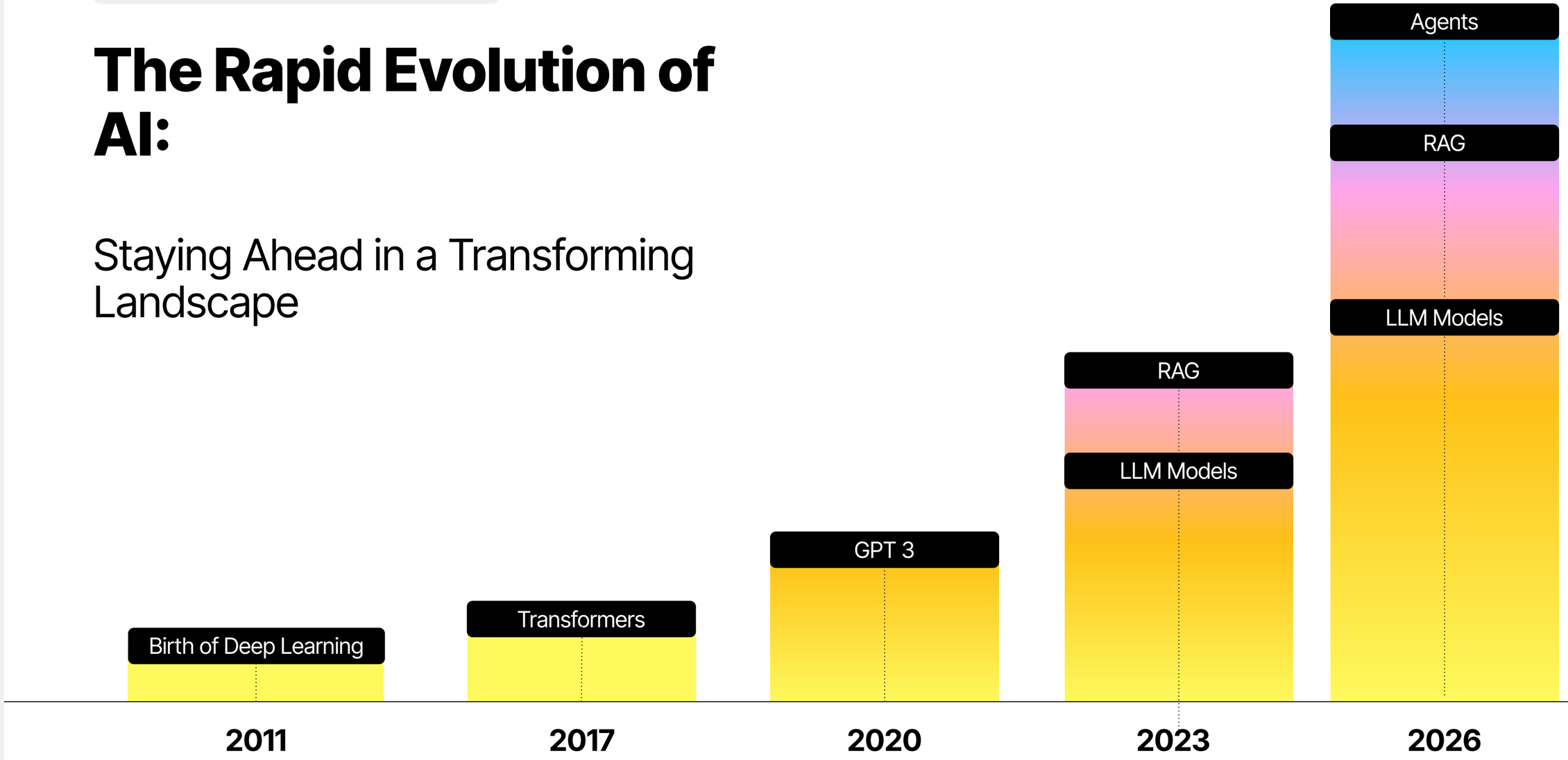

UNST
RUCT
URED
Hello!

The Rapid Evolution of AI:

Staying Ahead in a Transforming Landscape



UNSTRUCTURED

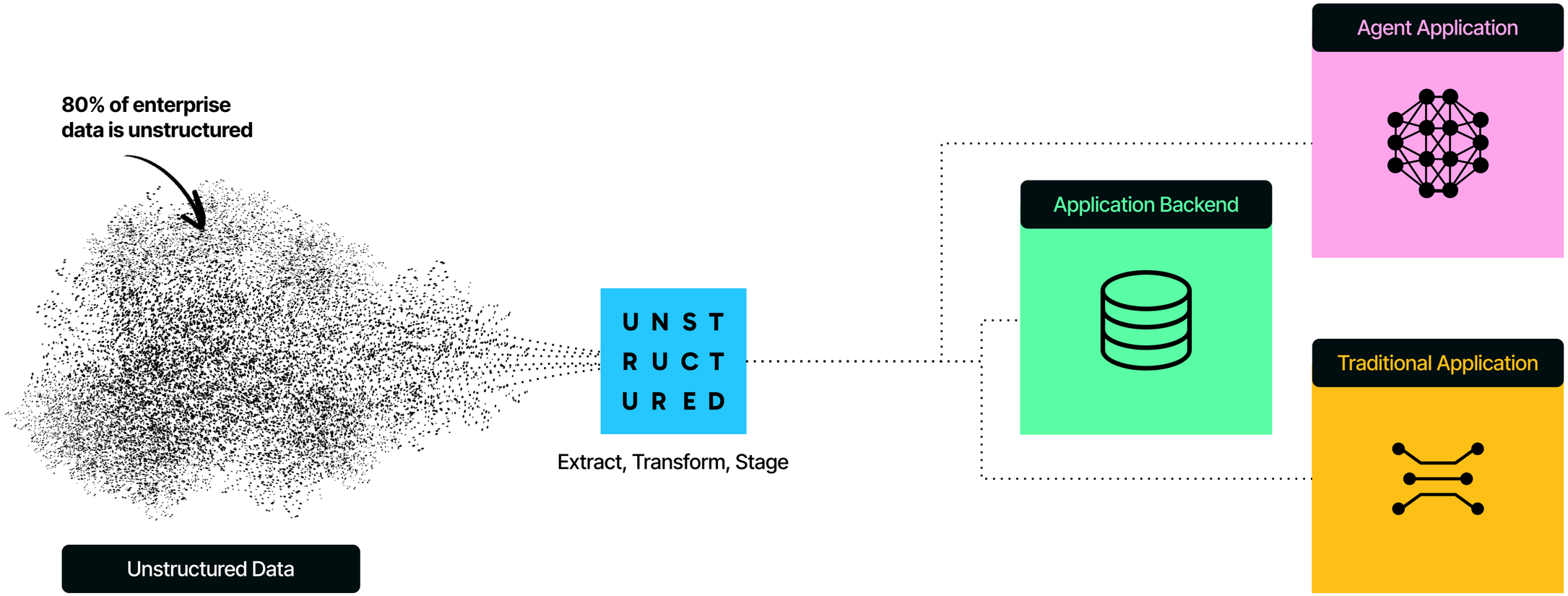
Consolidation to Use Cases with ROI

Data is the next GenAI bottleneck.

UNSTRUCTURED

Architecture

Where we fit in the tech stack.



Story time

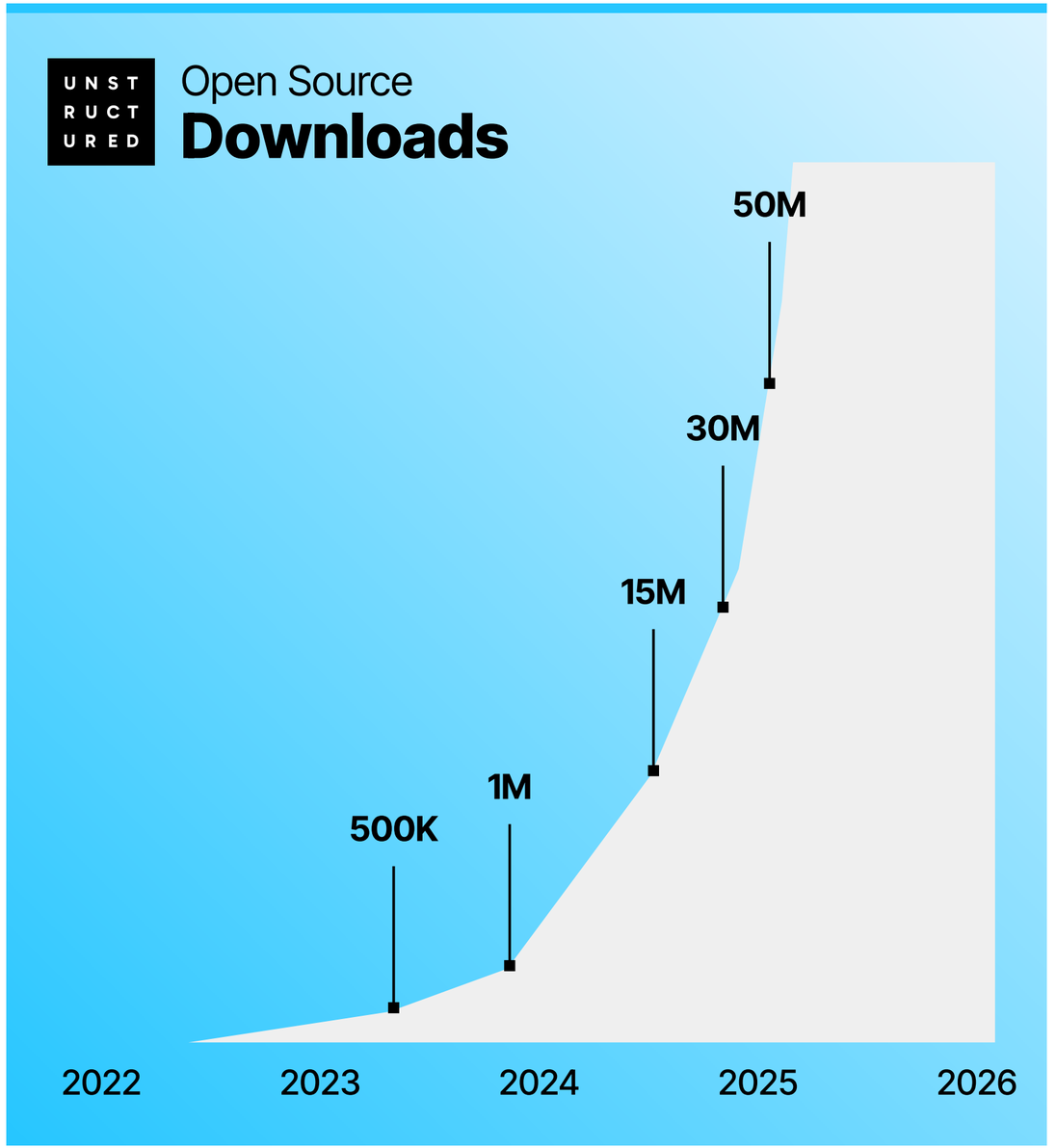
Back in 2022

We built an open source data transformation solution.

It went viral.

Thousands of Enterprises, Data Platforms, and Government organizations started using us.

Enterprises 217,000	Gov't Orgs 829	Data Platforms 81
Apple	Department of Defense	Cohesity
AT&T	Department of State	Couchbase
Boeing	Department of the Treasury	Elastic
Capital One Financial	Department of Justice	IBM
Cisco	Department of Commerce	NetApp
Dell Technologies	Commission (FCC)	Pure Storage
Goldman Sachs	Securities and Exchange	SAP
IBM	Commission (SEC)	Teradata
Intel	Federal Emergency	
Tesla	Management Agency	
The Walt Disney Company	U.S. Census Bureau	
+ more	**+ more**	**+ more**

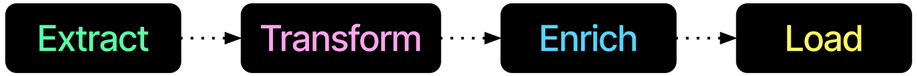


UNSTRUCTURED

Hi, We're Unstructured

The GenAI Data Company.

Enterprises turn to us when they move to production GenAI.



We make messy, unstructured enterprise data like PDFs, emails, images, videos and more, usable.

Trusted by 87% of the Fortune 1000.

Awards

FAST COMPANY
Most Innovative Companies

Forbes
Artificial Intelligence 50

CBINSIGHTS
Artificial Intelligence 100

+10,000
paying customers

+50M
product downloads

+87%
Fortune 1000 using Unstructured

$65M
raised since founding in 2022

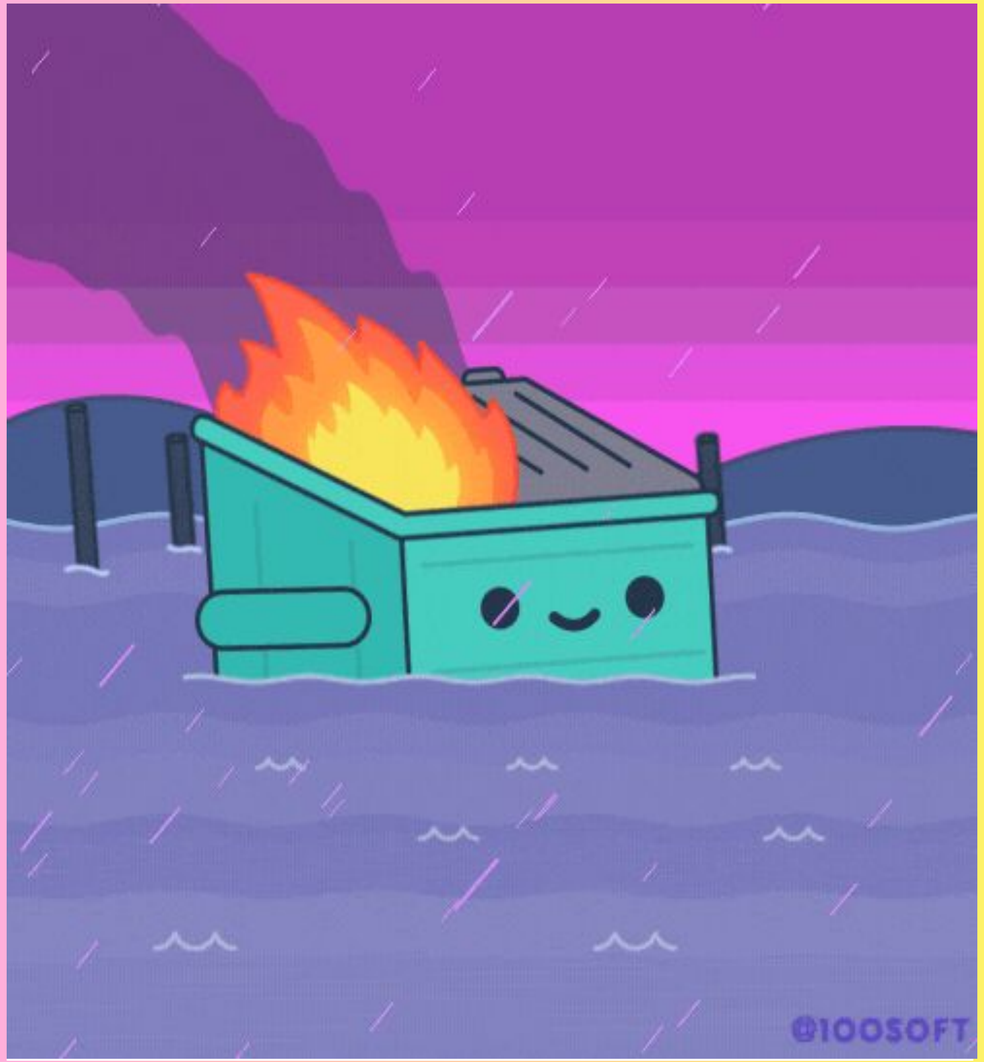
MENLO VENTURES
NVIDIA
IBM
BainCapital VENTURES
Madrona
SHIELD CAPITAL
databricks
MongoDB.
M12

UNSTRUCTURED

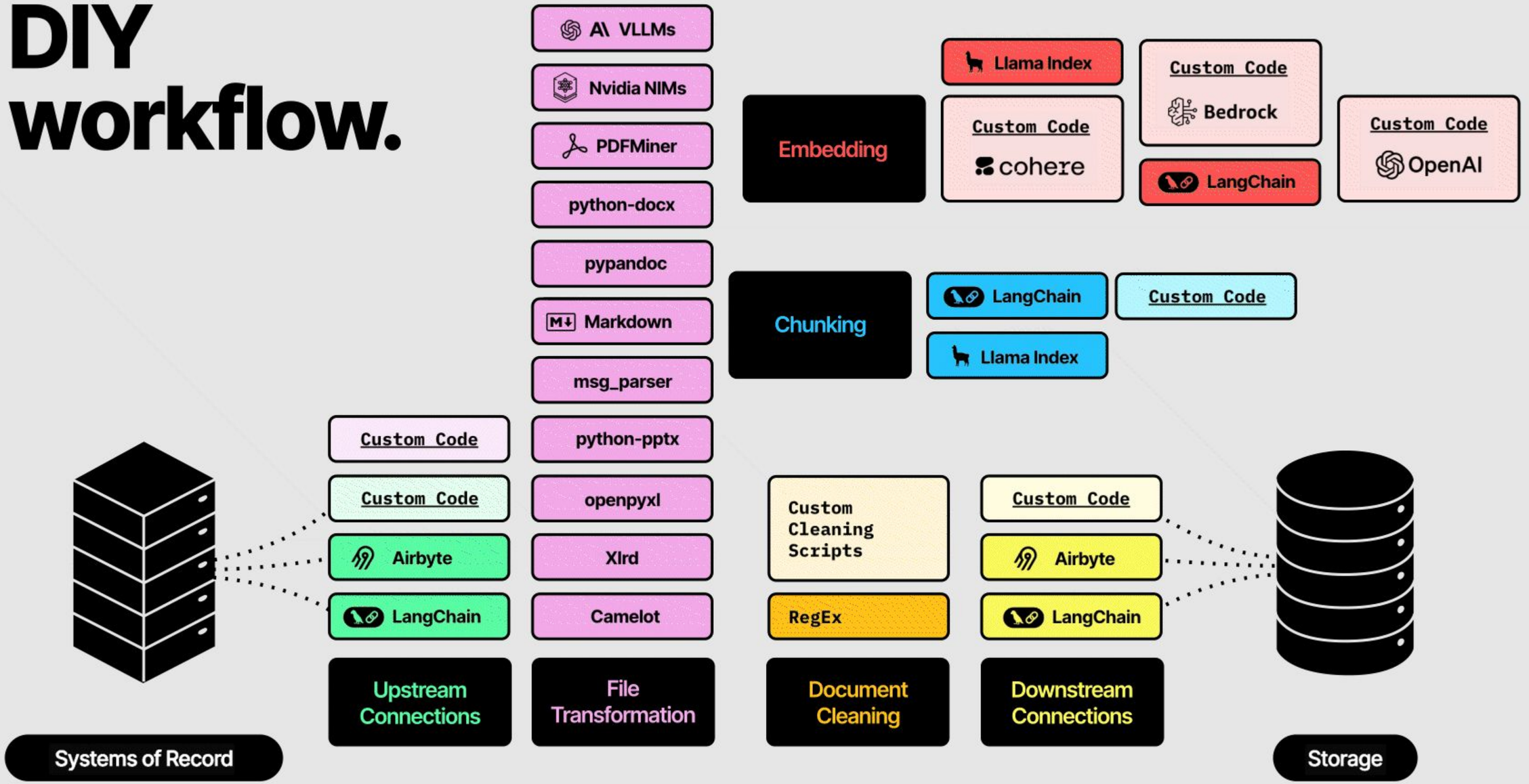
Problem Statement

Companies spent 2025 rapidly building AI into their platforms...

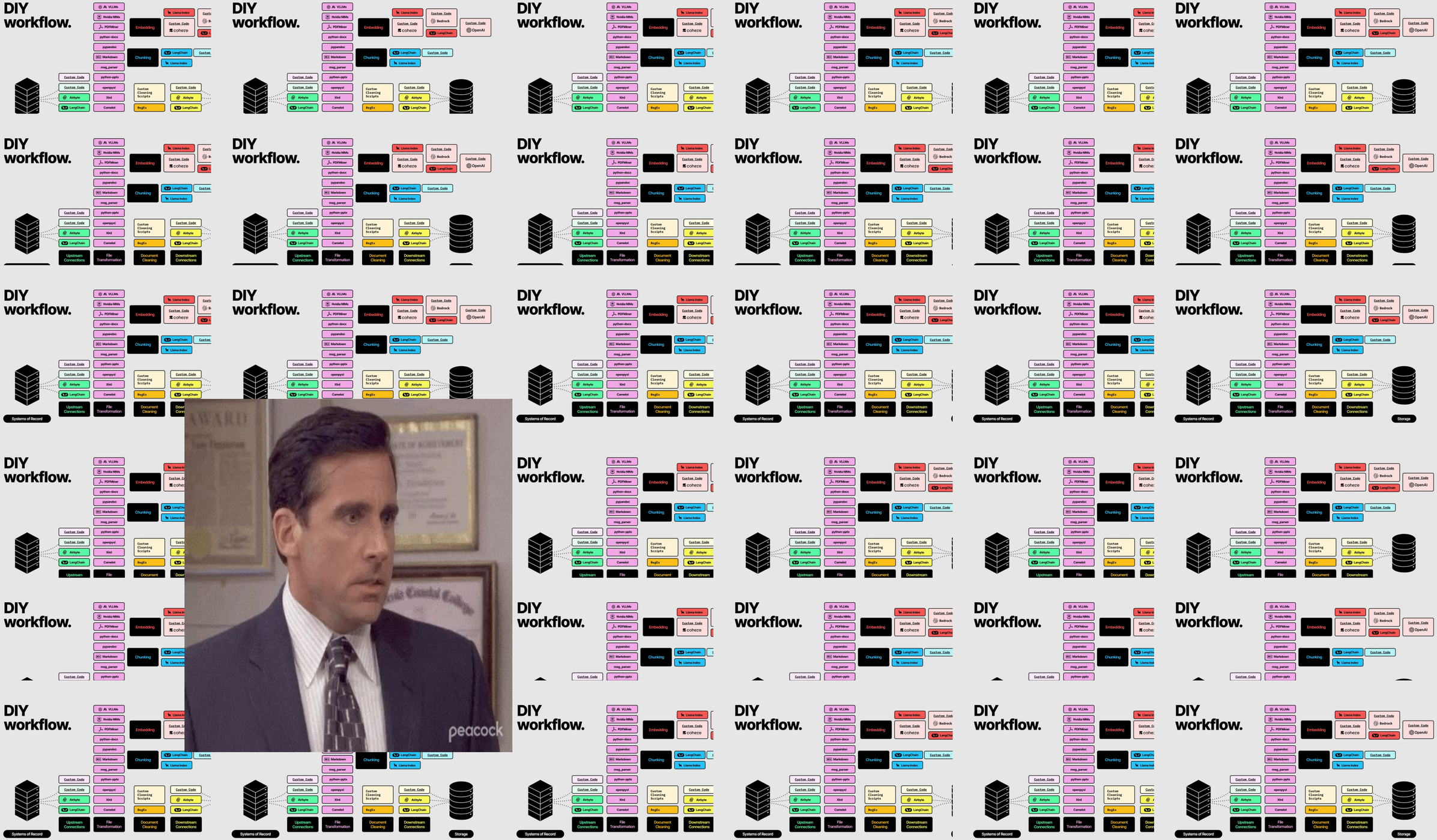
1. Prototypes aren't making it into production
2. The **OpEx** and **CapEx** to build AI products aren't showing value relative to costs
3. There is no standardization or ability to reuse pipelines across your organization



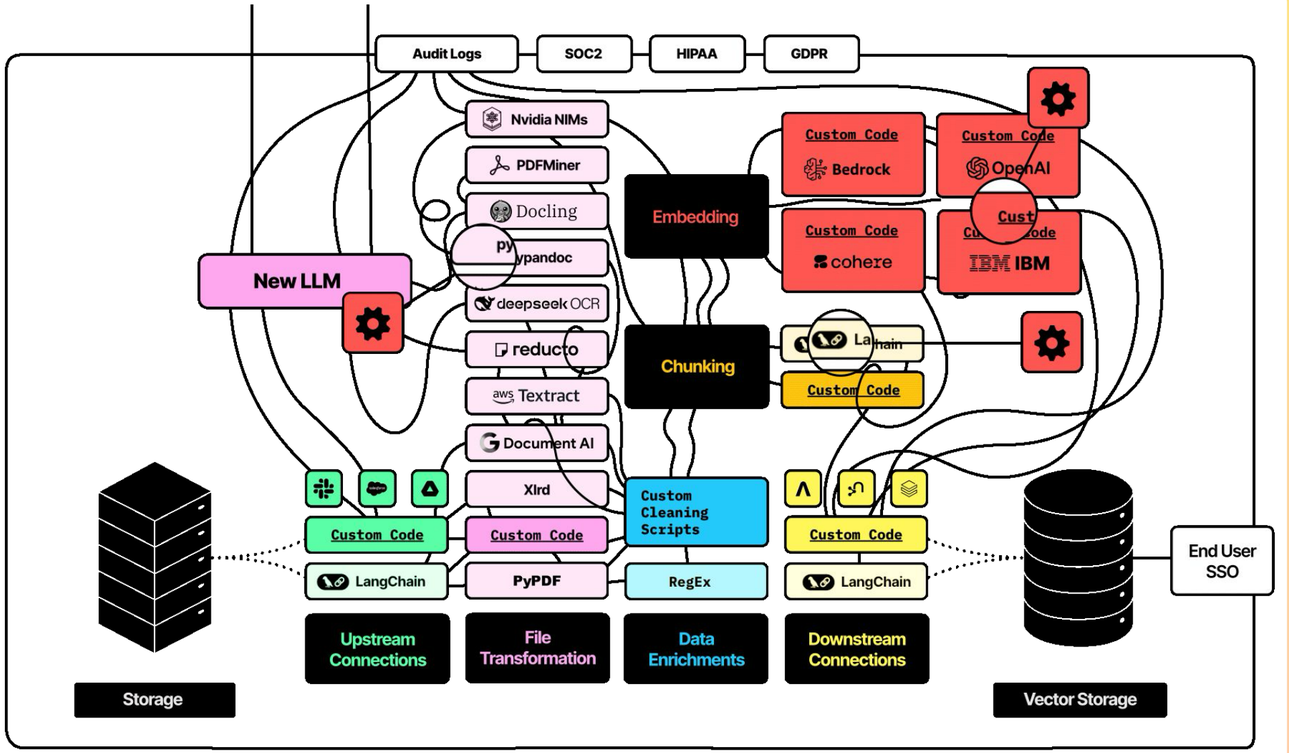
DIY workflow.



UNSTRUCTURED

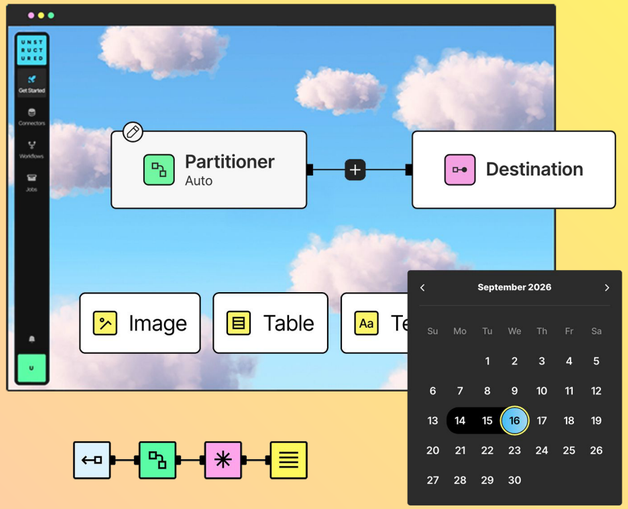


Delivering enterprise-grade GenAI data at scale.



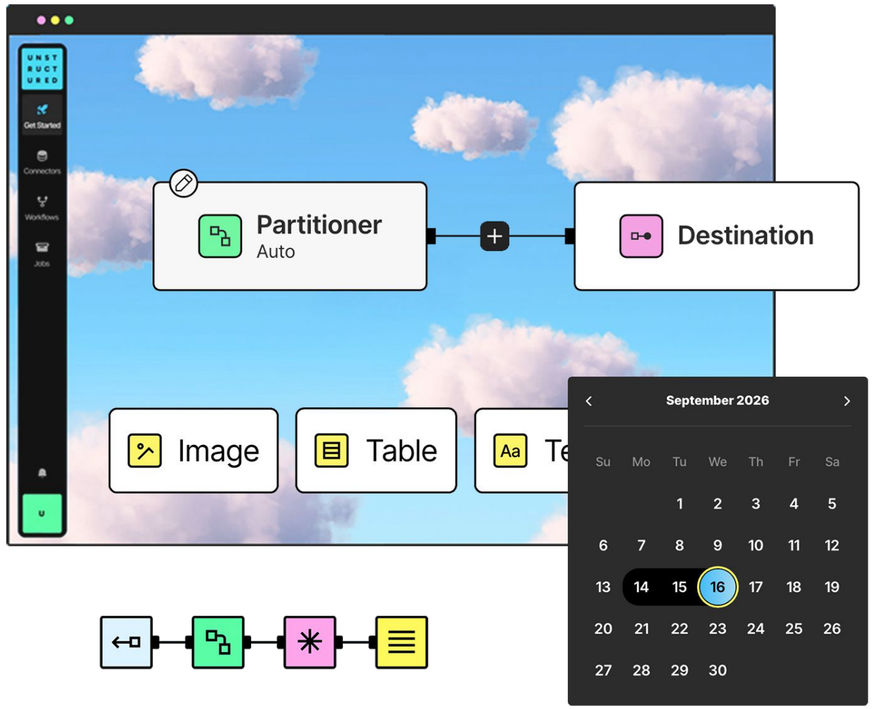
UNSTRUCTURED

Unstructured
Platform



No More “One-off” Solutions

One platform does it all.



How We Do It

✓ Connect: 35+ Connectors Supported

Bring in content from SharePoint, Confluence, Google Drive, S3, Salesforce, and more. Our connectors are secure, scalable, and built for production use.

✓ Support: 70+ File Types Supported

From PDFs and DOCX to emails and meeting recordings, Unstructured handles the formats your teams actually use.

✓ Normalize: Canonical JSON Format

All ingested content is normalized into a clean, consistent JSON schema. This makes it easy to use downstream and build on top of a stable foundation.

✓ All-in-One Data Layer

Unstructured acts as the Grand Central Station for your unstructured data. Ingest, structure, enrich, and embed all in one place.

✓ UI or API

Choose your own adventure!

UNSTRUCTURED

Document Processing:

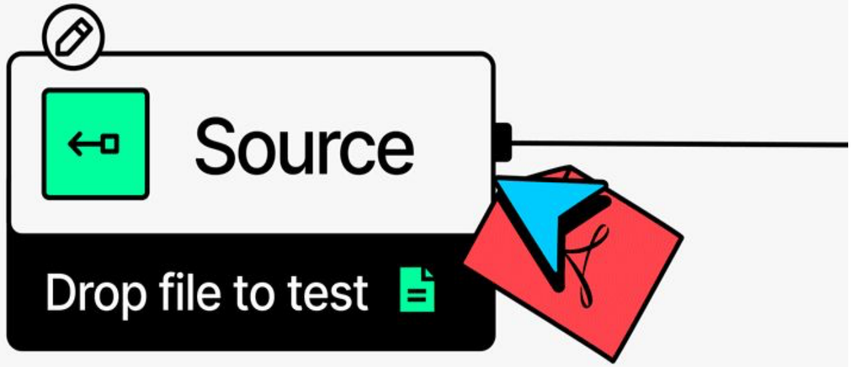
A Closer Look

UNSTRUCTURED

No More Fragmented Solutions

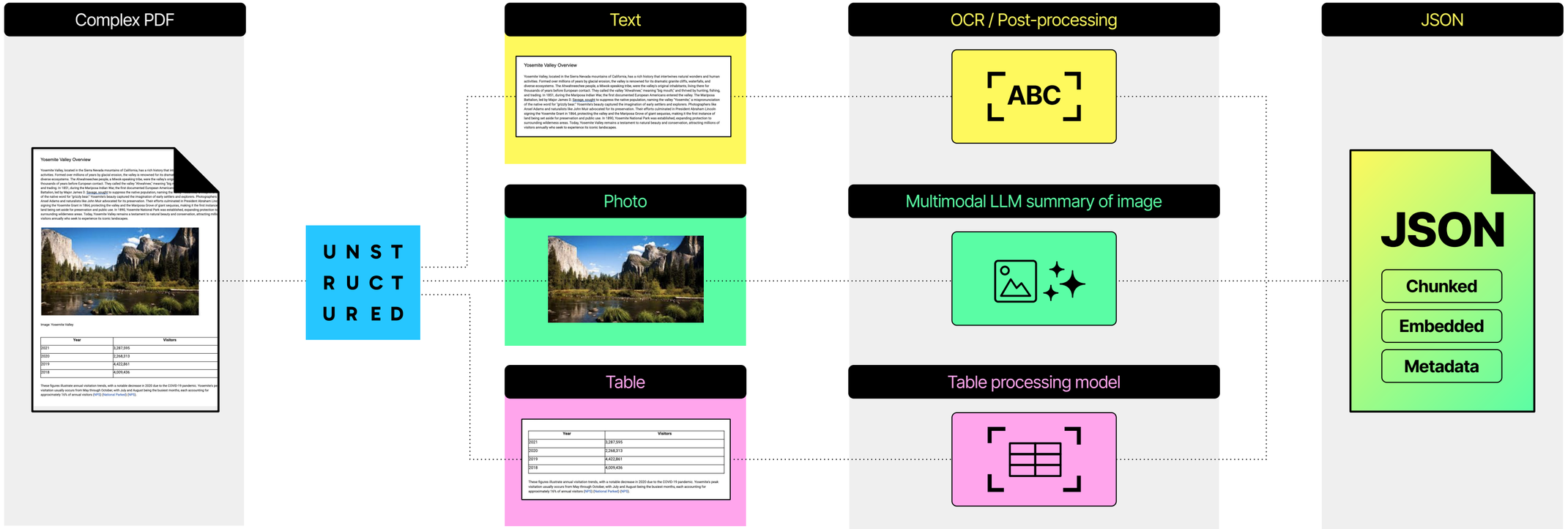
Partition, chunk, enrich, and embed 70+ file types.

Connecting to unstructured data shouldn't be complex. With 30+ built-in connectors, Unstructured pulls content from your systems of record and business applications—no custom code required. Every integration works the same way, so your data arrives clean, consistent, and ready to power your AI.



Superior Transformation

Integrates with any model for superior transformation.



Input

Classification

Runs object detection algorithm and classifies each element.

Processing

Integrates with any third-party model for optimal transformation.

Output

UNSTRUCTURED

With Unstructured: Simple, Stable, Scaleable

Ingestion & Preprocessing 2.0

Source Connectors

- Azure Blob Storage Search
- Box
- Confluence
- Couchbase
- Databricks Volumes
- Discord
- Dropbox
- ElasticSearch
- Github
- Gitlab
- Google BigQuery Source
- Google Drive
- Google GCS
- Kafka
- Loom
- Notion
- OneDrive
- OpenSearch
- Oracle
- Outlook
- Pure Storage
- S3 Source
- Salesforce
- Sharepoint
- SFTP
- Singlestore
- Slack
- Snowflake

File Types

- .CSV
- .DOCX
- .EML
- .EPUB
- .HTML
- .JPG
- .MD
- .OST
- .PDF
- .PNG
- .PST
- .RST
- .RTF
- .TSV
- .TXT
- .WAV
- .XLS
- .TIFF
- .XLSX
- .ZAVW

UNSTRUCTURED

Transform Strategies

Fast

For simple, text-only documents.

Hi-Res (OD+OCR)

For PDFs, images and complex file types.

Vision Language Model

For difficult docs including scanned and handwritten content with VLM API integration.

JSON [;]

Enrichment

- Chunked JSON
- Embeddings
- Structured Data
- Custom Integrations

JSON
Chunked
Embedded
Metadata

Custom Pipelines & Third-Party Integrations

- Amazon Bedrock
- Anthropic
- Azure AI Studio
- Azure Doc Integ's
- Gemini
- Open AI
- Together.ai
- VertexAI

Destination Connectors

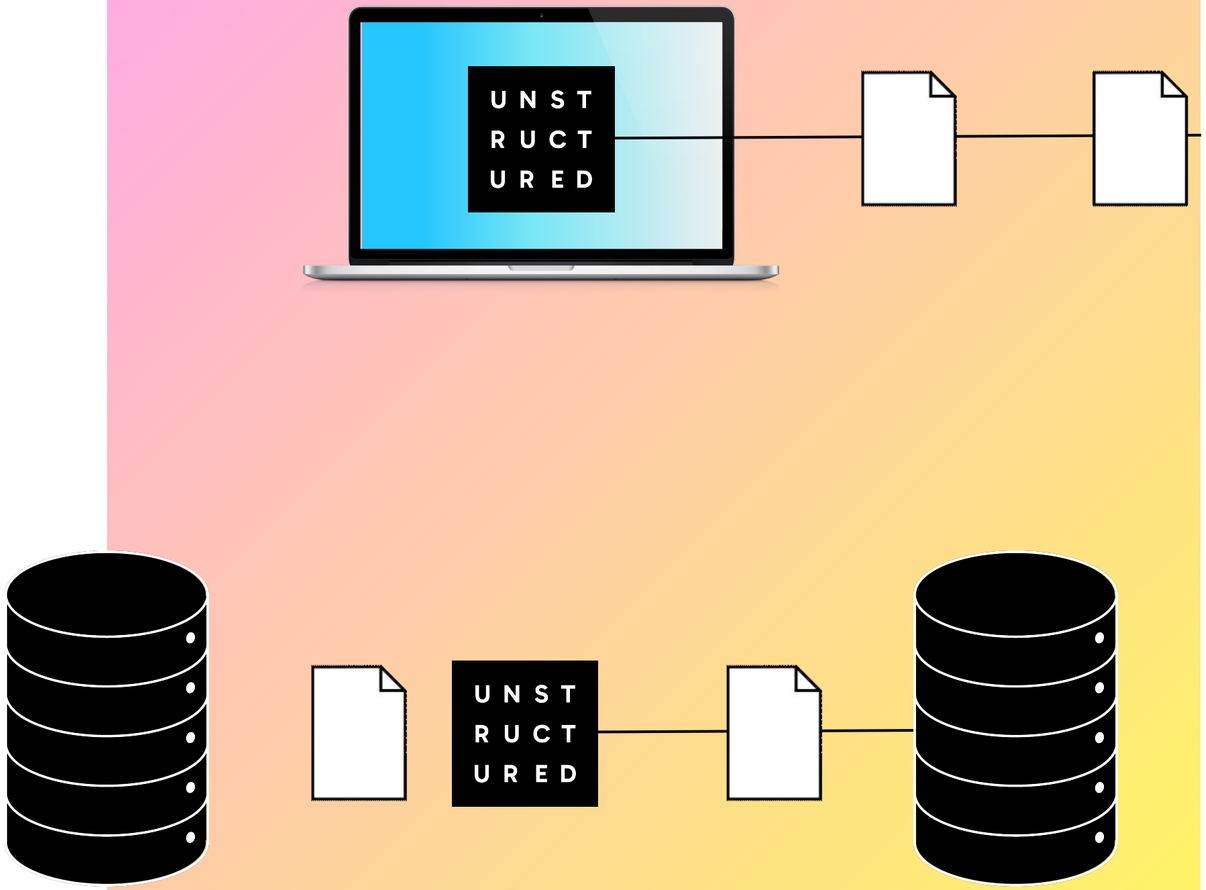
- Azure AI Search
- Chroma
- Couchbase
- Databricks Volumes
- Datastax AstraDB
- Delta Table
- **ElasticSearch**
- Google GCS
- Kafka
- KDB.ai
- Lance DB
- MongoDB
- Neo4j
- OneDrive
- OpenSearch
- Oracle
- PG Vector
- Pinecone Destination
- Postgres
- Pure Storage
- Qdrant
- Redis
- S3 Destination
- Singlestore
- Snowflake
- Turbo Puffer
- Weaviate
- Weaviate
- Notion
- Zilliz
- Milvus

Evaluation

Accuracy

Ok, you care about throughput, latency, cost, reliability, **but how do you measure transform quality?**

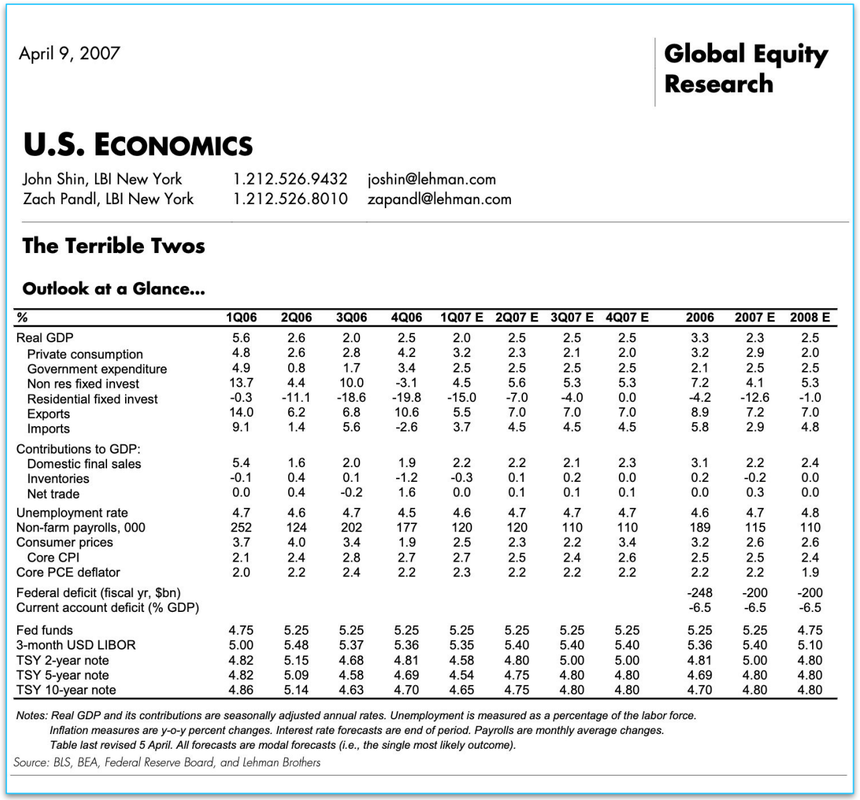
```
========== Evaluation Metrics ==========
                                mean    std
adjusted_cct                   0.876  0.156
cct                            0.874  0.158
cell_level_content_acc         0.687  0.253
cell_level_index_acc           0.709  0.258
detection_f                    0.986  0.066
element_alignment              0.705  0.275
overall                        0.783  0.179
percent_tokens_added           0.066  0.102
percent_tokens_found           0.932  0.104
shifted_cell_content_acc       0.703  0.255
table_teds                     0.715  0.262
table_teds_corrected           0.678  0.283
========================================
```



UNSTRUCTURED

Evaluation for Document Parsing

How To Evaluate Diverse Outputs?



April 9, 2007

Global Equity Research

U.S. ECONOMICS

John Shin, LBI New York 1.212.526.9432 joshin@lehman.com
Zach Pandl, LBI New York 1.212.526.8010 zapandl@lehman.com

The Terrible Twos

Outlook at a Glance...

%	1Q06	2Q06	3Q06	4Q06	1Q07 E	2Q07 E	3Q07 E	4Q07 E	2006	2007 E	2008 E
Real GDP	5.6	2.6	2.0	2.5	2.0	2.5	2.5	2.5	3.3	2.3	2.5
Private consumption	4.8	2.6	2.8	4.2	3.2	2.3	2.1	2.0	3.2	2.9	2.0
Government expenditure	4.9	0.8	1.7	3.4	2.5	2.5	2.5	2.5	2.1	2.5	2.5
Non res fixed invest	13.7	4.4	10.0	-3.1	4.5	5.6	5.3	5.3	7.2	4.1	5.3
Residential fixed invest	-0.3	-11.1	-18.6	-19.8	-15.0	-7.0	-4.0	0.0	-4.2	-12.6	-1.0
Exports	14.0	6.2	6.8	10.6	5.5	7.0	7.0	7.0	8.9	7.2	7.0
Imports	9.1	1.4	5.6	-2.6	3.7	4.5	4.5	4.5	5.8	2.9	4.8
Contributions to GDP:											
Domestic final sales	5.4	1.6	2.0	1.9	2.2	2.2	2.1	2.3	3.1	2.2	2.4
Inventories	-0.1	0.4	0.1	-1.2	-0.3	0.1	0.2	0.0	0.2	-0.2	0.0
Net trade	0.0	0.4	-0.2	1.6	0.0	0.1	0.1	0.1	0.0	0.3	0.0
Unemployment rate	4.7	4.6	4.7	4.5	4.6	4.7	4.7	4.7	4.6	4.7	4.8
Non-farm payrolls, 000	252	124	202	177	120	120	110	110	189	115	110
Consumer prices	3.7	4.0	3.4	1.9	2.5	2.3	2.2	3.4	3.2	2.6	2.6
Core CPI	2.1	2.4	2.8	2.7	2.7	2.5	2.4	2.6	2.5	2.5	2.4
Core PCE deflator	2.0	2.2	2.4	2.2	2.3	2.2	2.2	2.2	2.2	2.2	1.9
Federal deficit (fiscal yr, $bn)									-248	-200	-200
Current account deficit (% GDP)									-6.5	-6.5	-6.5
Fed funds	4.75	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	4.75
3-month USD LIBOR	5.00	5.48	5.37	5.36	5.35	5.40	5.40	5.40	5.36	5.40	5.10
TSY 2-year note	4.82	5.15	4.68	4.81	4.58	4.80	5.00	5.00	4.81	5.00	4.80
TSY 5-year note	4.82	5.09	4.58	4.69	4.54	4.75	4.80	4.80	4.69	4.80	4.80
TSY 10-year note	4.86	5.14	4.63	4.70	4.65	4.75	4.80	4.80	4.70	4.80	4.80

Notes: Real GDP and its contributions are seasonally adjusted annual rates. Unemployment is measured as a percentage of the labor force. Inflation measures are y-o-y percent changes. Interest rate forecasts are end of period. Payrolls are monthly average changes. Table last revised 5 April. All forecasts are modal forecasts (i.e., the single most likely outcome).

Source: BLS, BEA, Federal Reserve Board, and Lehman Brothers

U.S. ECONOMICS
April 9, 2007
Global Equity Research
John Shin, LBI New York,
1-212-526-9432, joshin@lehman.com
Zach Pandl, LBI New York,
1-212-526-8010, zapandl@lehman.com
The Terrible Twos
Outlook at a Glance...
1Q06 2Q06 3Q06 4Q06 1Q07 E 2Q07 E 3Q07 E 4Q07 E 2006 2007 E 2008 E
Real GDP 5.6 2.6 2.0 2.5 2.0 2.5 2.5 2.5 3.3 2.3 3.5
Personal consumption 4.8 2.6 2.8 4.2 3.5 2.5 2.5 2.5 3.2 2.8 2.5
Nonres fixed expenditure 7.5 8.9 11.7 5.8 2.5 2.0 1.5 0.5 7.2 1.6 4.0
Non res fixed invest. 18.1 1.1 10.3 -1.8 -5.0 -7.0 -8.0 -6.0 4.2 -6.6 -1.5
Res fixed invest. -16.2 -11.2 -18.6 -19.8 -15.5 -7.6 -4.0 0.0 -5.9 -7.3 3.0
Net exports 0.1 1.6 0.8 1.6 1.0 1.0 1.0 1.0 0.8 1.0 0.0
Government 3.1 1.4 2.2 1.9 2.0 2.0 2.0 2.0 2.1 2.0 1.5
Contributions to GDP:
Domestic final sales 5.4 1.6 2.3 1.9 2.2 2.2 2.1 2.1 3.1 2.2 3.4
Inventories -0.1 0.4 -0.1 1.2 0.3 0.1 0.2 0.2 0.0 0.2 0.0
Foreign trade 0.2 0.6 -0.2 -0.5 -0.5 0.2 0.2 0.1 0.2 0.0 0.0
Price indices - % chg: -0.0 0.4 0.2 1.4 1.0 0.1 0.2 0.2 0.6 0.3 0.0
Non-farm payrolls, '000s 282 194 200 177 190 100 110 110 189 115 110


```
[ {   "type": "Header",
      "text": "April 9, 2007",
      "metadata": {
         "coordinates": {
            "points": [
               [ 189.96, 119.20],
               [ 189.96, 155.14],
               [ 372.03, 155.14],
               [ 372.03, 119.20]
            ],
            "system": "PixelSpace",
            "layout_width": 1700,
            "layout_height": 2200
         },
         "filetype": "application/pdf",
         "languages": ["eng"],
         "page_number": 1,
      },
}, {  "type": "Header",
      "text": "Global Equity\nResearch",
      "metadata": {
         "coordinates": {
            "points": [
               [ 1265.90, 112.60 ],
               [ 1265.90, 210.77 ],
               [ 1556.16, 210.77 ],
               [ 1556.16, 112.60 ]
            ],
            "system": "PixelSpace",
            "layout_width": 1700,
            "layout_height": 2200
         },
         "filetype": "application/pdf",
         "languages": ["eng"],
         "page_number": 1,
      },
…
```



```
<body>
   <div class="header">
      <div class="date">April 9,
2007</div>
      <div class="title">U.S.
ECONOMICS</div>
      <div class="subtitle">Global Equity
Research</div>
      <div class="authors">
         <div>John Shin, LBI New York |
1-212-526-9432 |
joshin@lehman.com</div>
         <div>Zach Pandl, LBI New York
| 1-212-526-8010 |
zapandl@lehman.com</div>
      </div>
   </div>

   <div class="report-title">The Terrible
Twos</div>
   <div class="section-title">Outlook at
a Glance...</div>

   <table>
      <thead>
         <tr>
            <th></th>
            <th>1Q06</th>
            <th>2Q06</th>
            <th>3Q06</th>
            <th>4Q06</th>
            <th>1Q07 E</th>
            <th>2Q07 E</th>
            <th>3Q07 E</th>
            <th>4Q07 E</th>
            <th>2006</th>
            <th>2007 E</th>
```

UNST
RUCT
URED

Introducing: SCORE

SCORE: A Semantic Evaluation Framework for Generative Document Parsing

Renyu Li*[1], Antonio Jimeno Yepes[1], Yao You[1], Kamil Pluciński[1], Maximilian Operlejn[1], and Crag Wolfe[1]

[1]Unstructured Technologies
https://unstructured.io

Abstract

Multi-modal generative document parsing systems challenge traditional evaluation: unlike deterministic OCR or layout models, they often produce semantically correct yet structurally divergent outputs. Conventional metrics—CER, WER, IoU, or TEDS—misclassify such diversity as error, penalizing valid interpretations and obscuring system behavior.

We introduce **SCORE** (*Structural and COntent Robust Evaluation*), an interpretation-agnostic framework that integrates (i) adjusted edit distance for robust content fidelity, (ii) token-level diagnostics to distinguish hallucinations from omissions, (iii) table evaluation with spatial tolerance and semantic alignment, and (iv) hierarchy-aware consistency checks. Together, these dimensions enable evaluation that embraces representational diversity while enforcing semantic rigor.

Across **1,114 pages** spanning a holistic benchmark and a field dataset, SCORE consistently revealed cross-dataset performance patterns missed by standard metrics. In **2–5% of pages** with ambiguous table structures, traditional metrics penalized systems by **12–25% on average**, leading to distorted rankings. SCORE corrected these cases, recovering equivalence between alternative but valid interpretations. Moreover, by normalizing generative outputs into a format-agnostic representation, SCORE reproduces traditional scores (e.g., table F1 up to **0.93**) without requiring object-detection pipelines, demonstrating that generative parsing alone suffices for comprehensive evaluation.

Download Here!

https://arxiv.org/pdf/2509.19345

Li, R., Jimeno Yepes, A., You, Y., Pluciński, K., Operlejn, M., & Wolfe, C. (2025).
SCORE: A Semantic Evaluation Framework for Generative Document Parsing. *arXiv preprint arXiv:2509.19345.*

SCORE - How We Stack Up

System	Adjusted CCT	Tokens Added	Element Alignment	Table Cell Level Content Accuracy	Table Cell Level Spatial Accuracy	Table Overall
Unstructured	**0.917**	**0.027**	0.644	**0.795**	**0.786**	**0.844**
Databricks AI Parse	0.854	0.057	0.398	0.754	0.726	0.816
NVIDIA NeMo Retriever	0.689	0.043	0.315	0.614	0.677	0.709
Snowflake AI Parse	0.854	0.094	0.632	0.714	0.706	0.791
Reducto	0.885	0.070	**0.646**	0.789	0.776	0.839
LlamaParse	0.752	0.039	0.284	0.500	0.507	0.622
Docling	0.783	0.060	0.611	0.746	0.742	0.816

Thank you