



Invoice OCR



Keep track of the changes and updates for the Invoice OCR API.

Version 4

⚡ Features and Changes (July 17th, 2024)

We added the LiLT for **line_items** reconstruction on the Invoice API.

- 🔥 **Strong improvement on `line_items`**
Reduction in errors in terms of perfectly reconstructed lines of about **30 to 40%**.
- ⚡ **Strong reduction of processing time for invoices with many line items**
- ✨ **New field `unit_of_measure` in `line_items` that represents the unit of measurement for the item, such as kilograms, liters, units, etc.**

⚡ Features and Changes (July 11th, 2024)

We added the LiLT for **total_amount** , **total_tax** and **taxes** on the Invoice API.

- 🔥 Strong improvement on **total_amount** , **total_tax** and **taxes** .

We have observed a reduction in error rates as follows:

- 36% for **total_amount**
- 25% for **total_tax** (+7% on precision)
- 15% for **taxes**

⚡ Features and Changes (May 16th, 2024)

- 🐛 Fix date parsing for spanish/italian invoices
- 🐛 Fix reading errors for invoice number
- 🔥 Strong improvement on **document_type**
- ⚡ Taxes from line items

taxes field outputs taxes from the line items when no tax summary is present on the document.

⚡ Features and Changes (April 25th, 2024)

- 🚀 Extended latin alphabet support
We released new models for our generic text detection and recognition pipeline. This release has increased the overall performances on all fields and supports extended latin alphabet characters:

```
{' ', 'i', '¥', '¿', 'Á', 'Ã', 'Ä', 'Å', 'Æ', ' '}
```

-  **Strong improvement on `due_date` , `line_items`**

We have observed a reduction in error rates as follows:

- 20% for `due_date`
- 25% for `line_items`

-  **New fields:**

The API is now extracting the following fields:

`customer_id` : The identifier of the customer in the supplier's referential. It can also refer to the client ID, client / customer account number...

`supplier_phone_number` : The phone number of the supplier

`supplier_email` : The supplier email address

`supplier_website` : The supplier website URL

-  **General accuracy improvement**

Thanks to the improvement done on our generic text detection and recognition algorithms, we measured a reduction in error rates on all fields, especially for supplier and customer information.

Features and Changes (March 11th, 2024)

-  **Integration of company ID & logo database**

We have integrated a company ID database and a vector database featuring millions of logos. This enhancement enables our R&D team to efficiently rectify any issues with non-functional supplier names.

- 🔥 **Strong improvement on `supplier_name` , `customer_name` , and `invoice_number`**

We have observed a reduction in error rates as follows:

- 20% for `customer_name`
- 15% for `supplier_name`
- 10% for `invoice_number`

The improvement in `supplier_name` was achieved by incorporating information from the databases. The `customer_name` algorithm now mirrors the `supplier_name` one. `invoice_number` now employs an NLP modality to boost its precision.

- ✨ **New fields: `shipping_address` , `billing_address` and `tax_base` :**

The API now extracts the `shipping_address` and `billing_address` .

We have updated the `taxes` field to include the `tax_base` value. Including the `tax_base` value in the `taxes` field can help you calculate and report taxes more accurately.

⚡ **Features and Changes (January 30th, 2024)**

- 🚀 **Integration of a proprietary language model in the algorithm pipeline: LiLT**

LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding.

LiLT's design combines textual content with layout structure. This means it doesn't just read the text but also understands how the text is organized within the

document. For instance, it recognizes headings, paragraphs, tables, and other structural elements, which is a crucial aspect of context awareness in document processing.

The integration of this new language model in our pipeline helps us achieve better accuracy, and more flexibility when adding new supported fields.

- 🔥 **Strong improvement on `supplier_name`, `supplier_address`, and `supplier_company_registrations`**

The main focus of this release was to improve drastically the supplier information extraction.

We measured a decrease in error rates of:

- 42% for `supplier_name`
- 10% for `supplier_address`
- 10% for `supplier_company_registrations`

Moreover, the integration of the LILT offers more robustness in terms of languages thanks to its language-independent component and will help us improve all other fields in the next releases.

- ✨ **New field: `total_tax`**

The API is now extracting the total tax information, returned as a number. It corresponds to the total tax explicitly written in the document.

- 🔥 **General improvement for all fields**

More training data was added to our training set, including different geographies and more variability. We've measured an improvement in accuracy for all extracted fields.

⚡ **Features and Changes (September 1st, 2023)**

- ✨ **New field: Raw Value available for both Supplier Name and Customer Name.**

The Raw Value extracts the name without post processing nor formatting. It can thus be different from the Value.

Customer Name

Value: **JESSIE M HORNE**

Raw Value: **Jessie M Horne**

Confidence score : 0.77

⚡ Features and Changes (January 3rd, 2023)

- New field extracted: Reference numbers which is a list of variable size which may contain PO numbers, reference numbers or project numbers

⚡ Features and Changes (November 30th, 2022)

- Line items extraction. Line items are returned as a list in the json response. Each item includes:
 - description
 - quantity
 - unit price
 - total amount
 - tax rate
 - tax amount
 - product code
- Renaming of fields in the json response for more clarity:
 - `supplier` -> `supplier_name`

- `payment_details` -> `supplier_payment_details`
- `company_registration` -> `supplier_company_registrations`
- `customer_company_registration` -> `customer_company_registrations`
- `total_incl` -> `total_amount`
- `total_excl` -> `total_net`

Version 3

Features and Changes (March 24th, 2022)

- Update in response scheme with new orientation information available.
- Update in polygon coordinates and format.
- Improvement in extraction performance of company IDs.

Features and Changes (Feb 24th, 2022)

- Supports 4 new extracted fields:
 - Customer name
 - Customer address
 - Customer company registration
 - Supplier address

Version 2



Invoice V2 API Performance Update

For increased performance in the extraction of your fields and result, please upgrade to v3 as this version is currently not maintained.

Features and Changes (August 18, 2021)

- Supports the use of native PDF text content to extract fields.
- Improvement in accuracy for document-level prediction.
- Supports 17 additional currencies: AED, AUD, BRL, CNY, COP, CZK, DKK, GNF, HKD, HUF, JPY, NOK, NZD, PLN, SEK, SGD, XPF.
- Extraction performance improvement of these currencies: CAD, CHF, EUR, GBP, USD.
- Addition of international payment details (IBAN, routing number, SWIFT/BIC code, account number).
- Inconsistency in amount / tax fixed.
- Overall improvement in extraction performance.

Version 1

Invoice V1 API Depreciation

Support for the invoice V1 API is deprecated. Please use v3 instead.

Feature: First Release (August 6, 2020)

Extracted fields:

- Due date
- Invoice date
- Invoice number
- Locale & currency
- Supplier identification number (SIRET, EIN, VAT number...)
- Supplier name
- Taxes details
- Total amount including taxes
- Total amount excluding taxes

Questions?



[Join our Slack](#)



Updated 8 days ago

← Bank Account
Details OCR

Passport OCR →

Did this page help you? Yes No