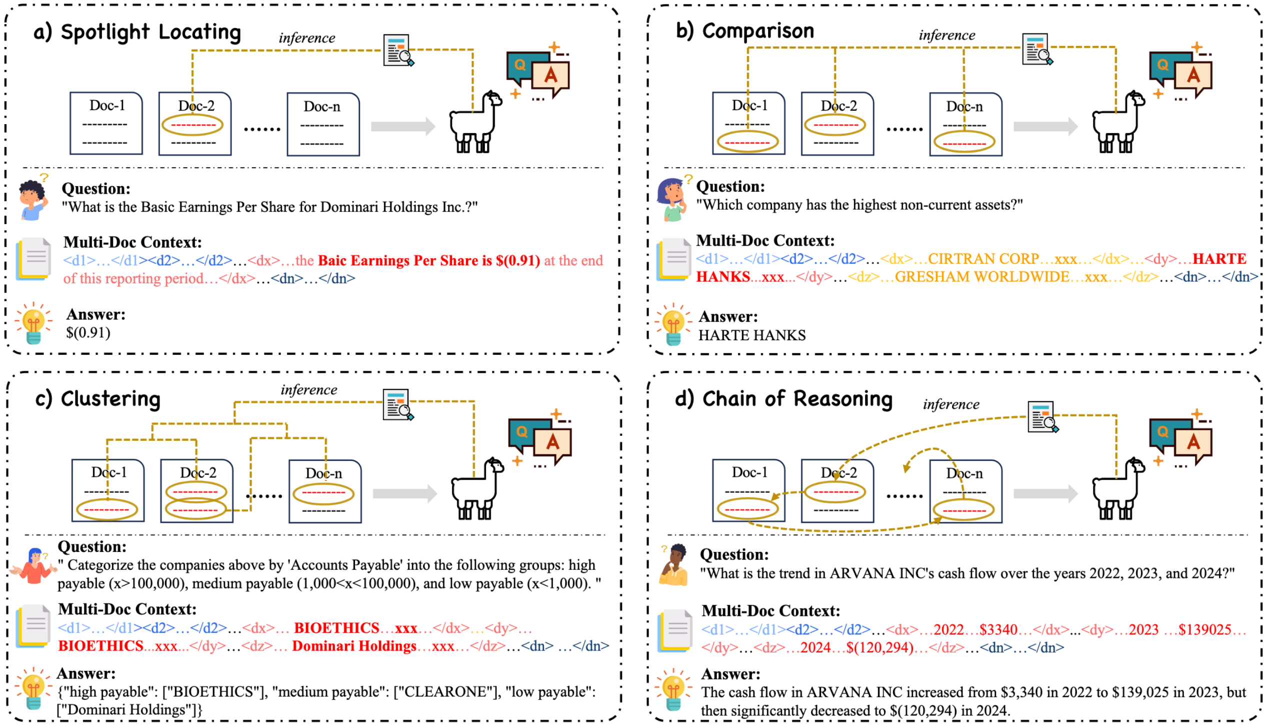
SUQL Question

DB Execution

Final Answer

Loong Benchmark

- Domains: Finance, Legal, Academic Papers
- Languages: English, Chinese



a) Spotlight Locating

Question:
"What is the Basic Earnings Per Share for Dominari Holdings Inc.?"

Multi-Doc Context:
<d1>…</d1><d2>…</d2>…<dx>…the **Baic Earnings Per Share is $(0.91)** at the end of this reporting period…</dx>…<dn>…</dn>

Answer:
$(0.91)

b) Comparison

Question:
"Which company has the highest non-current assets?"

Multi-Doc Context:
<d1>…</d1><d2>…</d2>…<dx>…CIRTRAN CORP…xxx…</dx>…<dy>…**HARTE HANKS**...xxx...</dy>…<dz>…GRESHAM WORLDWIDE…xxx…</dz>…<dn>…</dn>

Answer:
HARTE HANKS

c) Clustering

Question:
" Categorize the companies above by 'Accounts Payable' into the following groups: high payable (x>100,000), medium payable (1,000<x<100,000), and low payable (x<1,000). "

Multi-Doc Context:
<d1>…</d1><d2>…</d2>…<dx>… **BIOETHICS**…**xxx**…</dx>…<dy>… **BIOETHICS**...xxx...</dy>…<dz>… **Dominari Holdings**…xxx…</dz>…<dn> …</dn>

Answer:
{"high payable": ["BIOETHICS"], "medium payable": ["CLEARONE"], "low payable": ["Dominari Holdings"]}

d) Chain of Reasoning

Question:
"What is the trend in ARVANA INC's cash flow over the years 2022, 2023, and 2024?"

Multi-Doc Context:
<d1>…</d1><d2>…</d2>…<dx>…2022…$3340…</dx>...<dy>…2023 …$139025…</dy>…<dz>…2024…$(120,294)…</dz>…<dn>…</dn>

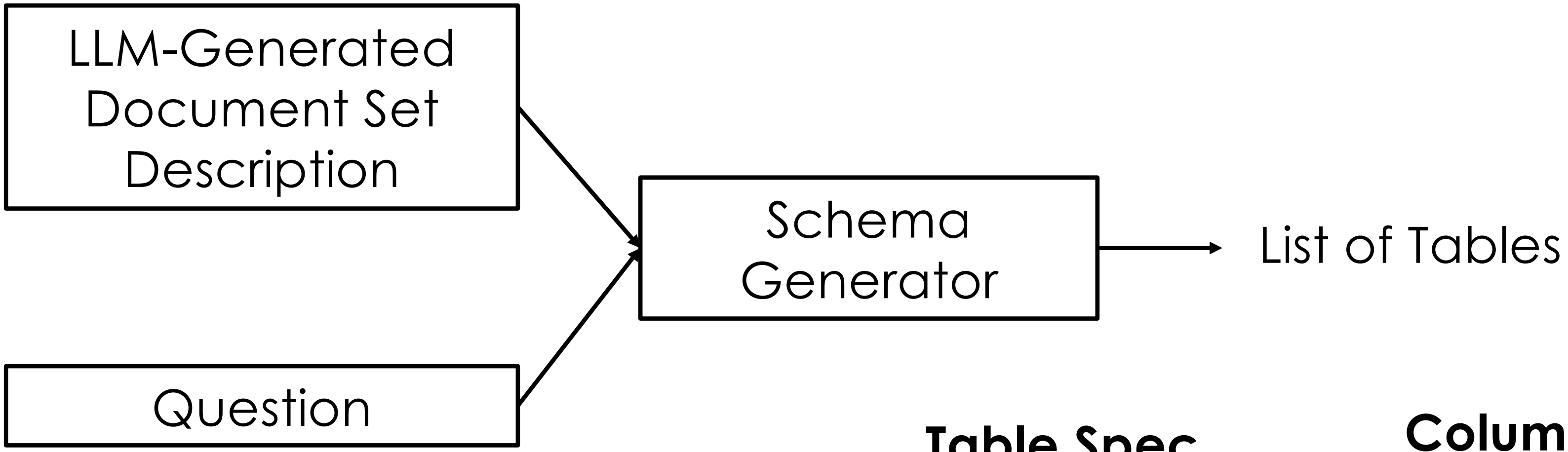
Answer:
The cash flow in ARVANA INC increased from $3,340 in 2022 to $139,025 in 2023, but then significantly decreased to $(120,294) in 2024.

Running Example

- Dataset: Loong (Finance, Set2: Comparison)
- Documents:
  - AIM ImmunoTech Inc. Q1 2024 10-Q Report
  - Dominari Holdings Inc. Q1 2024 10-Q Report
  - 1st Franklin Financial Corp Q1 2024 10-Q

- Question: Which company has the highest 'Total Shares Outstanding'?

LLM Induces Schema from the Question



LLM-Generated Document Set Description

Question

Schema Generator

List of Tables

Table Spec

Name

Description

Fields: List[Field]

Column Spec

Name

Description

Type

Unit

Scale

Normalization Rules

Example

Question	Which company has the highest Total Shares Outstanding'?
Table Name	SharesOutstanding
Table Description	Total shares outstanding for each company as reported in the financial statements, normalized to shares as of the reporting period end.”
Columns	company_name period_end_date total_shares_outstanding

Example

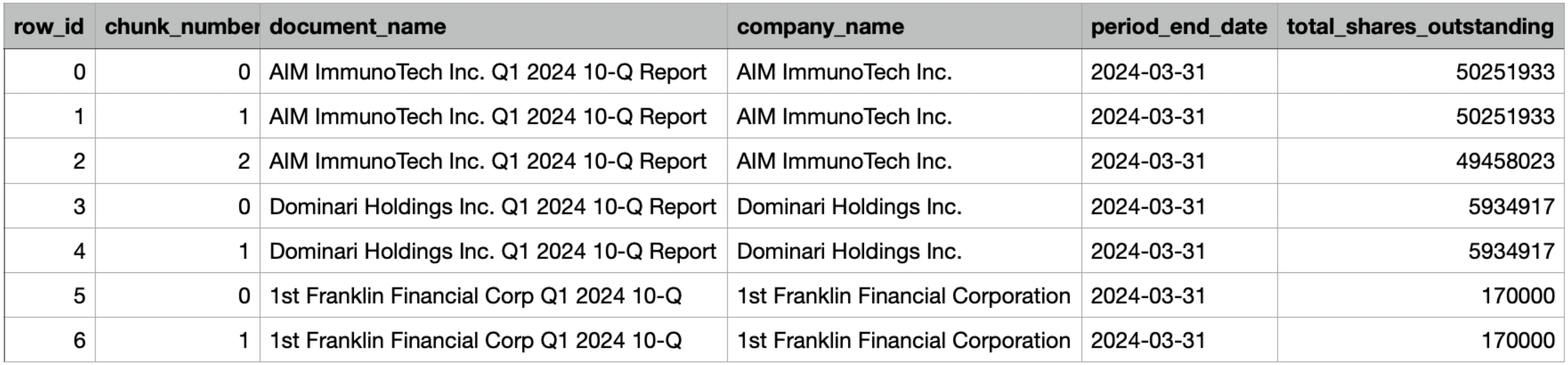
Which company has the highest 'Total Shares Outstanding'?

SharesOutstanding

Name	**company_name**	**period_end_date**	**total_shares_outstanding**
Data Type	str	str	float
Unit	None	None	shares
Scale	None	None	None
Description	"Name of the company as reported in the financial statements (e.g., 'AIM ImmunoTech Inc.')."	"End date of the reporting period for which shares outstanding is reported (e.g., '2024-03-31')."	"Total number of shares outstanding as of the period end date, as reported in the balance sheet or notes. Common surface forms: 'shares outstanding', 'common shares outstanding', 'total shares issued and outstanding'."
Normalization	None	Date_format: "YYYY-MM-DD"	None

Normalization

- For comparison and aggregation
  - Standardize units, scales, datatypes
  - Normalization rules to make all data consistent



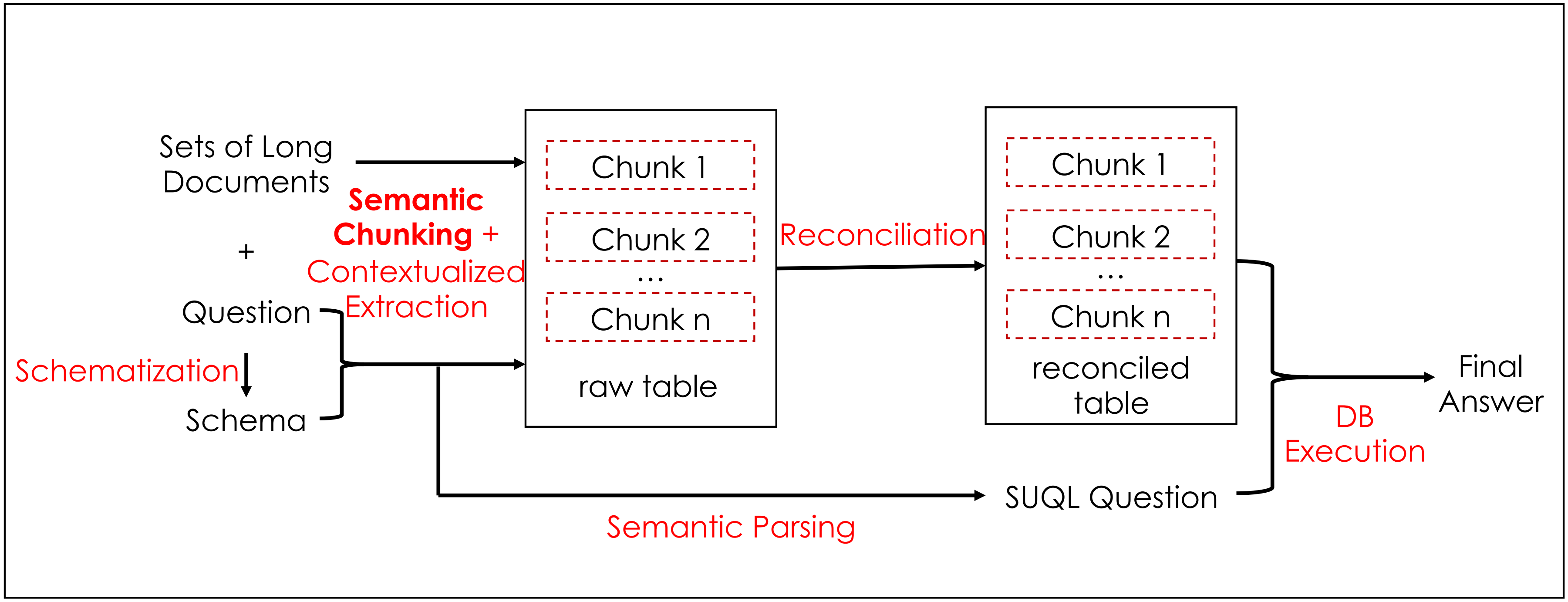
row_id	chunk_number	document_name	company_name	period_end_date	total_shares_outstanding
0	0	AIM ImmunoTech Inc. Q1 2024 10-Q Report	AIM ImmunoTech Inc.	2024-03-31	50251933
1	1	AIM ImmunoTech Inc. Q1 2024 10-Q Report	AIM ImmunoTech Inc.	2024-03-31	50251933
2	2	AIM ImmunoTech Inc. Q1 2024 10-Q Report	AIM ImmunoTech Inc.	2024-03-31	49458023
3	0	Dominari Holdings Inc. Q1 2024 10-Q Report	Dominari Holdings Inc.	2024-03-31	5934917
4	1	Dominari Holdings Inc. Q1 2024 10-Q Report	Dominari Holdings Inc.	2024-03-31	5934917
5	0	1st Franklin Financial Corp Q1 2024 10-Q	1st Franklin Financial Corporation	2024-03-31	170000
6	1	1st Franklin Financial Corp Q1 2024 10-Q	1st Franklin Financial Corporation	2024-03-31	170000

Schematization Discussion

- Automatic schematization appears to work well for 1 question
- Can schematize across multiple questions
    - To amortize the cost of extraction
- Domain expert can also improve the schema
    - Due to interpretability!

Note: Can handle arbitrarily many documents!
Support aggregation and comparison across documents
with SUQL!

SLIDERS



Sets of Long Documents

+

Question

Semantic Chunking + Contextualized Extraction

Schematization

Schema

Chunk 1

Chunk 2

…

Chunk n

raw table

Reconciliation

Chunk 1

Chunk 2

…

Chunk n

reconciled table

Semantic Parsing

SUQL Question

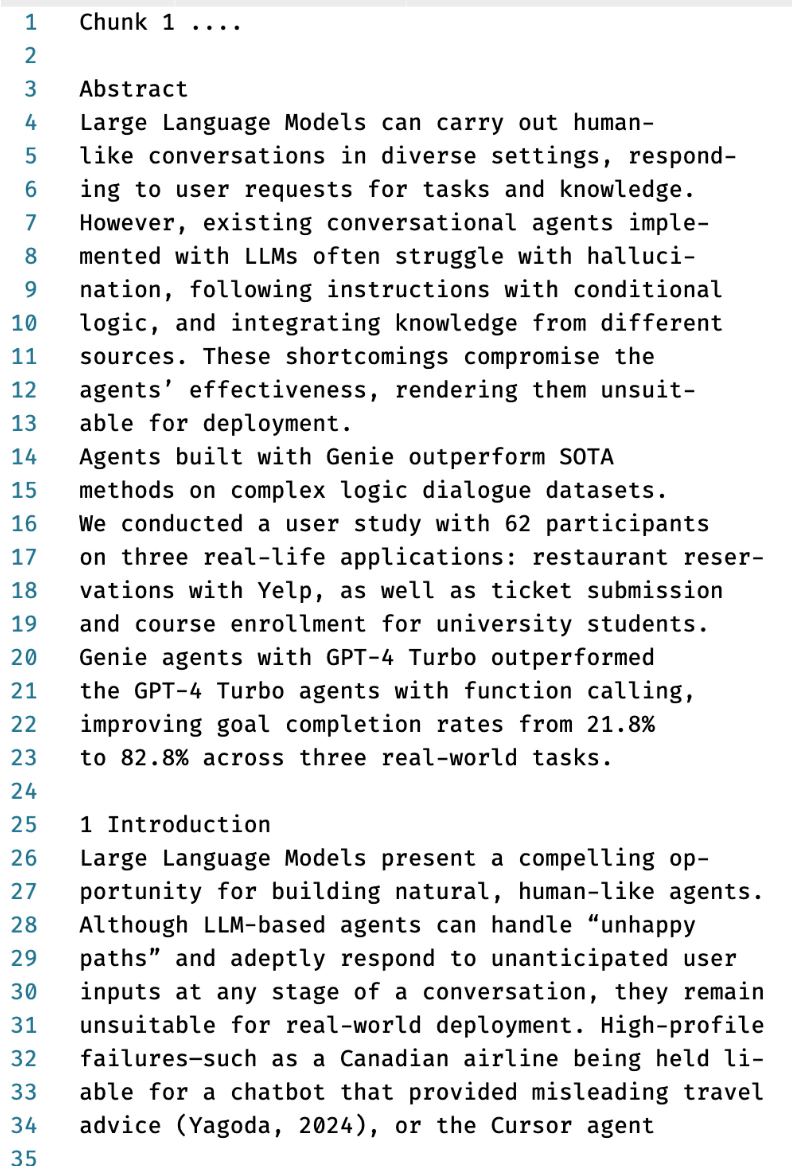
DB Execution

Final Answer

How Should We Chunk?

- For embedding-based search, chunks have fixed sizes
- Would this work for QA on long documents?

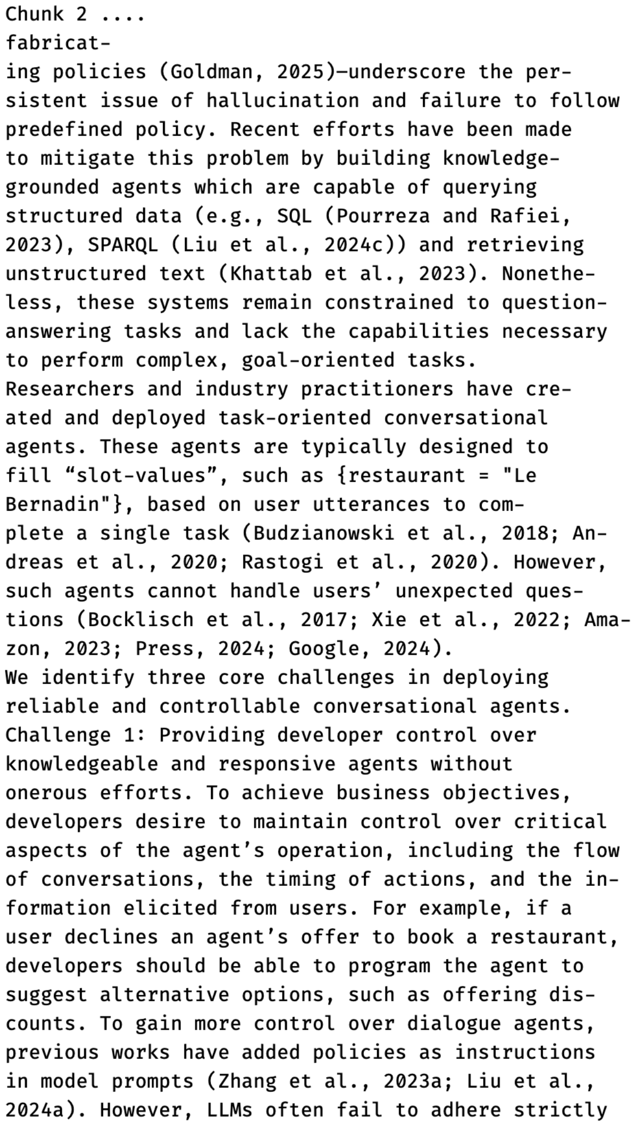
Fixed-Size Chunking Example



```
Chunk 1 ....

Abstract
Large Language Models can carry out human-
like conversations in diverse settings, respond-
ing to user requests for tasks and knowledge.
However, existing conversational agents imple-
mented with LLMs often struggle with halluci-
nation, following instructions with conditional
logic, and integrating knowledge from different
sources. These shortcomings compromise the
agents' effectiveness, rendering them unsuit-
able for deployment.
Agents built with Genie outperform SOTA
methods on complex logic dialogue datasets.
We conducted a user study with 62 participants
on three real-life applications: restaurant reser-
vations with Yelp, as well as ticket submission
and course enrollment for university students.
Genie agents with GPT-4 Turbo outperformed
the GPT-4 Turbo agents with function calling,
improving goal completion rates from 21.8%
to 82.8% across three real-world tasks.

1 Introduction
Large Language Models present a compelling op-
portunity for building natural, human-like agents.
Although LLM-based agents can handle "unhappy
paths" and adeptly respond to unanticipated user
inputs at any stage of a conversation, they remain
unsuitable for real-world deployment. High-profile
failures—such as a Canadian airline being held li-
able for a chatbot that provided misleading travel
advice (Yagoda, 2024), or the Cursor agent
```

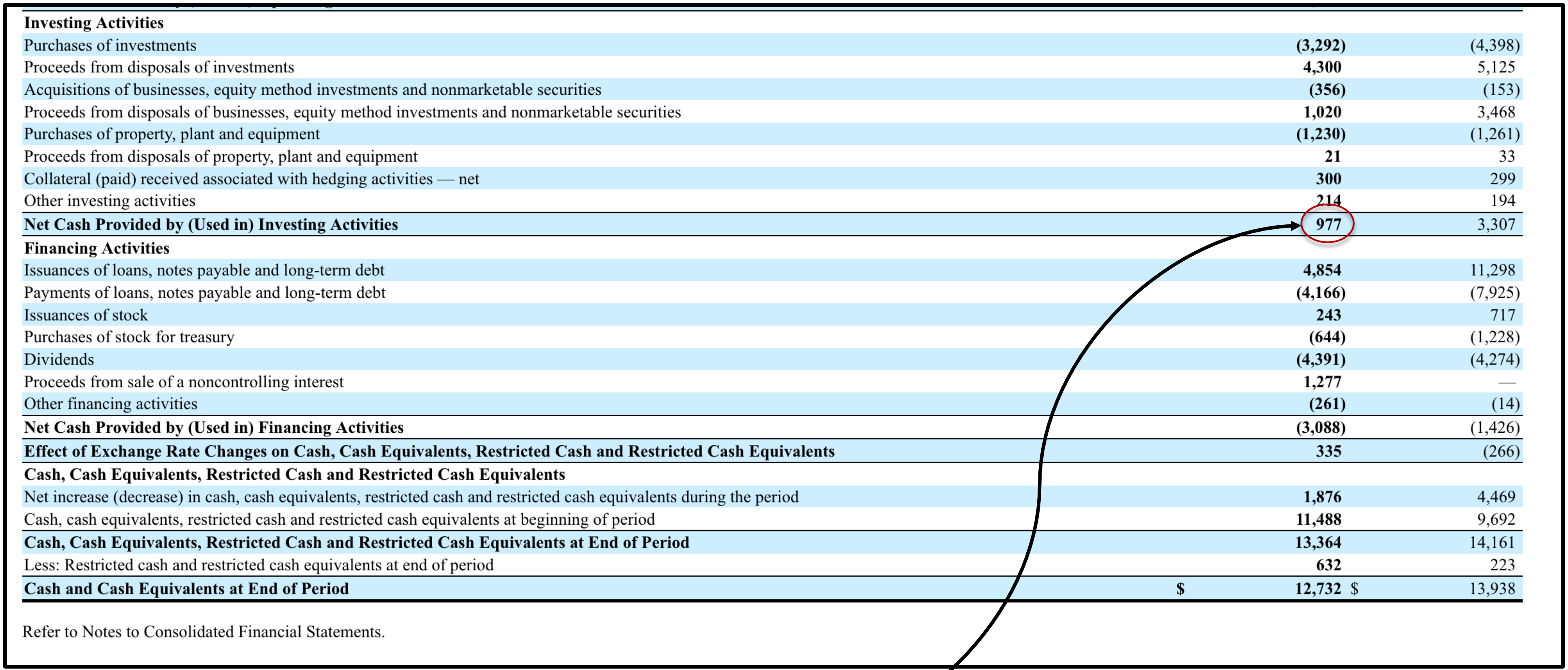


```
Chunk 2 ....
fabricat-
ing policies (Goldman, 2025)—underscore the per-
sistent issue of hallucination and failure to follow
predefined policy. Recent efforts have been made
to mitigate this problem by building knowledge-
grounded agents which are capable of querying
structured data (e.g., SQL (Pourreza and Rafiei,
2023), SPARQL (Liu et al., 2024c)) and retrieving
unstructured text (Khattab et al., 2023). Nonethe-
less, these systems remain constrained to question-
answering tasks and lack the capabilities necessary
to perform complex, goal-oriented tasks.
Researchers and industry practitioners have cre-
ated and deployed task-oriented conversational
agents. These agents are typically designed to
fill "slot-values", such as {restaurant = "Le
Bernadin"}, based on user utterances to com-
plete a single task (Budzianowski et al., 2018; An-
dreas et al., 2020; Rastogi et al., 2020). However,
such agents cannot handle users' unexpected ques-
tions (Bocklisch et al., 2017; Xie et al., 2022; Ama-
zon, 2023; Press, 2024; Google, 2024).
We identify three core challenges in deploying
reliable and controllable conversational agents.
Challenge 1: Providing developer control over
knowledgeable and responsive agents without
onerous efforts. To achieve business objectives,
developers desire to maintain control over critical
aspects of the agent's operation, including the flow
of conversations, the timing of actions, and the in-
formation elicited from users. For example, if a
user declines an agent's offer to book a restaurant,
developers should be able to program the agent to
suggest alternative options, such as offering dis-
counts. To gain more control over dialogue agents,
previous works have added policies as instructions
in model prompts (Zhang et al., 2023a; Liu et al.,
2024a). However, LLMs often fail to adhere strictly
```

Chunk 1:
- No document title (which paper is this?)
- In introduction, last sentence is arbitrarily cutoff.

Chunk 2:
- No document title
- Which section is this chunk part of?
- Incomplete sentences

Chunk *n*



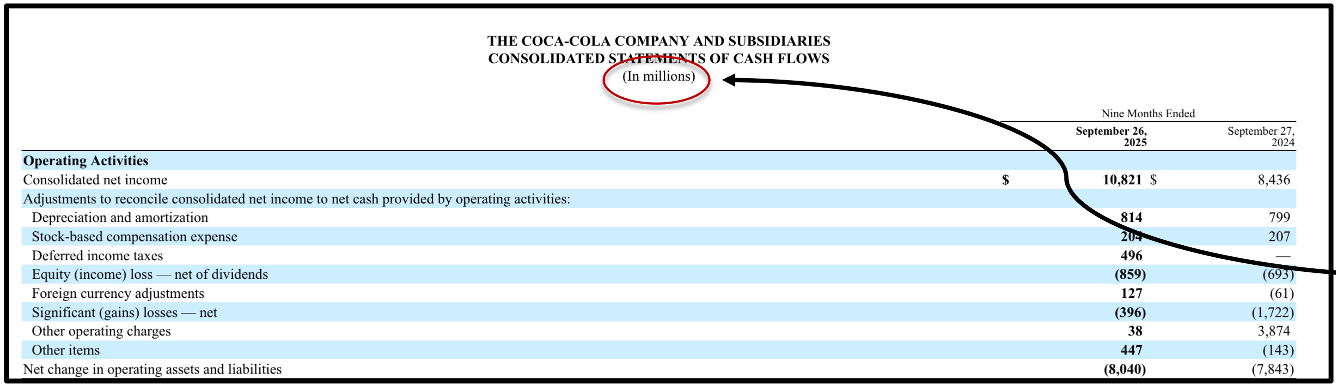
Investing Activities		
Purchases of investments	**(3,292)**	(4,398)
Proceeds from disposals of investments	**4,300**	5,125
Acquisitions of businesses, equity method investments and nonmarketable securities	**(356)**	(153)
Proceeds from disposals of businesses, equity method investments and nonmarketable securities	**1,020**	3,468
Purchases of property, plant and equipment	**(1,230)**	(1,261)
Proceeds from disposals of property, plant and equipment	**21**	33
Collateral (paid) received associated with hedging activities — net	**300**	299
Other investing activities	**214**	194
Net Cash Provided by (Used in) Investing Activities	**977**	3,307
Financing Activities		
Issuances of loans, notes payable and long-term debt	**4,854**	11,298
Payments of loans, notes payable and long-term debt	**(4,166)**	(7,925)
Issuances of stock	**243**	717
Purchases of stock for treasury	**(644)**	(1,228)
Dividends	**(4,391)**	(4,274)
Proceeds from sale of a noncontrolling interest	**1,277**	—
Other financing activities	**(261)**	(14)
Net Cash Provided by (Used in) Financing Activities	**(3,088)**	(1,426)
Effect of Exchange Rate Changes on Cash, Cash Equivalents, Restricted Cash and Restricted Cash Equivalents	**335**	(266)
Cash, Cash Equivalents, Restricted Cash and Restricted Cash Equivalents		
Net increase (decrease) in cash, cash equivalents, restricted cash and restricted cash equivalents during the period	**1,876**	4,469
Cash, cash equivalents, restricted cash and restricted cash equivalents at beginning of period	**11,488**	9,692
Cash, Cash Equivalents, Restricted Cash and Restricted Cash Equivalents at End of Period	**13,364**	14,161
Less: Restricted cash and restricted cash equivalents at end of period	**632**	223
Cash and Cash Equivalents at End of Period	**$ 12,732**	$ 13,938

Refer to Notes to Consolidated Financial Statements.

What is the Net Cash Provided by Investing Activities?

$977?

Chunk *n* - 1

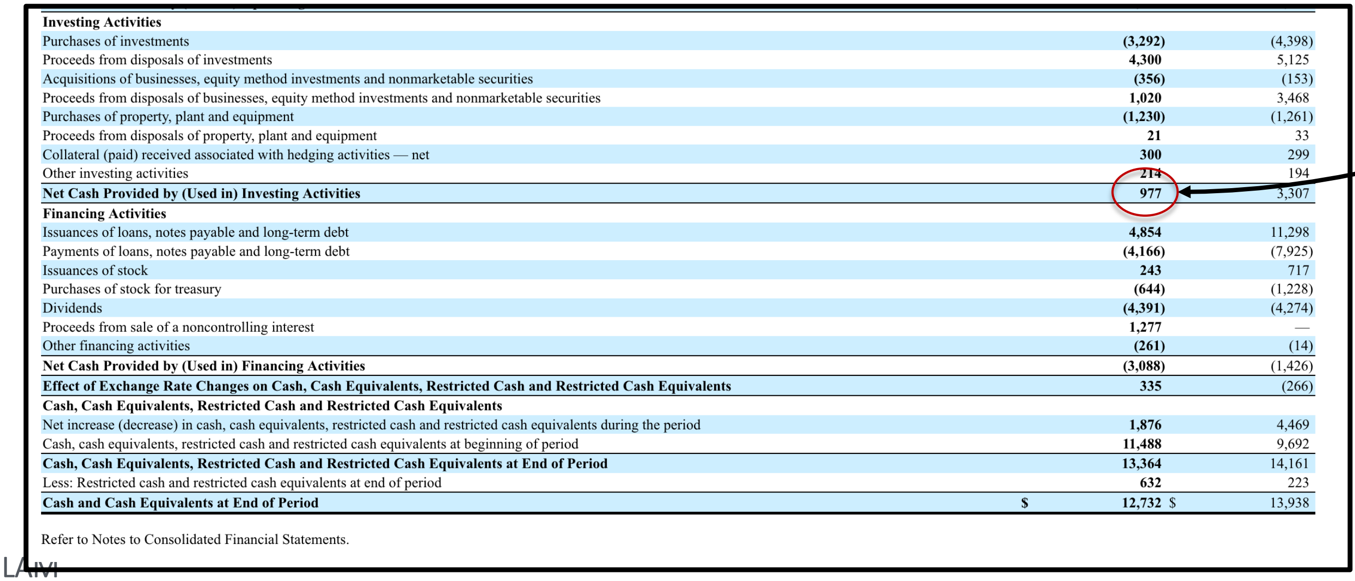


THE COCA-COLA COMPANY AND SUBSIDIARIES
CONSOLIDATED STATEMENTS OF CASH FLOWS
(In millions)

	Nine Months Ended	
	September 26, 2025	September 27, 2024
Operating Activities		
Consolidated net income	**$ 10,821**	$ 8,436
Adjustments to reconcile consolidated net income to net cash provided by operating activities:		
Depreciation and amortization	**814**	799
Stock-based compensation expense	**204**	207
Deferred income taxes	**496**	—
Equity (income) loss — net of dividends	**(859)**	(693)
Foreign currency adjustments	**127**	(61)
Significant (gains) losses — net	**(396)**	(1,722)
Other operating charges	**38**	3,874
Other items	**447**	(143)
Net change in operating assets and liabilities	**(8,040)**	(7,843)

What is the Net Cash Provided by Investing Activities?

Chunk *n*



Investing Activities		
Purchases of investments	**(3,292)**	(4,398)
Proceeds from disposals of investments	**4,300**	5,125
Acquisitions of businesses, equity method investments and nonmarketable securities	**(356)**	(153)
Proceeds from disposals of businesses, equity method investments and nonmarketable securities	**1,020**	3,468
Purchases of property, plant and equipment	**(1,230)**	(1,261)
Proceeds from disposals of property, plant and equipment	**21**	33
Collateral (paid) received associated with hedging activities — net	**300**	299
Other investing activities	**214**	194
Net Cash Provided by (Used in) Investing Activities	**977**	3,307
Financing Activities		
Issuances of loans, notes payable and long-term debt	**4,854**	11,298
Payments of loans, notes payable and long-term debt	**(4,166)**	(7,925)
Issuances of stock	**243**	717
Purchases of stock for treasury	**(644)**	(1,228)
Dividends	**(4,391)**	(4,274)
Proceeds from sale of a noncontrolling interest	**1,277**	—
Other financing activities	**(261)**	(14)
Net Cash Provided by (Used in) Financing Activities	**(3,088)**	(1,426)
Effect of Exchange Rate Changes on Cash, Cash Equivalents, Restricted Cash and Restricted Cash Equivalents	**335**	(266)
Cash, Cash Equivalents, Restricted Cash and Restricted Cash Equivalents		
Net increase (decrease) in cash, cash equivalents, restricted cash and restricted cash equivalents during the period	**1,876**	4,469
Cash, cash equivalents, restricted cash and restricted cash equivalents at beginning of period	**11,488**	9,692
Cash, Cash Equivalents, Restricted Cash and Restricted Cash Equivalents at End of Period	**13,364**	14,161
Less: Restricted cash and restricted cash equivalents at end of period	**632**	223
Cash and Cash Equivalents at End of Period	**$ 12,732**	$ 13,938

Refer to Notes to Consolidated Financial Statements.

$977 millions

If tables are split up, we lose key information, such as units

Chunking Problem and Solution

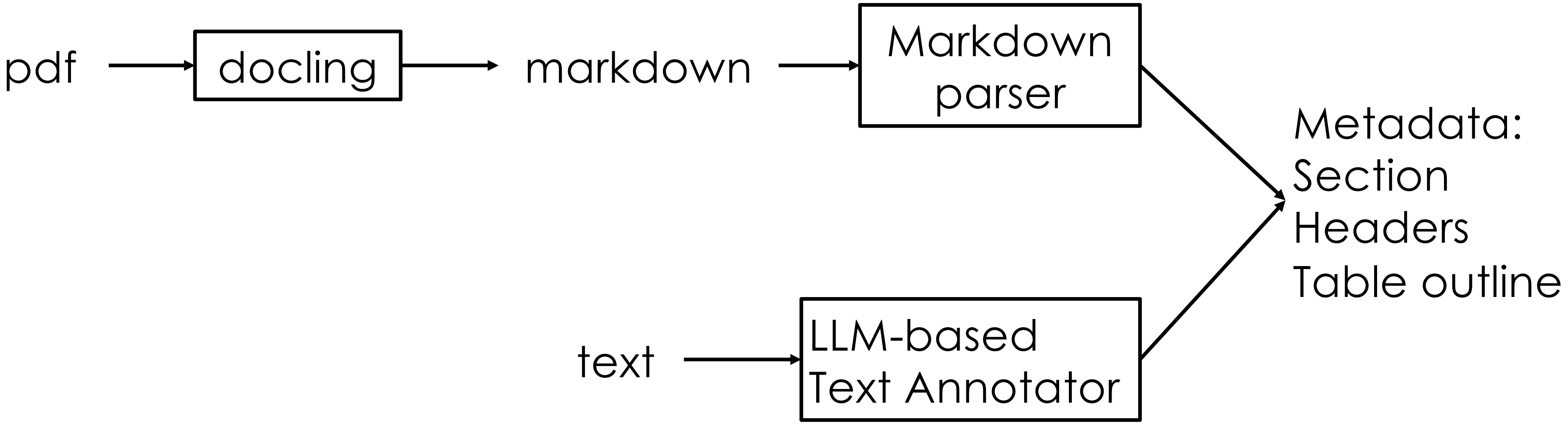
Boundary artifacts

Logical units (tables and paragraphs) may be split across chunks
→ fragmentation and a loss of coherence.

Solution: Semantics-Driven chunking
→ to create self-contained chunks for question-answering

Semantics-Driven Chunking

No paragraphs and tables are split.



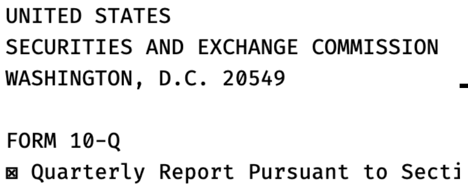
pdf → docling → markdown → Markdown parser

text → LLM-based Text Annotator

Metadata:
Section
Headers
Table outline

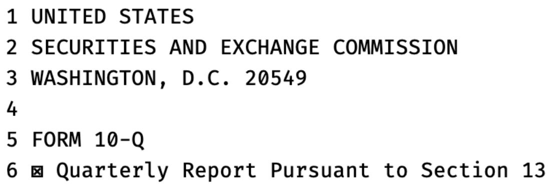
Docling: open-source toolkit to translates PDF files to markdown

LLM-Based Text Annotator

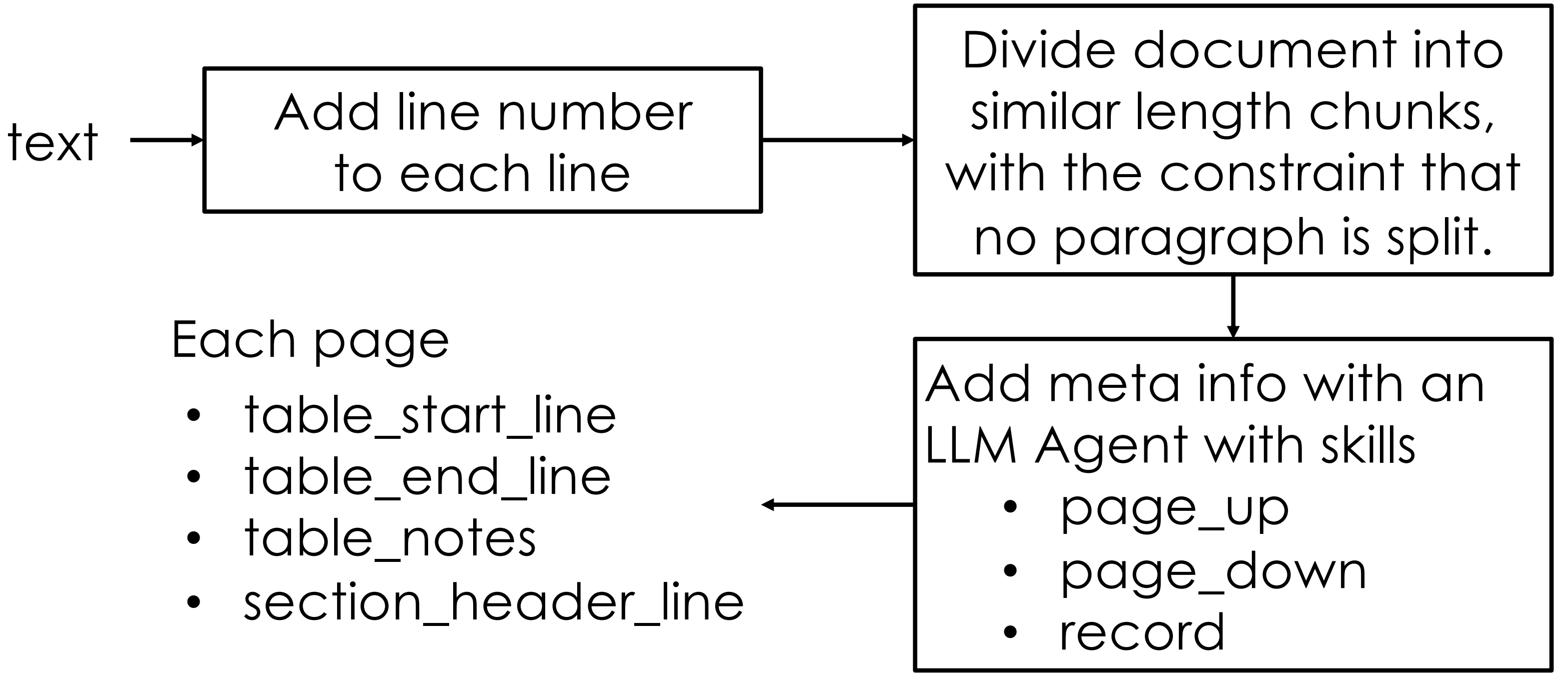


```
UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
WASHINGTON, D.C. 20549

FORM 10-Q
⊠ Quarterly Report Pursuant to Secti
```



```
1 UNITED STATES
2 SECURITIES AND EXCHANGE COMMISSION
3 WASHINGTON, D.C. 20549
4
5 FORM 10-Q
6 ⊠ Quarterly Report Pursuant to Section 13
```



text

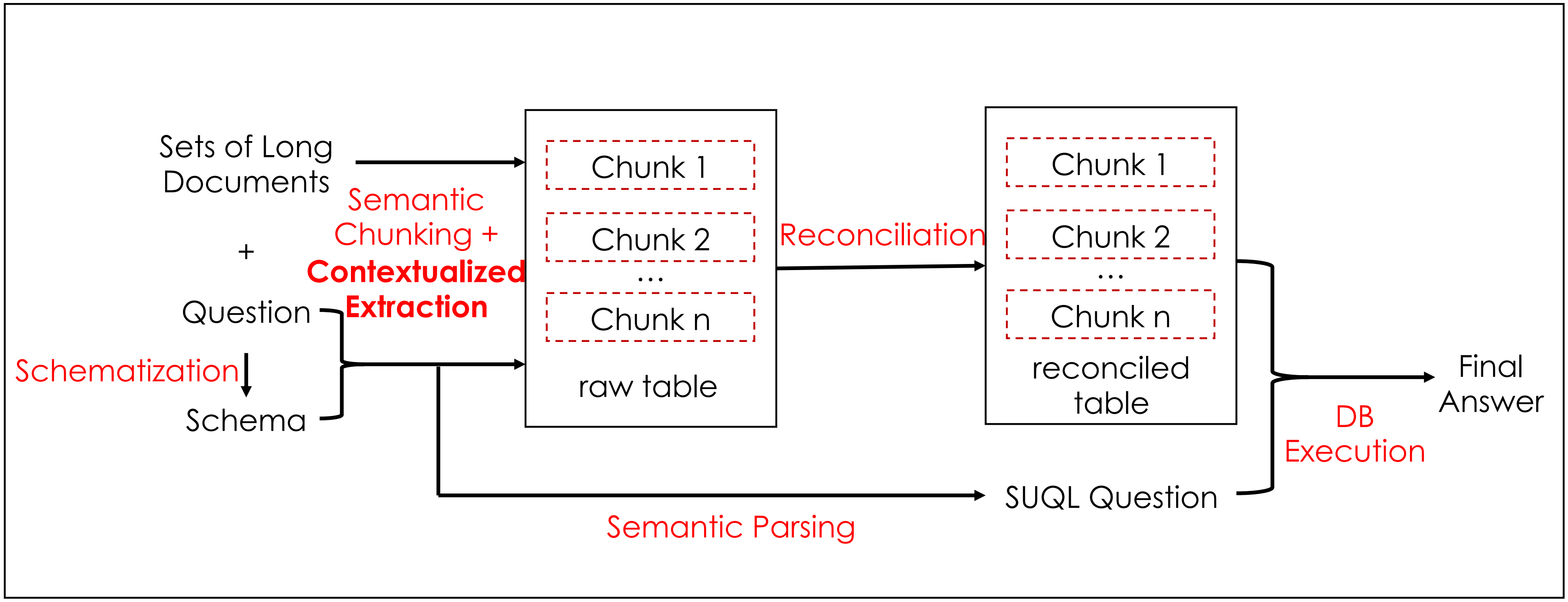
Add line number to each line

Divide document into similar length chunks, with the constraint that no paragraph is split.

Add meta info with an LLM Agent with skills
- page_up
- page_down
- record

Each page
- table_start_line
- table_end_line
- table_notes
- section_header_line

SLIDERS



Sets of Long Documents

+

Question

Schematization

Schema

Semantic Chunking +
Contextualized Extraction

Chunk 1
Chunk 2
…
Chunk n
raw table

Reconciliation

Chunk 1
Chunk 2
…
Chunk n
reconciled table

Semantic Parsing

SUQL Question

DB Execution

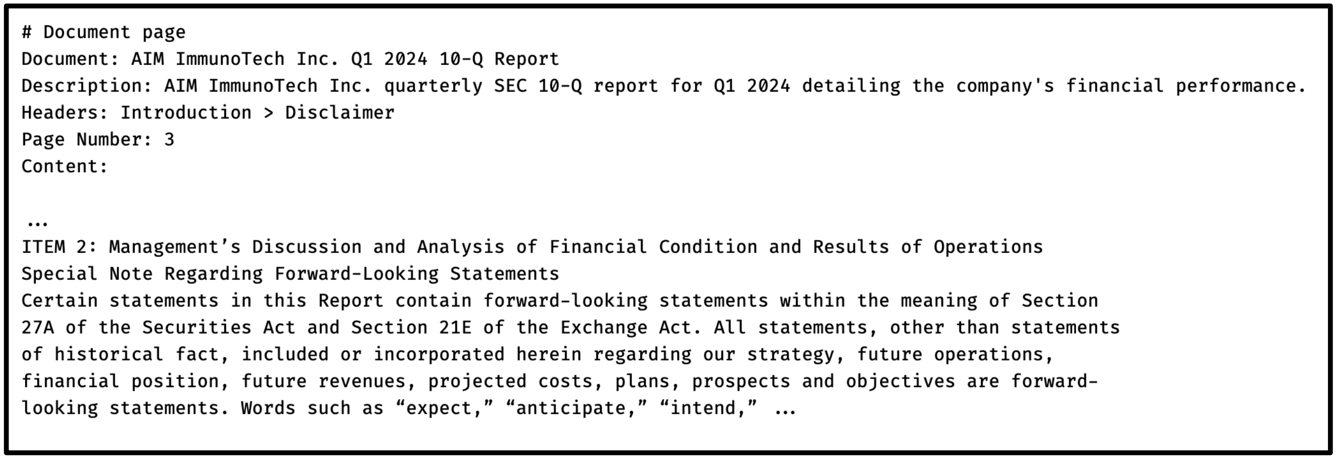
Final Answer

Contextualized Extraction Motivation

- Hard to interpret segments out of context
  - Title and section headings
  - Spatial layouts, …

Add to each chunk:

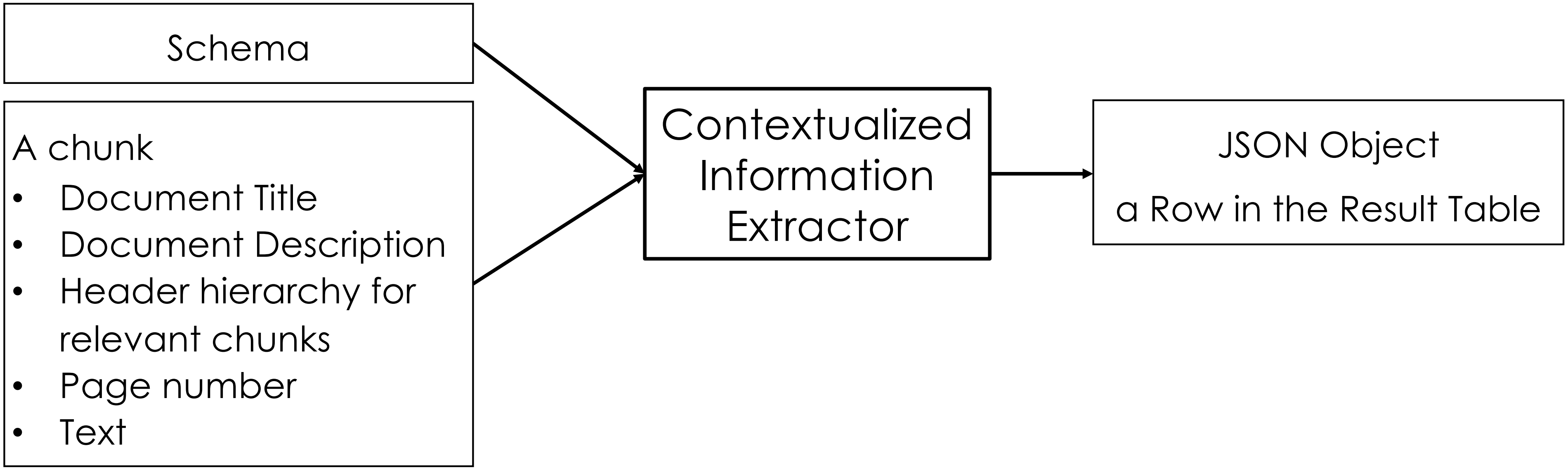
- Document Title
- Document Description
- Header hierarchy for relevant chunks
- Page number



```
# Document page
Document: AIM ImmunoTech Inc. Q1 2024 10-Q Report
Description: AIM ImmunoTech Inc. quarterly SEC 10-Q report for Q1 2024 detailing the company's financial performance.
Headers: Introduction > Disclaimer
Page Number: 3
Content:

 …
ITEM 2: Management’s Discussion and Analysis of Financial Condition and Results of Operations
Special Note Regarding Forward-Looking Statements
Certain statements in this Report contain forward-looking statements within the meaning of Section
27A of the Securities Act and Section 21E of the Exchange Act. All statements, other than statements
of historical fact, included or incorporated herein regarding our strategy, future operations,
financial position, future revenues, projected costs, plans, prospects and objectives are forward-
looking statements. Words such as “expect,” “anticipate,” “intend,”  …
```

Contextualized Extraction Overview



Schema

A chunk

- Document Title
- Document Description
- Header hierarchy for relevant chunks
- Page number
- Text

Contextualized Information Extractor

JSON Object

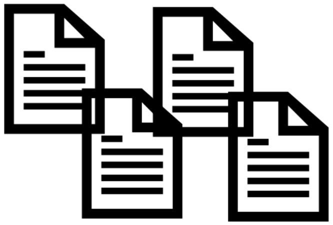
a Row in the Result Table

Example of Hallucination

We hope you will carefully study the provided papers and determine the citation relationships between them.

1. Reference: references are about what the given paper is using.
2. Citation: citations are about who is using the given paper.

The paper you need to analyze:
Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems



Given 4 academic papers

Schema:
- Source Paper Title
- Relationship Type
- Target Paper Title

LLM Generates a reference
- not in the chunk
- not one of the provided papers

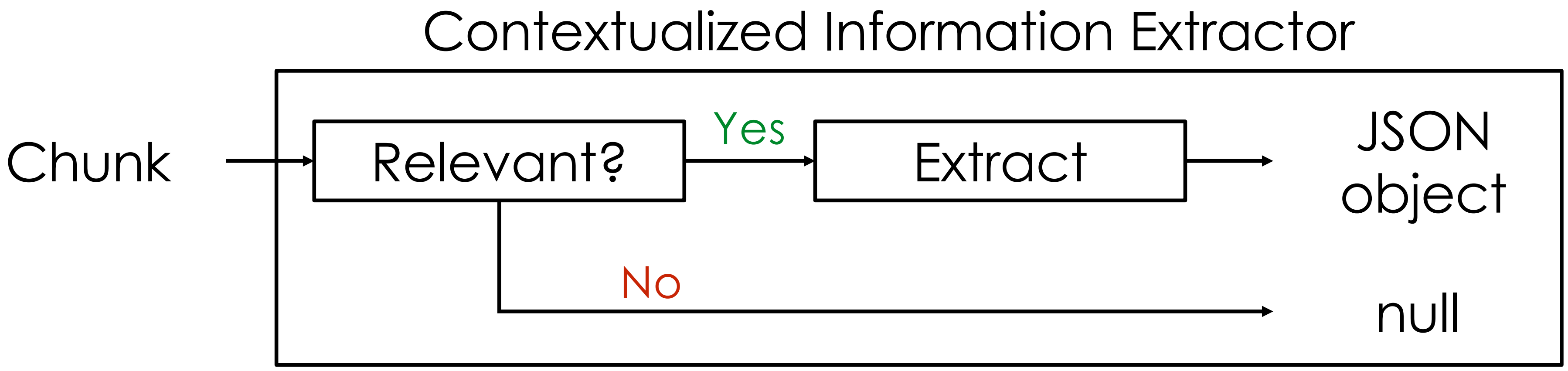
```
PaperCitationReference”: {
"fields": {
    "source_paper_title”: Risk Taxonomy, Mitigation,
and Assessment Benchmarks of Large Language
Model Systems ",
   "relationship_type": "reference",
   "target_paper_title": "Generative AI and Large
Language Models for Cyber Security: All Insights You
Need"},
    …
}
```

Contextualized Extraction Technique

- LLMs like to hallucinate
  - When given text contains no useful information (common for chunks)
  - When extracting a JSON object

→ Relevance gating.

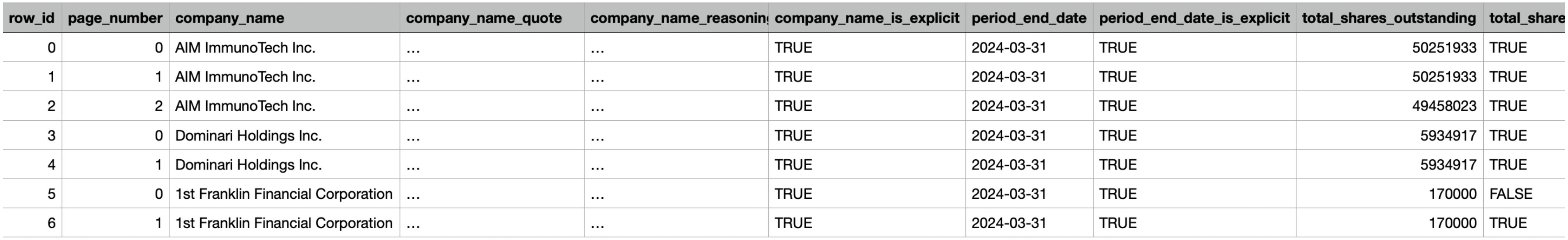
- Use a lightweight LLM to decide if a chunk is relevant



Contextualized Information Extractor

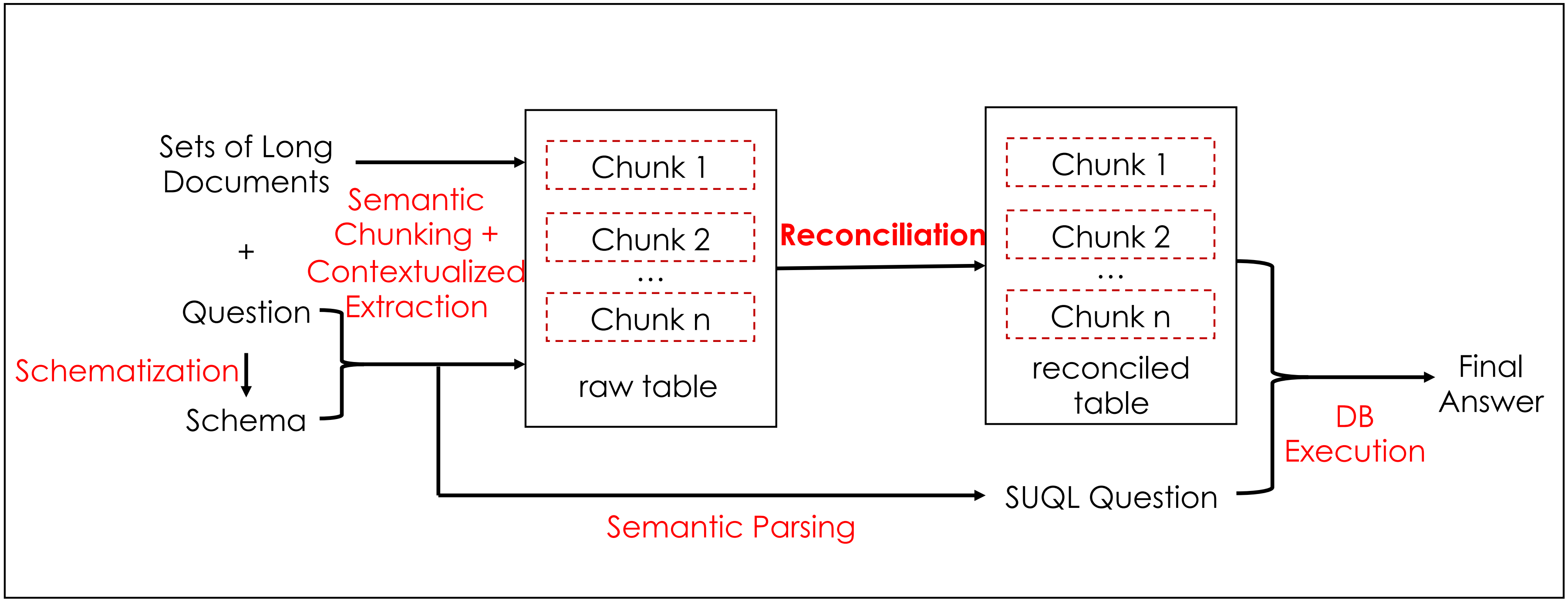
Contextualized Extraction

- To support the downstream answer-assembling task
  - JSON output for the given schema
  - Additionally:
    - Quotes from the text
    - Reasoning for selecting a value
    - Whether the information is explicitly mentioned



row_id	page_number	company_name	company_name_quote	company_name_reasonin	company_name_is_explicit	period_end_date	period_end_date_is_explicit	total_shares_outstanding	total_share
0	0	AIM ImmunoTech Inc.	...	...	TRUE	2024-03-31	TRUE	50251933	TRUE
1	1	AIM ImmunoTech Inc.	...	...	TRUE	2024-03-31	TRUE	50251933	TRUE
2	2	AIM ImmunoTech Inc.	...	...	TRUE	2024-03-31	TRUE	49458023	TRUE
3	0	Dominari Holdings Inc.	...	...	TRUE	2024-03-31	TRUE	5934917	TRUE
4	1	Dominari Holdings Inc.	...	...	TRUE	2024-03-31	TRUE	5934917	TRUE
5	0	1st Franklin Financial Corporation	...	...	TRUE	2024-03-31	TRUE	170000	FALSE
6	1	1st Franklin Financial Corporation	...	...	TRUE	2024-03-31	TRUE	170000	TRUE

SLIDERS


Sets of Long Documents
+
Question
Schematization
Schema
Semantic Chunking + Contextualized Extraction
Chunk 1
Chunk 2
...
Chunk n
raw table
Reconciliation
Chunk 1
Chunk 2
...
Chunk n
reconciled table
Semantic Parsing
SUQL Question
DB Execution
Final Answer

Why We Need Reconciliation

- Side effects from chunking and processing chunks independently
  - Overlapping information
  - Partial information
  - Conflicting information

Reconciliation Technique

For each kind of reconciliation (Overlapping, Partial, conflicting)

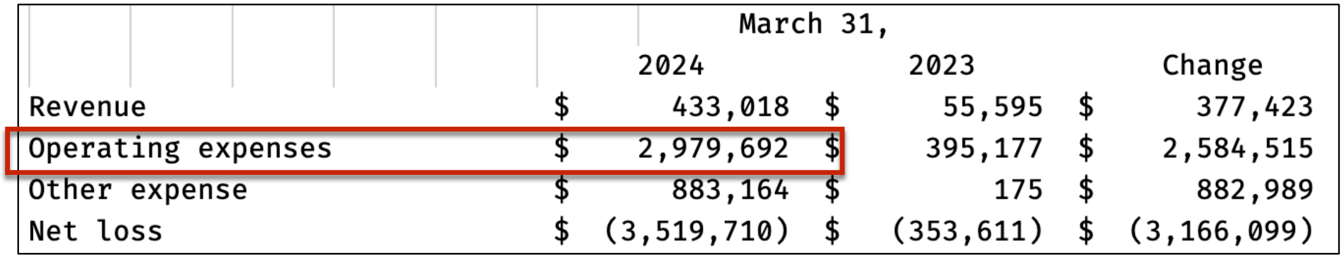
1. Reasoning: Currently use LLM to identify the issue (Needs to improve for large sets of documents).

2. Use SQL to generate the reconciled table

1. Overlapping Info

What are the Operating and Administrative Expenses?

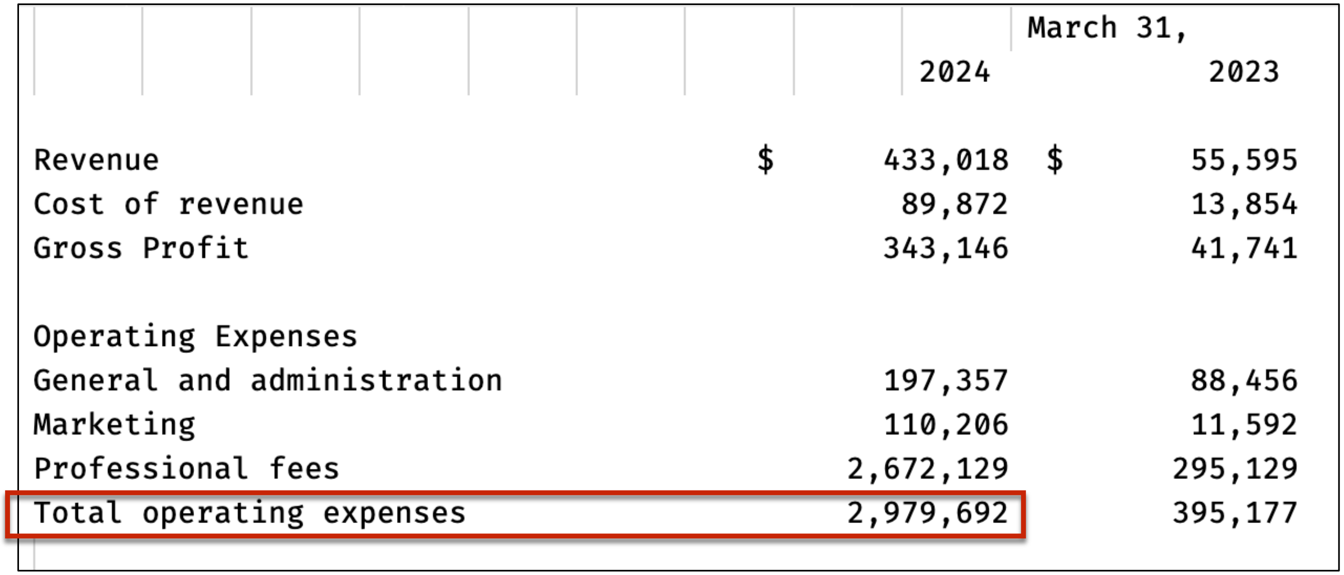
Chunk 2



	March 31, 2024		2023		Change
Revenue	$ 433,018	$	55,595	$	377,423
Operating expenses	$ 2,979,692	$	395,177	$	2,584,515
Other expense	$ 883,164	$	175	$	882,989
Net loss	$ (3,519,710)	$	(353,611)	$	(3,166,099)

2,979,692

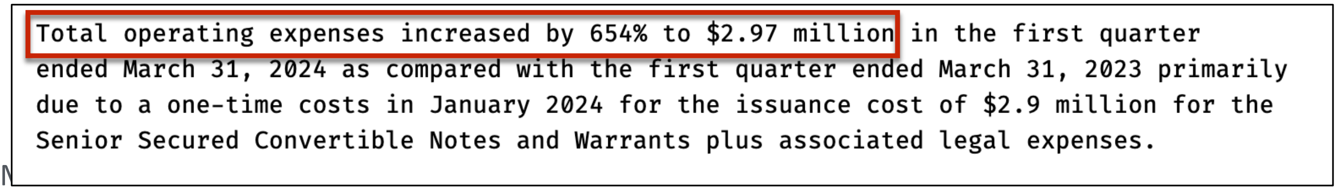
Chunk 8



		March 31, 2024		2023
Revenue	$	433,018	$	55,595
Cost of revenue		89,872		13,854
Gross Profit		343,146		41,741
Operating Expenses				
General and administration		197,357		88,456
Marketing		110,206		11,592
Professional fees		2,672,129		295,129
Total operating expenses		2,979,692		395,177

2,979,692

Chunk 11

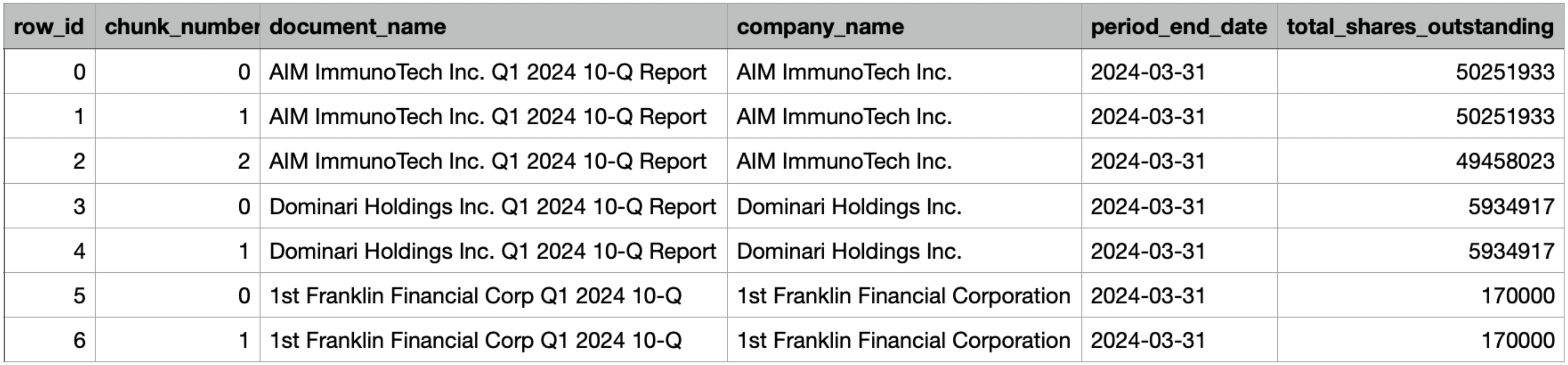


Total operating expenses increased by 654% to $2.97 million in the first quarter ended March 31, 2024 as compared with the first quarter ended March 31, 2023 primarily due to a one-time costs in January 2024 for the issuance cost of $2.9 million for the Senior Secured Convertible Notes and Warrants plus associated legal expenses.

2.97 million

What are the Operating and Administrative Expenses?

After Extraction



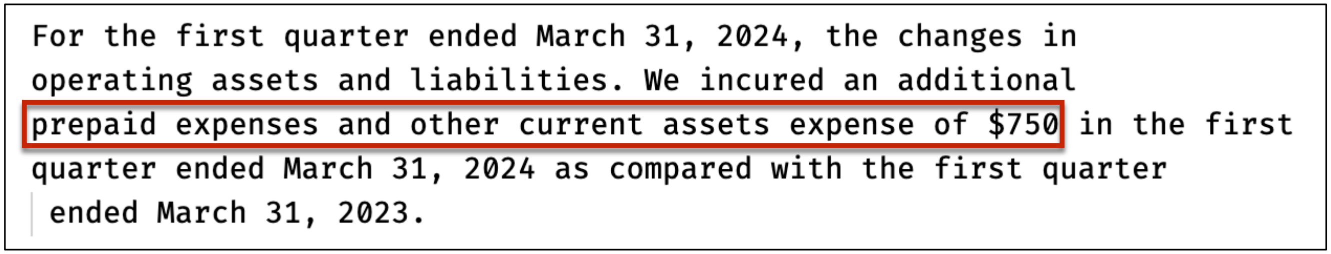
row_id	chunk_number	document_name	company_name	period_end_date	total_shares_outstanding
0	0	AIM ImmunoTech Inc. Q1 2024 10-Q Report	AIM ImmunoTech Inc.	2024-03-31	50251933
1	1	AIM ImmunoTech Inc. Q1 2024 10-Q Report	AIM ImmunoTech Inc.	2024-03-31	50251933
2	2	AIM ImmunoTech Inc. Q1 2024 10-Q Report	AIM ImmunoTech Inc.	2024-03-31	49458023
3	0	Dominari Holdings Inc. Q1 2024 10-Q Report	Dominari Holdings Inc.	2024-03-31	5934917
4	1	Dominari Holdings Inc. Q1 2024 10-Q Report	Dominari Holdings Inc.	2024-03-31	5934917
5	0	1st Franklin Financial Corp Q1 2024 10-Q	1st Franklin Financial Corporation	2024-03-31	170000
6	1	1st Franklin Financial Corp Q1 2024 10-Q	1st Franklin Financial Corporation	2024-03-31	170000

Reasoning: Deduplicate all rows

SQL: SELECT DISTINCT total_shares_outstanding, row_id, chunk_number, company_name, period_end_date, document_name FROM TableSharesOutstanding;

2. Partial Info: What is the Net Loss?

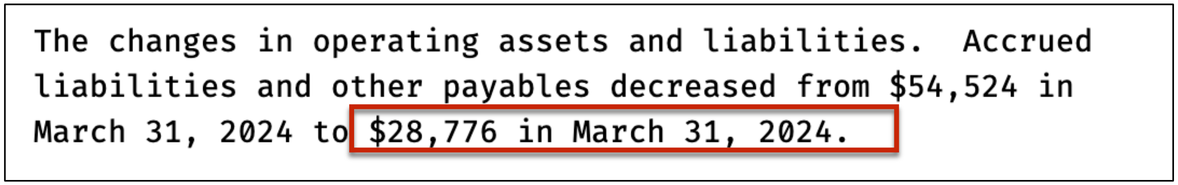
Chunk 3



For the first quarter ended March 31, 2024, the changes in operating assets and liabilities. We incured an additional prepaid expenses and other current assets expense of $750 in the first quarter ended March 31, 2024 as compared with the first quarter ended March 31, 2023.

- Prepaid expenses and other current assets: $750

Chunk 4



The changes in operating assets and liabilities. Accrued liabilities and other payables decreased from $54,524 in March 31, 2024 to $28,776 in March 31, 2024.

- Accrued liabilities and other payables: $28,776

SQL: SELECT SUM(net_loss), *
FROM NetLoss;

- Prepaid expenses and other current assets: $750
- Accrued liabilities and other payables: $28,776

Net Loss: $ 29,526

3. Conflicting Info: Where is Sandra?

Chunk 2

> At the same time he was curious to hear the end of this business. He did put on his hat and coat, however, and went out, stopping at the first convenient saloon to get a drink. Sandra moved to the kitchen. He took a car and went down to the club, strolling about the different rooms and chatting with several people whom he encountered. He was restless and irritated; and finally, after three hours of meditation, he took a cab and returned to his apartment. The distraught Jennie, sitting by her sleeping child, was at last made to realize, by its peaceful breathing that all danger was over. There was nothing more that she could do for Vesta, and now the claims of the home that she had deserted began to reassert themselves, the promise to Lester and the need of being loyal to her duties unto the very end.

Sandra is in the kitchen

Chunk 26

> It was true that Robert had not cut him out of his one-fourth of his father's estate, but certainly he had not helped him to get it, and now Robert was thinking that this offer of his might mend things. Sandra went back to the bathroom. It hurt him--Lester--a little. "I can't see it, Robert," he said finally and determinedly. "I can appreciate the motive that prompts you to make this offer. But I can't see the wisdom of my taking it. We can make all the changes you suggest if you take the stock. I'm perfectly willing to talk with you from time to time. This other thing is simply a sop with which to plaster an old wound. You want my friendship and so far as I'm concerned you have that. I don't

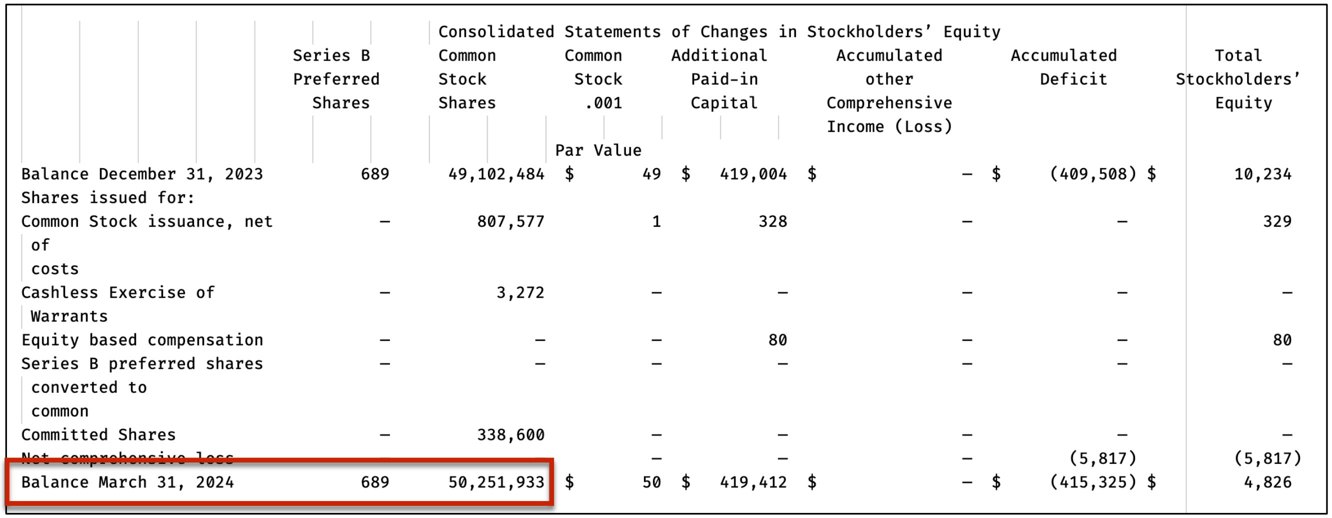
Sandra is in the bathroom

SQL: SELECT * FROM SandraLocation
ORDER BY page_number DESC LIMIT 1;

Example from babilong benchmark

Conflicting Info: What is the total # of outstanding shares?

Chunk 1

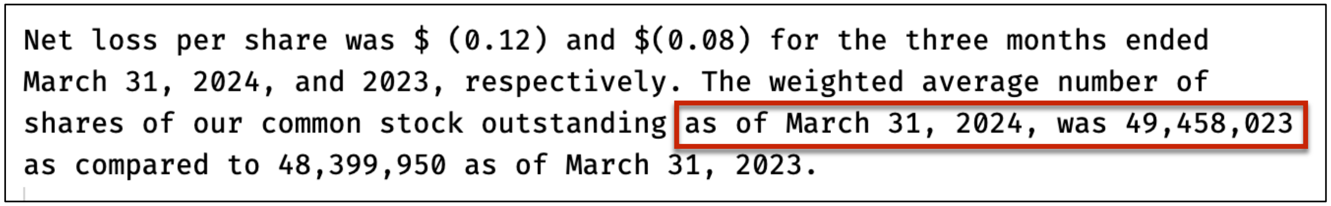


Consolidated Statements of Changes in Stockholders' Equity

	Series B Preferred Shares	Common Stock Shares	Common Stock .001 Par Value	Additional Paid-in Capital	Accumulated other Comprehensive Income (Loss)	Accumulated Deficit	Total Stockholders' Equity
Balance December 31, 2023	689	49,102,484	$ 49	$ 419,004	$ –	$ (409,508)	$ 10,234
Shares issued for:							
Common Stock issuance, net of costs	–	807,577	1	328	–	–	329
Cashless Exercise of Warrants	–	3,272	–	–	–	–	–
Equity based compensation	–	–	–	80	–	–	80
Series B preferred shares converted to common	–	–	–	–	–	–	–
Committed Shares	–	338,600	–	–	–	–	–
Net comprehensive loss	–	–	–	–	–	(5,817)	(5,817)
Balance March 31, 2024	689	50,251,933	$ 50	$ 419,412	$ –	$ (415,325)	$ 4,826

Total issued and Outstanding shares 50,251,933

Chunk 8

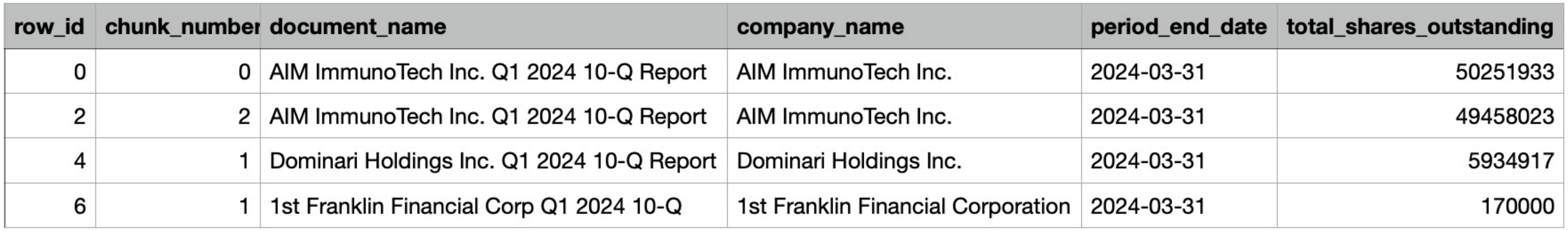


Net loss per share was $ (0.12) and $(0.08) for the three months ended March 31, 2024, and 2023, respectively. The weighted average number of shares of our common stock outstanding as of March 31, 2024, was 49,458,023 as compared to 48,399,950 as of March 31, 2023.

Weighted average number of outstanding shares 49,458,023

What is the total # of outstanding shares?

After Deduplication



row_id	chunk_number	document_name	company_name	period_end_date	total_shares_outstanding
0	0	AIM ImmunoTech Inc. Q1 2024 10-Q Report	AIM ImmunoTech Inc.	2024-03-31	50251933
2	2	AIM ImmunoTech Inc. Q1 2024 10-Q Report	AIM ImmunoTech Inc.	2024-03-31	49458023
4	1	Dominari Holdings Inc. Q1 2024 10-Q Report	Dominari Holdings Inc.	2024-03-31	5934917
6	1	1st Franklin Financial Corp Q1 2024 10-Q	1st Franklin Financial Corporation	2024-03-31	170000

From Extraction:
Total_shares_outstanding_reasoning

Note 11(b) explicitly reports outstanding shares;
The March 31, 2024 value (50,251,933) was selected.

MD&A provides the weighted average shares outstanding:
49,458,023 was recorded.

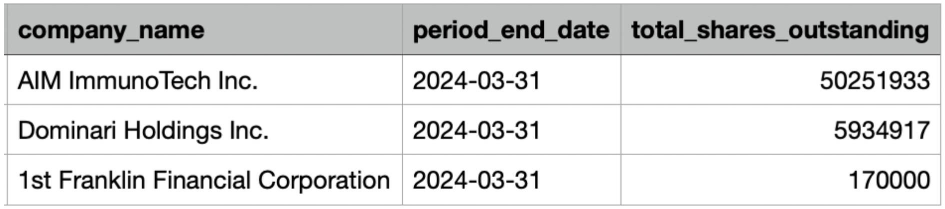
Reconciliation Reasoning:

Based on the quote and rationale,
I should keep 50251933 instead of 49458023 since
50251933 is explicitly mentioned in the balance sheet and
49458023 is given as weighted average shares

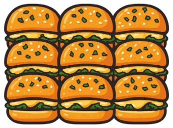
SQL:

```
SELECT * FROM TableSharesOutstanding;
WHERE row_id IN (0, 4, 6)
```

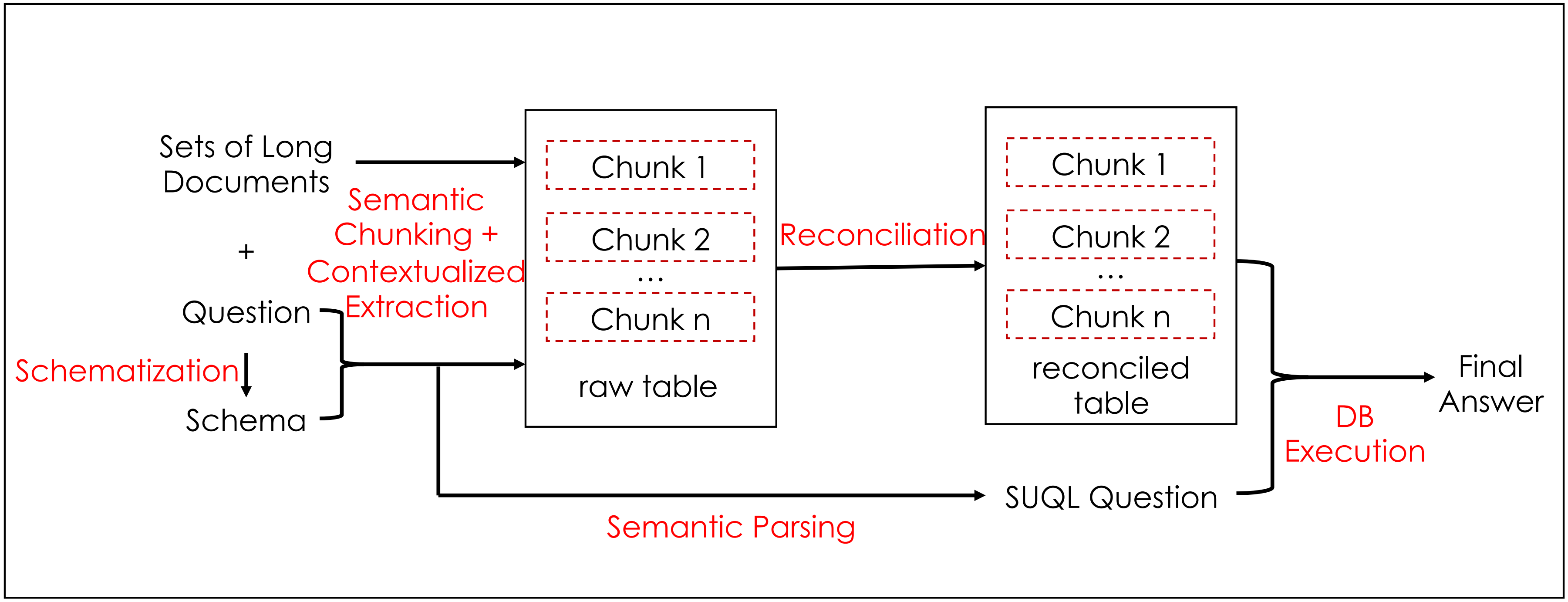
Reconciled Table



company_name	period_end_date	total_shares_outstanding
AIM ImmunoTech Inc.	2024-03-31	50251933
Dominari Holdings Inc.	2024-03-31	5934917
1st Franklin Financial Corporation	2024-03-31	170000



SLIDERS



Sets of Long Documents

+

Question

Schematization

Schema

Semantic Chunking + Contextualized Extraction

Chunk 1

Chunk 2

…

Chunk n

raw table

Reconciliation

Chunk 1

Chunk 2

…

Chunk n

reconciled table

Semantic Parsing

SUQL Question

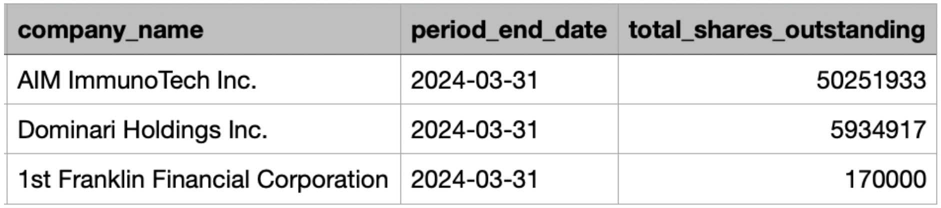
DB Execution

Final Answer

DB Query: What is the total # of outstanding shares?

Semantic-parse NL to SQL to compute the final answer ✓

Reconciled Table



company_name	period_end_date	total_shares_outstanding
AIM ImmunoTech Inc.	2024-03-31	50251933
Dominari Holdings Inc.	2024-03-31	5934917
1st Franklin Financial Corporation	2024-03-31	170000

```
SELECT company_name, total_shares_outstanding
FROM SharesOutstanding
ORDER BY total_shares_outstanding DESC LIMIT 1;
```

final answer

Let's review the data from the table:

- AIM ImmunoTech Inc. has 50,251,933 total shares outstanding as of March 31, 2024.
- 1st Franklin Financial Corporation has 170,000 total shares outstanding as of March 31, 2024.
- Dominari Holdings Inc. has 5,934,917 total shares outstanding as of March 31, 2024.

Comparing these numbers, AIM ImmunoTech Inc. has the highest total shares outstanding.

Answer:
AIM ImmunoTech Inc. has the highest 'Total Shares Outstanding' with 50,251,933 shares as of March 31, 2024.

EVALUATION

Long-Context QA Benchmarks

Documents for each question fit in the LLM context window

- FinanceBench
  - Benchmark for Financial Question Answering
  - 150 Questions – 95k tokens
- Loong Benchmark
  - Three domains: finance, academic papers, legal
  - Two languages: English and Chinese
  - Four types of questions: spotlight, comparison, clustering, chain of reasoning
  - 1600 Question – up to 250k tokens

Preliminary Evaluation

Models	**Method**	**FinanceBench**	**Loong**
GPT 4o	LLM inference	78.67	53.58
Gemini 1.5 Pro	LLM inference	-	55.36
GPT 4.1	LLM inference	84.00	77.23
SLIDERS	**chunk-based**	**90.67**	**78.34**

Note: SLIDERS can handle
large sets of long documents not fitting in LLM context

Preliminary Error Analysis

- Found several wrong gold annotations in Loong
- Found a couple of wrong annotations in FinanceBench
- Based on different interpretation of the question,
  the answers can be different.

Example of Incorrect Gold Label

AppTech Payments Corp.

Ameritek Ventures, Inc.

BATTALION OIL CORP

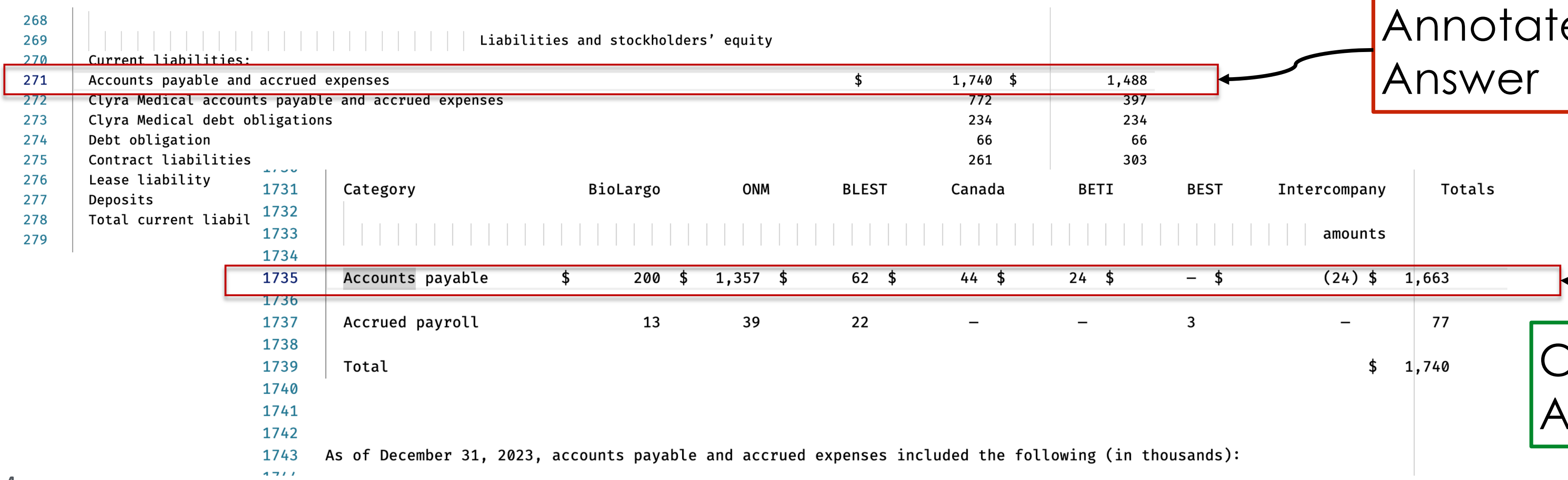
Agape ATP Corp

BIOLARGO, INC

1847 Holdings LLC

What is the accounts payable of BIOLARGO, INC.?

GPT and the Annotated answer say its 1,740



Liabilities and stockholders' equity

Current liabilities:

Accounts payable and accrued expenses	$ 1,740	$ 1,488
Clyra Medical accounts payable and accrued expenses	772	397
Clyra Medical debt obligations	234	234
Debt obligation	66	66
Contract liabilities	261	303
Lease liability		
Deposits		
Total current liabil		

Category	BioLargo	ONM	BLEST	Canada	BETI	BEST	Intercompany amounts	Totals
Accounts payable	$ 200	$ 1,357	$ 62	$ 44	$ 24	$ —	$ (24)	$ 1,663
Accrued payroll	13	39	22	—	—	3	—	77
Total								$ 1,740

As of December 31, 2023, accounts payable and accrued expenses included the following (in thousands):

Annotated Answer

Correct Answer

Code Release

- https://github.com/stanford-oval/sliders
- Step 1: Preprocess the pdfs (scripts/pdf_to_markdown.py)
- Step 2: Run Sliders
- Step 3: Evaluate the steps (step_viewer.py)

SLIDERS is work in progress

- Open Issues on the repo
- Make pull requests

Conclusions

- **Important to analyze across large sets of long documents**
  - Need to scale beyond model context limit
- **Precision degrades with document length in the context**
- Approaches
  - **Training: for precision – not successful**
  - **Chunking: for precision and scalability**
  - SLIDERS:
    - **Leverages SUQL: text → databases**
    - **Semantic chunking, contextualized extraction, reconciliation**
- Long-context benchmarks: SOTA (despite annotation errors)
- **Scale beyond long-context benchmarks**