OCR or Not? Rethinking Document Information Extraction in the MLLMs Era with Real-World Large-Scale Datasets

EACL 2026

Jiyuan Shen[1], Peiyue Yuan[1], Atin Ghosh[1], Yifan Mai[2], Daniel Dahlmeier[1]
[1]SAP [2]Stanford University
{jiyuan.shen, peiyue.yuan, atin.ghosh, d.dahlmeier}@sap.com yifan@cs.stanford.edu

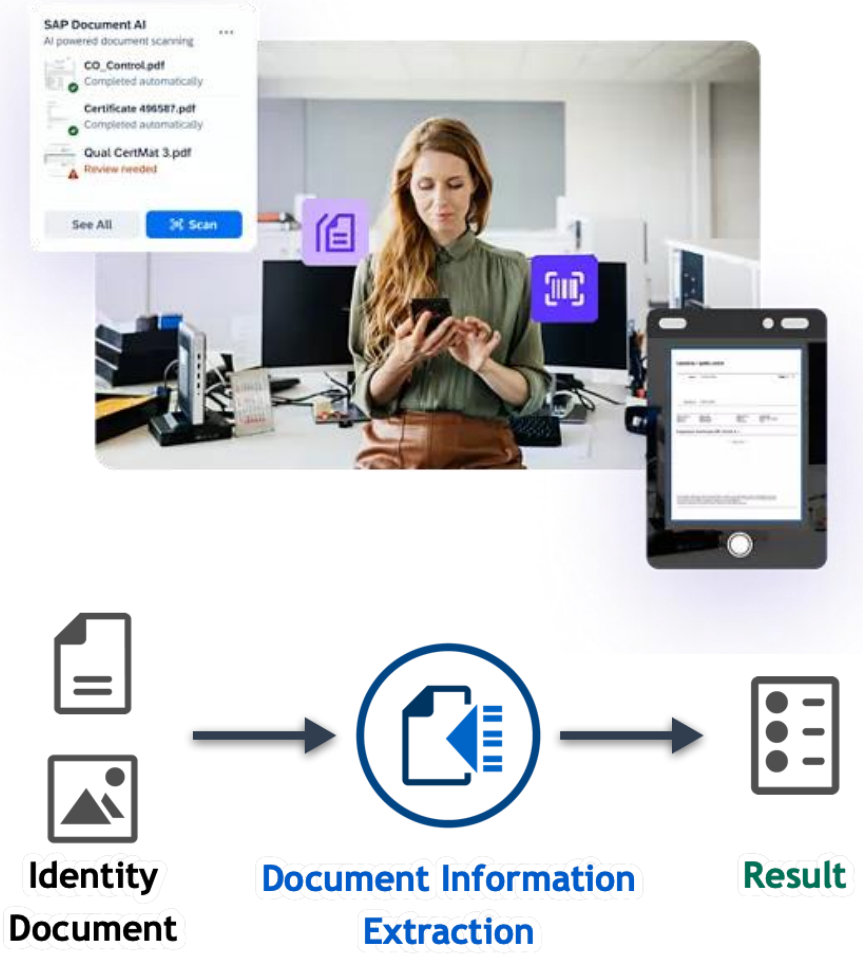
Background

The usage of IDP[1] in industry is still under-explored within the MLLMs era.

Strong market demand: Automating data / information extraction from diverse documents is increasingly critical.

Traditional method works but complex: Industries mostly rely on a two-stage pipeline (OCR + specific extraction model).

Rising Interest in MLLMs: MLLMs enable simpler, unified extraction pipelines.



SAP Document AI Service

[1] IDP stands for intelligent document processing.

Research Direction

Comprehensive benchmarking & systematic analysis

A comprehensive and large-scale benchmarking across flagship models is lacking.

Is OCR necessary for MLLM-based document information extraction?

Can MLLMs serve as a promising path for streamlining the pipeline?

What factors drive MLLM-based method's performance?

Real-World Industrial Document Dataset

Two domains:

- C1: Supply Chain domain documents
- C2: Finance domain documents

Scale: ~1,000 documents with manual ground-truth labels.

Challenges:

Multilingual Diversity:

- 30+ currencies
- Multi-page documents across regions
- Wide language distribution *(see Appendix B for statistics)*

Structural Complexity:

- Nested information
- Stacked / merged cells within line items
- Heterogeneous header layouts
- Cross-section dependencies *(Head / Line items / Tail)*

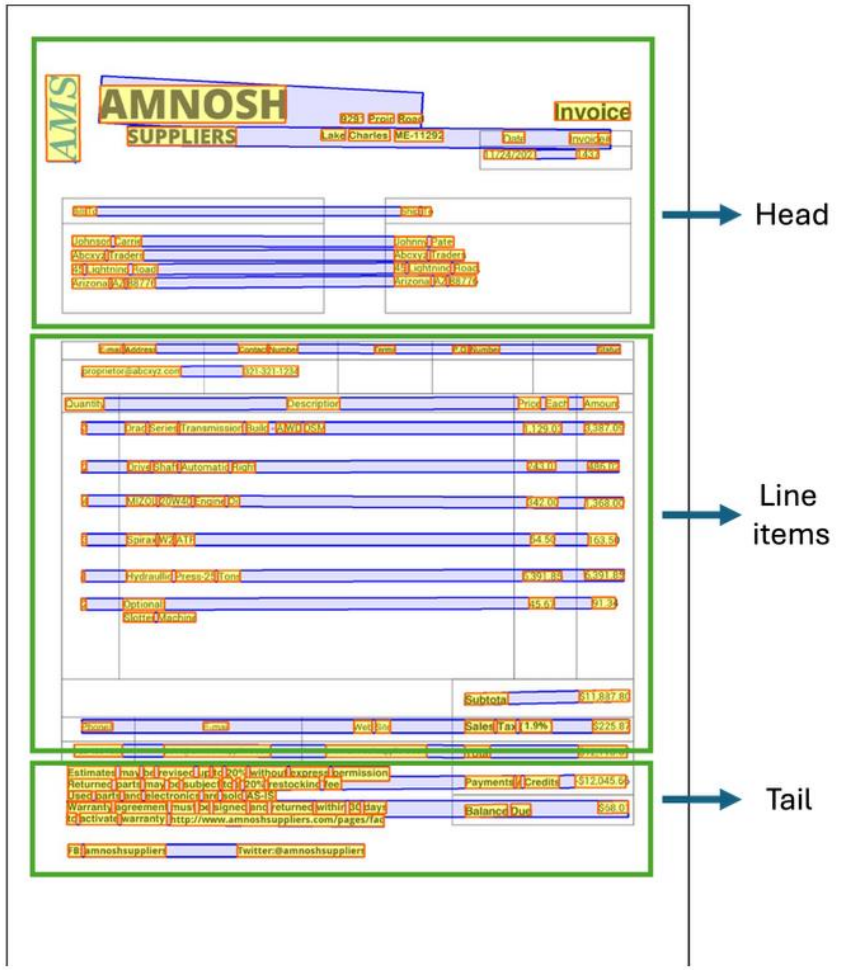


Figure 1: Example of a document page extracted using our OCR engine.

Evaluation Pipeline and Metrics

Text Extraction (OCR): Run OCR to extract text with positional layout information (skipped for image-only settings).

Structured Extraction (Schema-guided): Use an MLLM prompt template with format instructions + document schema to extract header fields + line items; output as JSON key–value pairs.

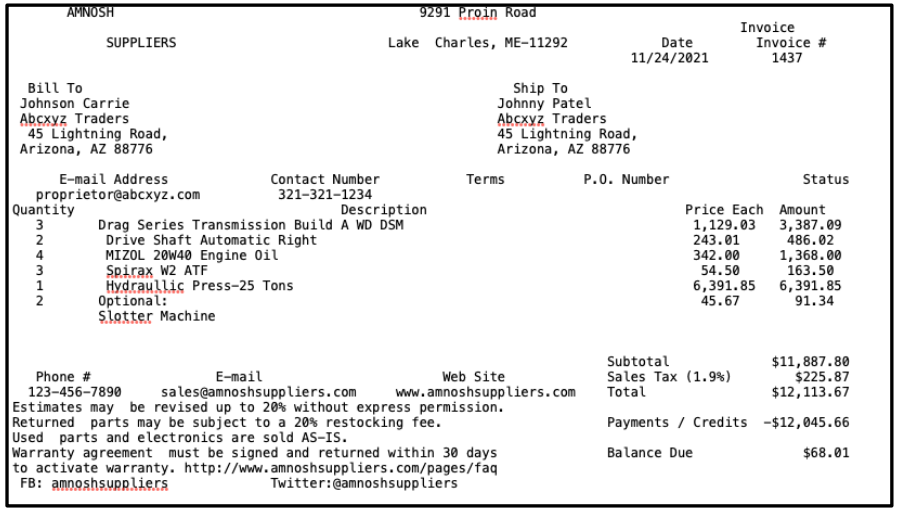
Metric (F1 Score): Compute precision and recall over all predicted key–value pairs, then report the overall F1 score.

OCR

```
      AMNOSH                                    9291 Proin Road
                                                                                  Invoice
           SUPPLIERS                            Lake  Charles, ME-11292          Date        Invoice #
                                                                                  11/24/2021      1437

 Bill To                                                          Ship To
Johnson Carrie                                                   Johnny Patel
Abcxyz Traders                                                   Abcxyz Traders
 45 Lightning Road,                                              45 Lightning Road,
Arizona, AZ 88776                                                Arizona, AZ 88776

     E-mail Address            Contact Number          Terms          P.O. Number                  Status
  proprietor@abcxyz.com         321-321-1234
Quantity                                Description                                    Price Each  Amount
   3       Drag Series Transmission Build A WD DSM                                      1,129.03   3,387.09
   2        Drive Shaft Automatic Right                                                 243.01      486.02
   4        MIZOL 20W40 Engine Oil                                                      342.00     1,368.00
   3        Spirax W2 ATF                                                                54.50      163.50
   1        Hydraullic Press-25 Tons                                                  6,391.85   6,391.85
   2       Optional:                                                                     45.67       91.34
           Slotter Machine



                                                                        Subtotal                  $11,887.80
   Phone #              E-mail                    Web Site             Sales Tax (1.9%)             $225.87
  123-456-7890     sales@amnoshsuppliers.com     www.amnoshsuppliers.com Total                     $12,113.67
Estimates may  be revised up to 20% without express permission.
Returned  parts may be subject to a 20% restocking fee.                 Payments / Credits        -$12,045.66
Used  parts and electronics are sold AS-IS.
Warranty agreement  must be signed and returned within 30 days          Balance Due                   $68.01
to activate warranty. http://www.amnoshsuppliers.com/pages/faq
 FB: amnoshsuppliers            Twitter:@amnoshsuppliers
```

OCR-extracted text results

OCR-extracted content:



AMNOSH 9291 Proin Road

SUPPLIERS Lake Charles, ME-11292

Invoice

Date 11/24/2021 Invoice # 1437

Bill To
Johnson Carrie
Abcxyz Traders
45 Lightning Road,
Arizona, AZ 88776

Ship To
Johnny Patel
Abcxyz Traders
45 Lightning Road,
Arizona, AZ 88776

E-mail Address	Contact Number	Terms	P.O. Number	Status
proprietor@abcxyz.com	321-321-1234			

Quantity	Description	Price Each	Amount
3	Drag Series Transmission Build A WD DSM	1,129.03	3,387.09
2	Drive Shaft Automatic Right	243.01	486.02
4	MIZOL 20W40 Engine Oil	342.00	1,368.00
3	Spirax W2 ATF	54.50	163.50
1	Hydraullic Press-25 Tons	6,391.85	6,391.85
2	Optional: Slotter Machine	45.67	91.34

Subtotal	$11,887.80
Sales Tax (1.9%)	$225.87
Total	$12,113.67
Payments / Credits	-$12,045.66
Balance Due	$68.01

Phone #	E-mail	Web Site
123-456-7890	sales@amnoshsuppliers.com	www.amnoshsuppliers.com

Estimates may be revised up to 20% without express permission.
Returned parts may be subject to a 20% restocking fee.
Used parts and electronics are sold AS-IS.
Warranty agreement must be signed and returned within 30 days to activate warranty. http://www.amnoshsuppliers.com/pages/faq
FB: amnoshsuppliers Twitter:@amnoshsuppliers

MLLM prompt template:

You are a warehouse manager receiving a delivery. As an expert, you go through the attached delivery note and …

The document may be in English, German or any other language. Some of the fields that you need may be indicated by abbreviations in the language of the document. It is important that…

Instructions: {format instructions}.
{document schema}.

Return date fields in YYYY-MM-DD format. For country and currency use ISO format. Do not include the schema in the answer. Return missing values as empty string…

Here is the document: {OCR extracted content}

Information extraction

Response Example:

```
{
"deliveryDate": [""],
"deliveryNoteNumber": ["ID"],
"documentDate": ["YYYY-MM-DD"],
"purchaseOrderNumber": [""],
"supplierId": [""],
"lineItems": [
{
"lineItem.customerMaterialNumber": "",
"lineItem.itemNumber": "1",
"lineItem.purchaseOrderItemNumber": "",
"lineItem.purchaseOrderNumber": "",
"lineItem.quantity": "QUANTITY",
"lineItem.supplierMaterialNumber": "MATERIAL CODE",
"lineItem.unitOfMeasure": ""
},
...
]
}
```

Key-value pair results

Experiment Results

1. OCR-only provides strong and stable baseline

Company	Model	Image-only			OCR-only			Image + OCR		
		C1	**C2**	Mean	**C1**	**C2**	Mean	**C1**	**C2**	Mean
Meta	Llama 4 Scout	67.4	69.3	68.4	68.1	69.7	68.9	67.3	69.8	68.6
	Llama 4 Maverick	62.8	68.2	65.5	63.9	68.1	66	62.9	68.2	65.5
MistralAI	Pixtral Large (2411)	68.7	57.4	63.1	75.3	71.2	73.3	72.7	68	70.4
Amazon	Nova Pro	77.9	65.1	71.5	68.7	65.1	66.9	77.5	66.6	72.1
OpenAI	GPT-4o mini	68.3	64.9	66.6	66.1	70.5	68.3	71.6	70.5	71.1
	GPT-4o	75.5	68.9	70.1	76	69.5	72.8	76.7	69.3	73
Anthropic	Claude 3 Opus	43.8	56.4	50.1	72	68.2	70.1	74	69.1	71.5
	Claude 3.5 Sonnet	65	69.3	67.2	73.7	**72.6**	72.8	73.6	69.6	71.6
Google	Gemini 1.5 Pro	**87.3**	66.4	**76.8**	**78.4**	69.8	**74.1**	**86.2**	65	**75.6**
	Gemini 2.0 Pro	75.2	**73.3**	74.3	77.6	69.5	73.6	77.1	**73.2**	75.2
	Gemini 2.5 Flash	73.9	71.2	72.6	74.6	69.6	72.1	73	71.4	72.2

C1: delivery note, C2: payment advice

- Mean F1 consistently **66%–74%**.
- **Low variance** across models.

Experiment Results

2. Image-only shows larger model disparity

Company	Model	Image-only			OCR-only			Image + OCR		
		C1	**C2**	Mean	**C1**	**C2**	Mean	**C1**	**C2**	Mean
Meta	Llama 4 Scout	67.4	69.3	68.4	68.1	69.7	68.9	67.3	69.8	68.6
	Llama 4 Maverick	62.8	68.2	65.5	63.9	68.1	66	62.9	68.2	65.5
MistralAI	Pixtral Large (2411)	68.7	57.4	63.1	75.3	71.2	73.3	72.7	68	70.4
Amazon	Nova Pro	77.9	65.1	71.5	68.7	65.1	66.9	77.5	66.6	72.1
OpenAI	GPT-4o mini	68.3	64.9	66.6	66.1	70.5	68.3	71.6	70.5	71.1
	GPT-4o	75.5	68.9	70.1	76	69.5	72.8	76.7	69.3	73
Anthropic	Claude 3 Opus	43.8	56.4	50.1	72	68.2	70.1	74	69.1	71.5
	Claude 3.5 Sonnet	65	69.3	67.2	73.7	**72.6**	72.8	73.6	69.6	71.6
Google	Gemini 1.5 Pro	**87.3**	66.4	**76.8**	**78.4**	69.8	**74.1**	**86.2**	65	**75.6**
	Gemini 2.0 Pro	75.2	**73.3**	74.3	77.6	69.5	73.6	77.1	**73.2**	75.2
	Gemini 2.5 Flash	73.9	71.2	72.6	74.6	69.6	72.1	73	71.4	72.2

C1: delivery note, C2: payment advice

- Performance varies significantly across providers.
- The Google Gemini series achieves the strongest overall performance, followed by the Nova models.

Experiment Results

3. Multimodal input improves stability

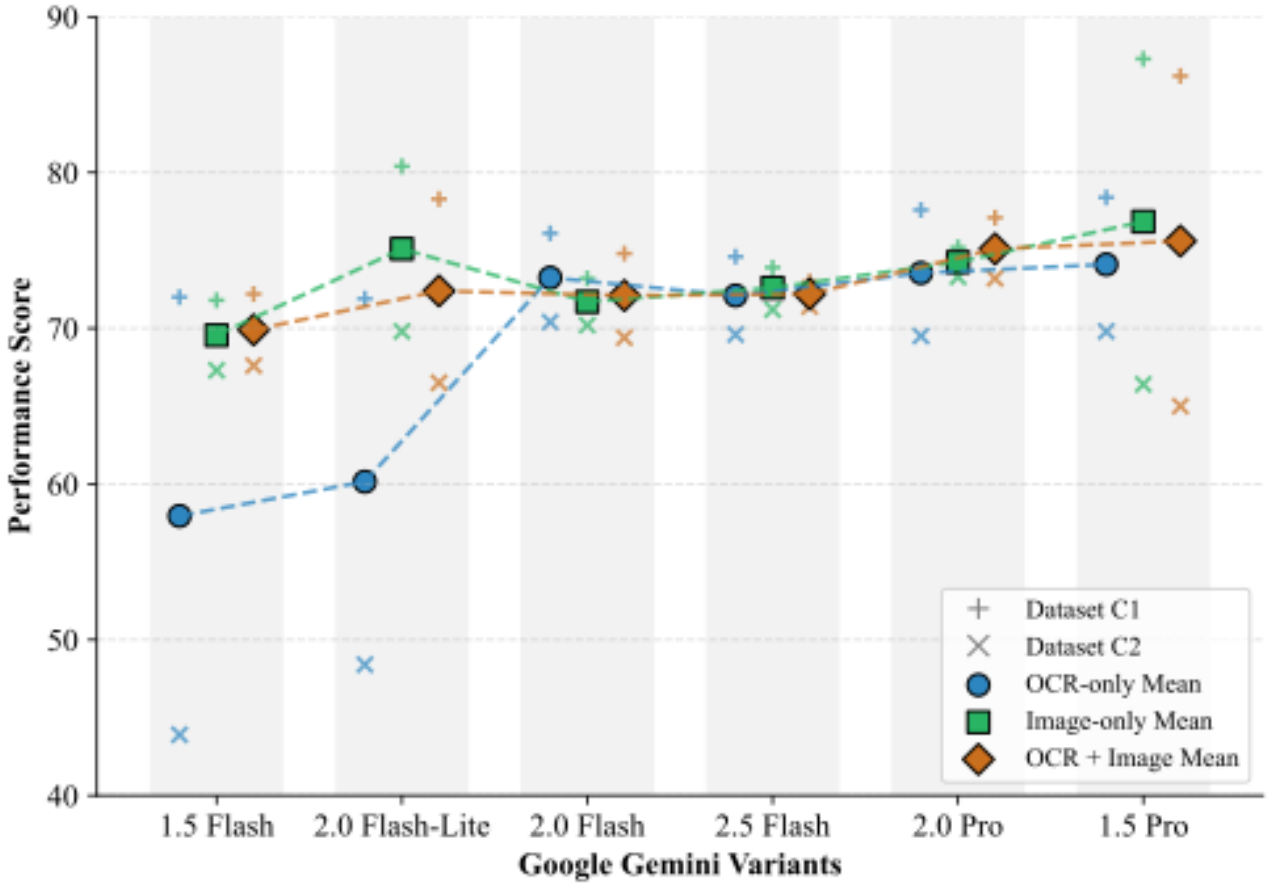
Company	Model	Image-only			OCR-only			Image + OCR		
		C1	C2	Mean	C1	C2	Mean	C1	C2	Mean
Meta	Llama 4 Scout	67.4	69.3	68.4	68.1	69.7	68.9	67.3	69.8	68.6
	Llama 4 Maverick	62.8	68.2	65.5	63.9	68.1	66	62.9	68.2	65.5
MistralAI	Pixtral Large (2411)	68.7	57.4	63.1	75.3	71.2	73.3	72.7	68	70.4
Amazon	Nova Pro	77.9	65.1	71.5	68.7	65.1	66.9	77.5	66.6	72.1
OpenAI	GPT-4o mini	68.3	64.9	66.6	66.1	70.5	68.3	71.6	70.5	71.1
	GPT-4o	75.5	68.9	70.1	76	69.5	72.8	76.7	69.3	73
Anthropic	Claude 3 Opus	43.8	56.4	50.1	72	68.2	70.1	74	69.1	71.5
	Claude 3.5 Sonnet	65	69.3	67.2	73.7	**72.6**	72.8	73.6	69.6	71.6
Google	Gemini 1.5 Pro	**87.3**	66.4	**76.8**	**78.4**	69.8	**74.1**	**86.2**	65	**75.6**
	Gemini 2.0 Pro	75.2	**73.3**	74.3	77.6	69.5	73.6	77.1	**73.2**	75.2
	Gemini 2.5 Flash	73.9	71.2	72.6	74.6	69.6	72.1	73	71.4	72.2

C1: delivery note, C2: payment advice

- Mean F1 concentrates in a narrower **70%–75%** band.
- Multimodal input enhances **robustness and consistency**.

Experiment Results

4. OCR is not necessary for strong models.

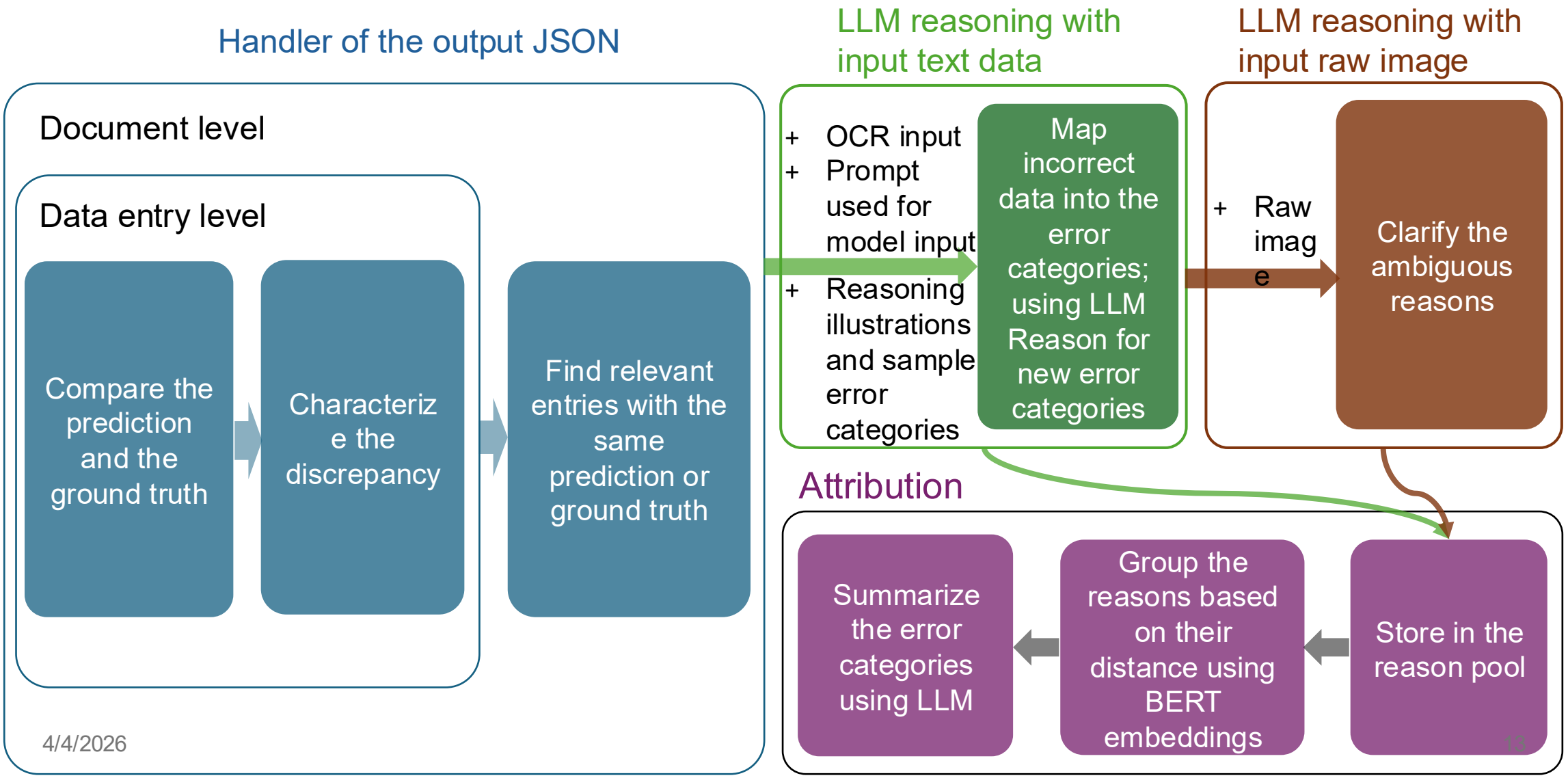


- Advanced multimodal models are capable of **directly extracting structured information from document images** and comprehending textual content effectively, without the need for OCR as an intermediary.
- In particular, for the Gemini models, OCR-generated text appears to **provide little to no** additional benefit.
- The overall performance **improves as the size of the model increases** accordingly[2].

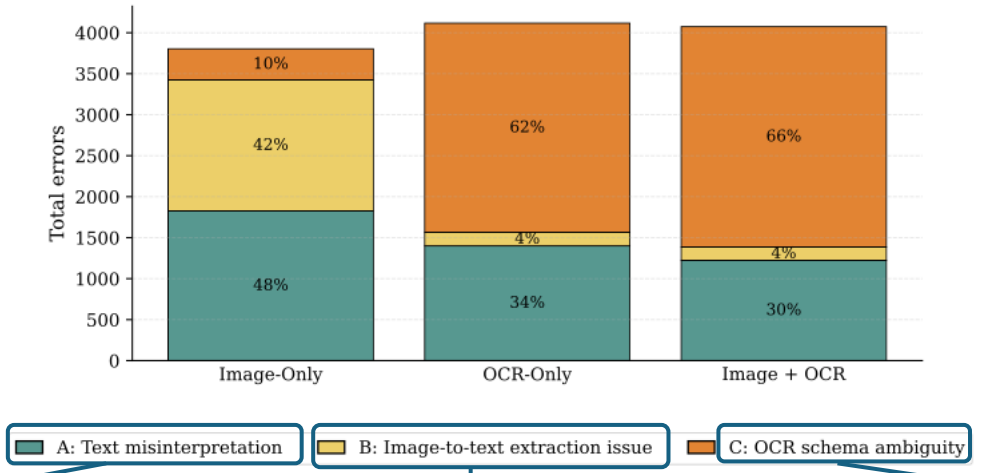
[2] Google does not disclose the exact parameter sizes for each variant, but the size relationship can be partially inferred from the model naming.

Hierarchical Error Analysis Framework

Explore why errors happen and how to further improve.



Error analysis can help optimize prompt and thus help improve the final performance.



Example 1:

Data entry: `"lineItem.itemNumber"`
Ground truth: ["2"]
Prediction: `"002"`
Cause: *"Error due to misreading or misunderstanding the text format"*

➞ Insight:
Absence of standardized itemNumber formatting may lead to value padding errors.

Example 4:

Data entry: `"lineItem.supplierMaterialNumber"`
Ground truth: ["KL-840I"]
Prediction: `"KL-8401"`
Cause: *"The model misinterpreted the quantity field as the item number due to their close proximity within the document."*

➞ Insight:
Visual similarity between characters (e.g., I and 1) can result in misclassification.

Example 9:

Data entry: `"lineItem.quantity"`
Ground truth: ["13"]
Prediction: `"7"`
Cause: *"The OCR data extracted the itemNumber and quantity as adjacent fields, which can lead to misinterpretation by the LLM."*

➞ Insight:
Loss of spatial structure in OCR output can impair accurate field association.

Prompt optimization based on the error analysis

1. Prompt Optimization: Introducing explicit emphasis and reasoning cues to encourage a more thoughtful generation.
2. Format Refinement: Strengthening format constraints to reduce output inconsistencies.
3. Schema Adjustment: Clarifying schema descriptions to minimize ambiguity.

Gemini 1.5 Pro	Initial	Final
Dataset C1	87.3	**89.1**
Dataset C2	66.4	**68.6**
Mean	76.8	**78.9**

Prompt Template:

You are a warehouse manager receiving a delivery. As an expert, you go through the attached delivery note and carefully extract the data that you require to receive the shipped goods and process them in your ERP system. So it is important to focus on the actually received goods and quantities.

The document may be in English, German or any other language. Some of the fields that you need may be indicated by abbreviations in the language of the document. It is important that you carefully extract the information and that you only retrieve information actually on the document. If you have any doubts on a field, skip the field.

Instructions: {format instructions}.
{document schema}.

Return date fields in YYYY-MM-DD format.
For country and currency use ISO format.
Do not include the schema in the answer.
Return missing values as empty string.
Always return valid json and don't wrap you response in backticks!
Do not include a comma before the closing curly bracket.

Here is the document: {OCR extracted content}

Here is the image:

Before

Optimized

Prompt Template for Image-only Input:

You are a warehouse manager receiving a delivery. As an expert, you will go through the attached delivery note and carefully extract the data required to receive the shipped goods and process them in your ERP system. Focus on the actually received goods and quantities.

The document may be in English, German, or any other language. Some fields may be indicated by abbreviations. Extract only the information present in the document. If you have doubts about a field, skip it.

Format instructions: {modified format instructions}.
{modified document schema}.

Return date fields in YYYY-MM-DD format. For country and currency, use ISO format. Do not include the schema in the answer. Ensure that all fields are returned as valid values or empty strings (‘""’), rather than null. If a field does not have a value, return it as an empty string.

Always return valid JSON and do not wrap your response in backticks! Ensure that the JSON structure is valid and does not contain any extra commas or brackets. Each object should be properly closed without trailing commas.

Be attentive to abbreviations and language variations in the document, and ensure that you extract the correct information based on context. Validate the JSON structure before returning the output, checking for any syntax errors. Accuracy in the extraction process is crucial, ensuring that all relevant details are captured accurately.

Emphasize the importance of accuracy in the extraction process and encourage the model to double-check its outputs against the provided schema. Pay special attention to context clues in the document to accurately extract and interpret abbreviations and language variations. Your output must reflect the exact information present in the document, as inaccuracies can lead to significant operational issues.

Here is the document image:

After

Contributions

- Conduct a comprehensive benchmark over 11 state-of-the-art MLLMs.
- Image-only input emerges as a promising direction for document information extraction.
- Scaling law exists.
- General-purpose MLLMs lack task-specific knowledge for structured extraction.
- A carefully designed schemas, exemplars, and instructions improve performance. This prompt optimization can be systematically discovered by leveraging automatic error analysis.