

5.4 Extract data from forms with Azure Document intelligence

Document intelligence **uses machine learning technology to identify and extract key-value pairs and table data from form documents** with accuracy, at scale. This module teaches you how to use the **Azure Document intelligence cognitive service**.

Learning objectives

In this module, you learn how to:

- Identify how **Document intelligence's layout service, prebuilt models, and custom models** can automate processes.
- Use Document intelligence's capabilities with **SDKs, REST API, and Document Intelligence Studio**.
- Develop and test custom models.

Introduction

Forms are used to communicate information in every industry, every day. Many people still manually **extract data from forms** to exchange information.

Consider some of the instances when a person needs to process form data:

- When filing claims
- When enrolling new patients in an online management system
- When entering data from receipts to an expense report
- When reviewing an operations report for anomalies
- When selecting data from a report to give to a stakeholder

Without AI services, people need to manually sort through form documents to identify important information and then manually reenter data to record it. Some may also need to complete these tasks in real-time with a customer.

Azure Document Intelligence services provide the building blocks for automation by using intelligent services to extract data at scale and with accuracy. **Azure Document Intelligence is a Vision API that extracts key-value pairs and table data from form documents.**

Uses of the Azure Document Intelligence service include:

- Process automation
- Knowledge mining

- Industry-specific applications

In this module, you'll learn how to:

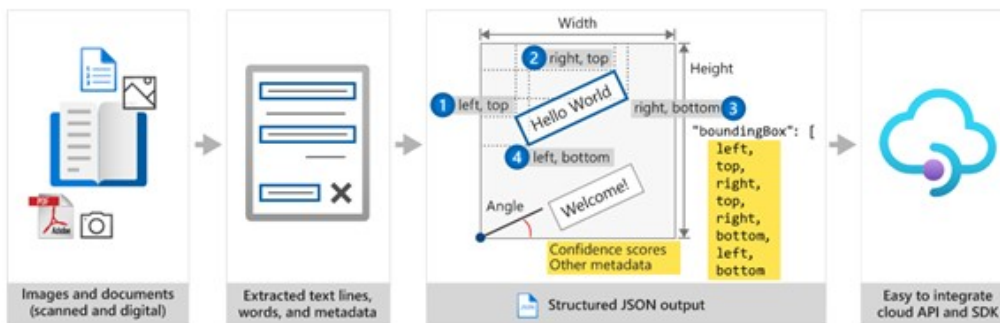
- Identify how Azure Document Intelligence's document analysis, prebuilt, and custom models can automate processes
- Use **Azure Document Intelligence's Optical Character Recognition (OCR)** capabilities with **SDKs and REST API**.
- Develop and test a custom Azure Document Intelligence model

To complete this module, you'll need a Microsoft Azure subscription. If you don't already have one, you can sign up for a free trial at <https://azure.microsoft.com>.

What is Azure Document Intelligence?

Azure Document Intelligence is one of many Azure AI Services, cloud-based artificial intelligence (AI) services with *REST APIs and client library SDKs* that can be used to build intelligence into your applications.

Azure Document Intelligence uses **Optical Character Recognition (OCR) capabilities and deep learning models** to **extract text, key-value pairs, selection marks, and tables from documents**.



OCR captures document structure by **creating bounding boxes around detected objects in an image**. The locations of the bounding boxes are recorded as coordinates in relation to the rest of the page. Azure Document Intelligence services return bounding box data and other information in a structured form with the relationships from the original file.

Bounding Boxes On Invoice



JSON Response



To build a high-accuracy model from scratch, people need to build deep learning models, use a large amount of compute resources, and face long model training times. These factors could make a project infeasible. Azure Document Intelligence provides underlying models that have been trained on thousands of form examples. The underlying models enable you to do high-accuracy data extraction from your forms with little to no model training.

Azure Document Intelligence service components

Azure Document Intelligence is composed of the following services:

- **Document analysis models:** which take an input of JPEG, PNG, PDF, and TIFF files and return a JSON file with the location of text in bounding boxes, text content, tables, selection marks (also known as checkboxes or radio buttons), and document structure.
- **Prebuilt models:** which detect and extract information from document images and return the extracted data in a structured JSON output. Azure Document Intelligence currently supports prebuilt models for several forms, including:
 - W-2 forms
 - Invoices
 - Receipts
 - ID documents
 - Business cards
- **Custom models:** custom models extract data from forms specific to your business. Custom models can be trained through the Azure Document Intelligence Studio.

Note: Some Azure Document Intelligence features are in preview, as of the time this content was authored, and as a result, features and usage details might change. You should refer to the [official page](#) for up-to-date information.

Access services

You can access Azure Document Intelligence services in several ways. These options include using:

- A **REST API**
- Client library **SDKs**
- **Azure Document Intelligence Studio**
- **Azure AI Foundry**

Tip: This module's exercise focuses on the Python and .NET SDKs. The underlying REST services can be used by any language.

Check out the [documentation](#) for quick start guides on all the available SDKs and the REST API.

Get started with Azure Document Intelligence

To start a project with Azure Document Intelligence services, you need an Azure resource and selection of form files for data extraction.

Subscribe to a resource

You can access Azure Document Intelligence services via:

- An **Azure AI Service resource**: a *multi-service subscription key* (used across multiple Azure AI Services)

OR

- An **Azure Document Intelligence resource**: a *single-service subscription key* (used only with a specific Azure AI Service)

Note: Create an **Azure AI Services resource** if you plan to access multiple Azure AI services **under a single endpoint/key**. For Azure Document Intelligence access only, create an Azure Document Intelligence resource.

Note that **you need a single-service resource if you intend to use Microsoft Entra authentication**.

You can subscribe to a service in the Azure portal or with the Azure Command Line Interface (CLI). You can learn more about the CLI commands [here](#).

Understand Azure Document Intelligence file input requirements

Azure Document Intelligence works on input documents that meet these requirements:

- Format must be **JPG, PNG, BMP, PDF (text or scanned), or TIFF**.
- The file size must be **less than 500 MB for paid (S0) tier and 4 MB for free (F0) tier**.
- Image dimensions must be **between 50 x 50 pixels and 10,000 x 10,000 pixels**.
- The total size of the training data set must be **500 pages or less**.

More input requirements can be found in the [documentation](#) for specific models.

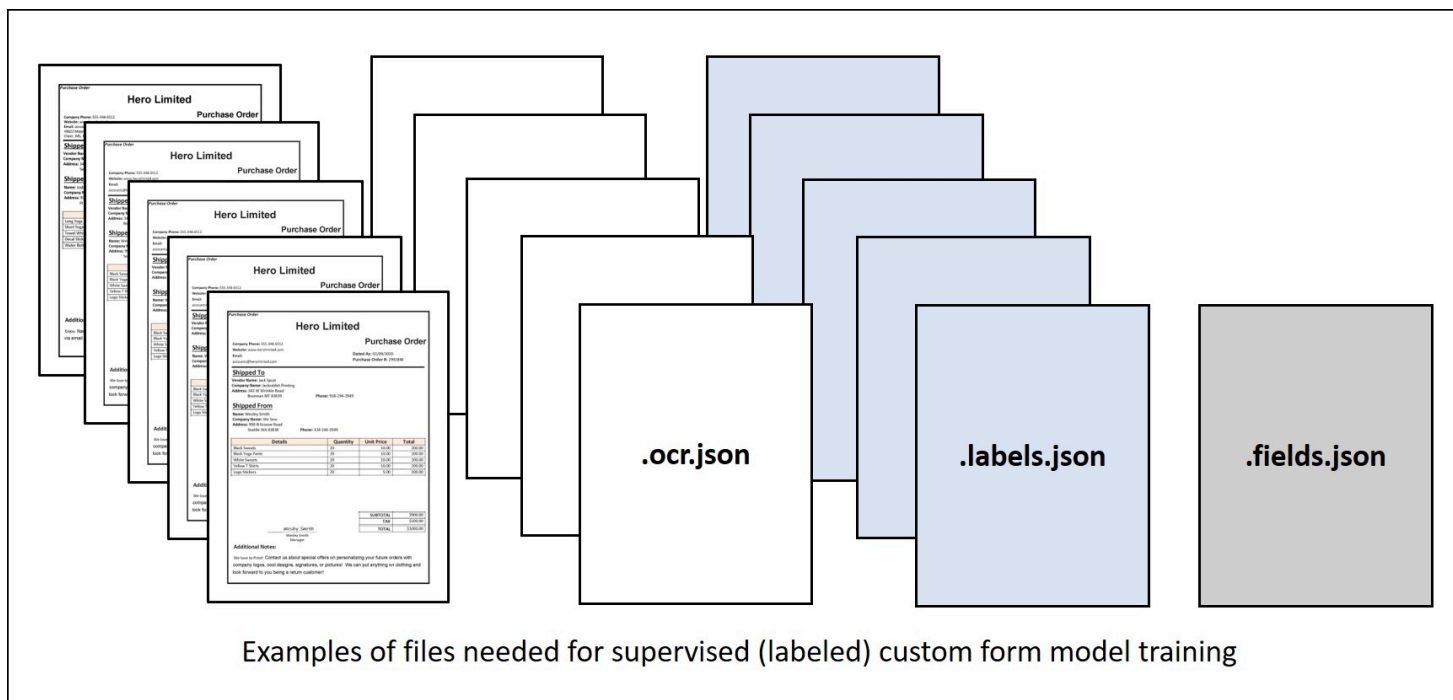
Decide what component of Azure Document Intelligence to use

After you collect your files, decide what you need to accomplish.

Use case	Recommended features to use
Use OCR capabilities to capture document analysis	Use the Layout model, Read model, or General Document model .
Create an application that extracts data from W-2s, Invoices, Receipts, ID documents, Health insurance, vaccination, and business cards	Use a prebuilt model . These models don't need to be trained. Azure Document Intelligence services analyze the documents and return a JSON output .
Create an application to extract data from your industry-specific forms	Create a custom model . This model needs to be trained on sample documents. After you train the custom model, it can analyze new documents and return a JSON output .

Train custom models

Azure's **Azure Document Intelligence service supports supervised machine learning**. You can train custom models and create composite models with form documents and JSON documents that contain labeled fields.



To train a custom model:

- Store sample forms in an **Azure blob container**, along with **JSON files containing layout and label field information**.
 - You can generate an **.ocr.json** file for each sample form using the Azure Document Intelligence's Analyze document function. Additionally, you need a single **fields.json** file describing **the fields you want to extract**, and a **labels.json** file for **each sample form mapping the fields to their location in that form**.
- Generate a **shared access security (SAS) URL** for the container.
- Use the **Build model REST API function (or equivalent SDK method)**.
- Use the **Get model REST API function (or equivalent SDK method)** to get **the trained model ID**.

OR

- Use the Azure Document Intelligence Studio to label and train. There are two types of underlying models for custom forms custom template models or custom neural models.
 - **Custom template models** accurately extract **labeled key-value pairs, selection marks, tables, regions, and signatures from documents**. Training only takes a few minutes, and more than 100 languages are supported.
 - **Custom neural models** are deep learned models that combine layout and language features to accurately **extract labeled fields from documents**. This model is **best for semi-structured or unstructured documents**.

Use Azure Document Intelligence models

Using the API

To extract form data using a custom model, use the analyze document function of either a **supported SDK, or the REST API**, while **supplying model ID** (generated during model training). This function starts the form analysis, which you can then request the result to get the analysis.

Example code to call your model:

```
endpoint = "YOUR_DOC_INTELLIGENCE_ENDPOINT"
key = "YOUR_DOC_INTELLIGENCE_KEY"

model_id = "YOUR_CUSTOM_BUILT_MODEL_ID"
formUrl = "YOUR_DOCUMENT"

document_analysis_client = DocumentAnalysisClient(    endpoint=endpoint, credential=AzureKeyCredential(key) )

# Make sure your document's type is included in the list of document types the custom model can analyze
task = document_analysis_client.begin_analyze_document_from_url(model_id, formUrl)
result = task.result()
```

A successful JSON response contains **analyzeResult** that contains **the content extracted** and **an array of pages containing information about the document content**.

Example analyze document JSON response:

Explore the documentation for [supported language quickstarts](#).

Understanding confidence scores

If the confidence values of the **analyzeResult** are low, try to improve the quality of your input documents.

You want to make sure that the form you're analyzing has a similar appearance to forms in the training set if the confidence values are low. If the form appearance varies, **consider training more than one model, with each model focused on one form format**.

Depending on the use case, you might find that **a confidence score of 80% or higher is acceptable for a low-risk application**. **For more sensitive cases, like reading medical records or billing statements, a score of 100% is recommended**.

Use the Azure Document Intelligence Studio

In addition to SDKs and the REST API, **Azure Document Intelligence** services can be accessed through a user interface called the **Azure Document Intelligence Studio**, an online tool for visually exploring, understanding, and

integrating features from the Azure Document Intelligence service. The Studio can be used to **analyze form layouts, extract data from prebuilt models, and train custom models.**

Azure AI | Document Intelligence Studio

ⓘ Azure Form Recognizer is now Azure AI Document Intelligence. [Learn more](#) about the latest updates to the service and the Studio experience.


[Form Recognizer Studio](#)

Get started with Document Intelligence Studio

Extract text, key-value pairs, tables, and structures from forms and documents using common layouts and prebuilt models, or create your own custom models. [Learn more](#)

Document analysis


Extract text, tables, structure, key-value pairs, and named entities from documents.



Read

Extract printed and handwritten text along with barcodes, formulas and font styles from images and documents.


[Try it out](#)



Layout

Extract tables, check boxes, and text from forms and documents.

[Try it out](#)




General documents

Extract key value pairs and structure like tables and selection marks from any form or document.

[Try it out](#)


Prebuilt models

Extract data from unique document types using the following prebuilt models.




Invoices

Extract invoice ID, customer details, vendor




Receipts

Extract time and date of the transaction,



Business cards

Extract person name, job title, address,



Identity documents

Extract name, expiration date, machine

The Azure Document Intelligence Studio currently supports the following projects:

- **Document analysis models**
 - **Read:** Extract **printed and handwritten text lines, words, locations, and detected languages from documents and images.**
 - **Layout:** Extract **text, tables, selection marks, and structure information from documents (PDF and TIFF) and images (JPG, PNG, and BMP).**
 - **General Documents:** Extract **key-value pairs, selection marks, and entities from documents.**
- **Prebuilt models**
- **Custom models**

Build Document analysis model projects

To **extract text, tables, structure, key-value pairs, and named entities** with **document analysis models**:

- Create an Azure Document Intelligence or Azure AI Services resource.

- Select either "**Read**", "**Layout**", or "**General Documents**" under the Document analysis models category.
- Analyze your document. You'll need your Azure Document Intelligence or Azure AI service **endpoint and key**.

Build prebuilt model projects

To extract data from common forms with prebuilt models:

- Create an **Azure Document Intelligence** or **Azure AI Services** resource
- Select one of the "**prebuilt models**" including **W-2s, Invoices, Receipts, ID documents, Health insurance, vaccination, and business cards**.
- Analyze your document. You'll need your Azure Document Intelligence or Azure AI service **endpoint and key**.

Build custom model projects

You can use **Azure Document Intelligence Studio's custom service** for the entire process of training and testing custom models.

When you use Azure Document Intelligence Studio to build custom models, the **ocr.json** files, **labels.json** files, and **fields.json** file needed for training are automatically created and stored in your storage account.

To train a custom model and use it to extract data with custom models:

- Create an Azure Document Intelligence or Azure AI Services resource
- Collect **at least 5-6 sample forms** for training and upload them to your storage account container.
- Configure **cross-domain resource sharing (CORS)**. CORS enables Azure Document Intelligence Studio to store labeled files in your storage container.
- **Create a custom model project in Azure Document Intelligence Studio**. You'll need to provide configurations linking your storage container and Azure Document Intelligence or Azure AI Service resource to the project.
- Use Azure Document Intelligence Studio to **apply labels to text**.
- Train your model. Once the model is trained, you'll receive a **Model ID and Average Accuracy for tags**.
- Test your model by analyzing a new form that wasn't used in training.

Exercise - Extract data from custom forms

In this exercise, you'll use the Azure Document Intelligence service to train and test a custom model using the Python or .NET SDK.

To complete the exercise for this module, launch the VM and follow the instructions.

Extract Data from Forms

Suppose a company currently requires employees to manually purchase order sheets and enter the data into a database. They would like you to utilize AI services to improve the data entry process. You decide to build a machine

learning model that will read the form and produce structured data that can be used to automatically update a database.

Azure AI Document Intelligence is an Azure AI service that enables users to build automated data processing software. This software can **extract text, key/value pairs, and tables from form documents using optical character recognition (OCR)**. Azure AI Document Intelligence has **pre-built models for recognizing invoices, receipts, and business cards**. The service also provides the capability to train **custom models**. In this exercise, we will focus on building custom models.

Module assessment

In this module, you have learned how to use the Azure Document Intelligence service to extract data from forms.

Consider the following review questions to check your understanding of the topics discussed in this module.

1. A person plans to use an Azure Document Intelligence prebuilt invoice model. To extract document data using the model and REST API language, what are two calls they need to make to the API? **Analyze Invoice and Get Analyze Invoice Result.**
2. A person needs to build an application that submits expense claims and extracts the merchant, date, and total from scanned receipts. What's the best way to build the application? **Use Azure Document Intelligence's prebuilt receipts model.**
3. A person is building a custom model with Azure Document Intelligence services. What is required to train a model? **Along with the form to analyze, JSON files need to be provided.**

Summary

This module focused on Azure Document Intelligence's **prebuilt service, custom service, and its client library SDKs and REST API**. We also introduced the **Azure Document Intelligence Studio** to label and train your model.

Azure Document Intelligence services can be integrated with other Azure AI Services. For example, you can try using [this tutorial](#) with **Azure Document Intelligence and Cognitive Search**.

Document intelligence is just one part of the overall Vision API in Azure AI Services. Azure Document Intelligence services are ever evolving. You can read about the latest updates [here](#).