Document Intelligence read model

Article • 11/21/2023

ⓘ Important

- Document Intelligence public preview releases provide early access to features that are in active development.
- Features, approaches, and processes may change, prior to General Availability (GA), based on user feedback.
- The public preview version of Document Intelligence client libraries default to REST API version **2023-10-31-preview**.
- Public preview version **2023-10-31-preview** is currently only available in the following Azure regions:
- **East US**
- **West US2**
- **West Europe**

This content applies to: ✓ v4.0 (preview) | Previous versions: ✓ v3.1 (GA) ✓ v3.0 (GA)

ⓘ Note

For extracting text from external images like labels, street signs, and posters, use the **Azure AI Vision v4.0 preview Read** feature optimized for general, non-document images with a performance-enhanced synchronous API that makes it easier to embed OCR in your user experience scenarios.

Document Intelligence Read Optical Character Recognition (OCR) model runs at a higher resolution than Azure AI Vision Read and extracts print and handwritten text from PDF documents and scanned images. It also includes support for extracting text from Microsoft Word, Excel, PowerPoint, and HTML documents. It detects paragraphs, text lines, words, locations, and languages. The Read model is the underlying OCR engine for other

Document Intelligence prebuilt models like Layout, General Document, Invoice, Receipt, Identity (ID) document, Health insurance card, W2 in addition to custom models.

What is OCR for documents?

Optical Character Recognition (OCR) for documents is optimized for large text-heavy documents in multiple file formats and global languages. It includes features like higher-resolution scanning of document images for better handling of smaller and dense text; paragraph detection; and fillable form management. OCR capabilities also include advanced scenarios like single character boxes and accurate extraction of key fields commonly found in invoices, receipts, and other prebuilt scenarios.

Development options

Document Intelligence v4.0 (2023-10-31-preview) supports the following tools, applications, and libraries:

⛶ Expand table

Feature	Resources	Model ID
Read OCR model	• Document Intelligence Studio • REST API • C# SDK • Python SDK • Java SDK • JavaScript SDK	prebuilt-read

Input requirements

- For best results, provide one clear photo or high-quality scan per document.
- Supported file formats:

⛶ Expand table

Model	PDF	Image: JPEG/JPG, PNG, BMP, TIFF, HEIF	Microsoft Office: Word (DOCX), Excel (XLSX), PowerPoint (PPTX), and HTML
Read	✓	✓	✓
Layout	✓	✓	✓ (2023-10-31-preview)
General Document	✓	✓	
Prebuilt	✓	✓	
Custom	✓	✓	

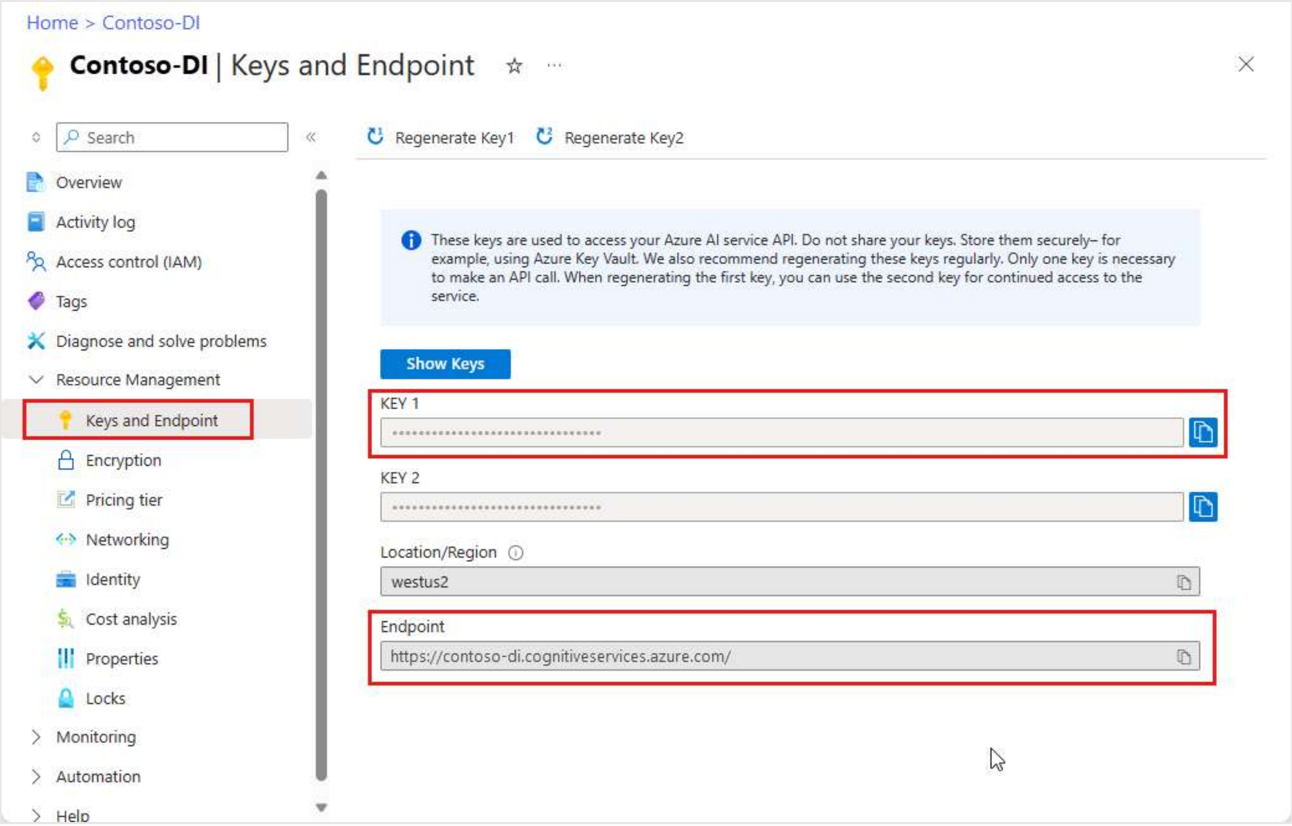
- For PDF and TIFF, up to 2000 pages can be processed (with a free tier subscription, only the first two pages are processed).
- The file size for analyzing documents is 500 MB for paid (S0) tier and 4 MB for free (F0) tier.
- Image dimensions must be between 50 x 50 pixels and 10,000 px x 10,000 pixels.
- If your PDFs are password-locked, you must remove the lock before submission.
- The minimum height of the text to be extracted is 12 pixels for a 1024 x 768 pixel image. This dimension corresponds to about `8` -point text at 150 dots per inch (DPI).
- For custom model training, the maximum number of pages for training data is 500 for the custom template model and 50,000 for the custom neural model.
  - For custom extraction model training, the total size of training data is 50 MB for template model and 1G-MB for the neural model.
  - For custom classification model training, the total size of training data is `1GB` with a maximum of 10,000 pages.

Read model data extraction

Try extracting text from forms and documents using the Document Intelligence Studio. You need the following assets:

- An Azure subscription—you can create one for free

- A Document Intelligence instance in the Azure portal. You can use the free pricing tier (`F0`) to try the service. After your resource deploys, select **Go to resource** to get your key and endpoint.



These keys are used to access your Azure AI service API. Do not share your keys. Store them securely– for example, using Azure Key Vault. We also recommend regenerating these keys regularly. Only one key is necessary to make an API call. When regenerating the first key, you can use the second key for continued access to the service.

KEY 1: ································

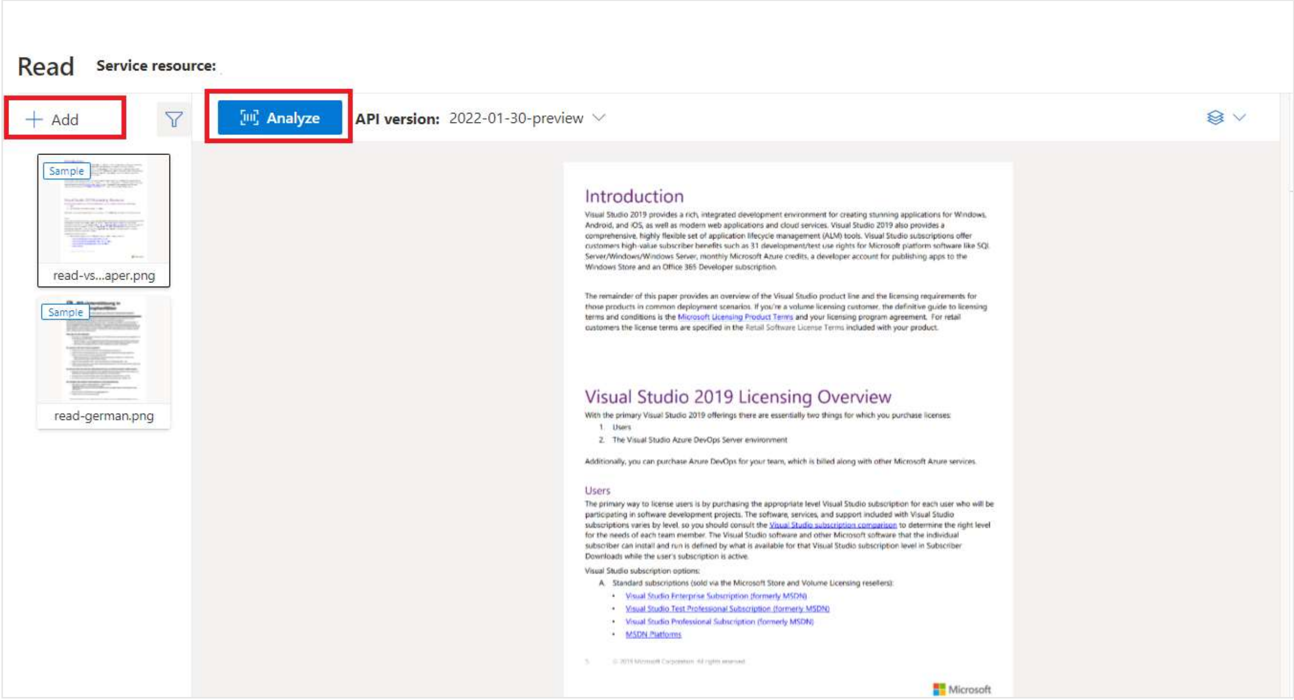
KEY 2: ································

Location/Region: westus2

Endpoint: https://contoso-di.cognitiveservices.azure.com/

> **Note**
>
> Currently, Document Intelligence Studio doesn't support Microsoft Word, Excel, PowerPoint, and HTML file formats.

Sample document processed with Document Intelligence Studio



API version: 2022-01-30-preview

Introduction

Visual Studio 2019 provides a rich, integrated development environment for creating stunning applications for Windows, Android, and iOS, as well as modern web applications and cloud services. Visual Studio 2019 also provides a comprehensive, highly flexible set of application lifecycle management (ALM) tools. Visual Studio subscriptions offer customers high-value subscriber benefits such as 31 development/test use rights for Microsoft platform software like SQL Server/Windows/Windows Server, monthly Microsoft Azure credits, a developer account for publishing apps to the Windows Store and an Office 365 Developer subscription.

The remainder of this paper provides an overview of the Visual Studio product line and the licensing requirements for those products in common deployment scenarios. If you're a volume licensing customer, the definitive guide to licensing terms and conditions is the Microsoft Licensing Product Terms and your licensing program agreement. For retail customers the license terms are specified in the Retail Software License Terms included with your product.

Visual Studio 2019 Licensing Overview

With the primary Visual Studio 2019 offerings there are essentially two things for which you purchase licenses:

1. Users
2. The Visual Studio Azure DevOps Server environment

Additionally, you can purchase Azure DevOps for your team, which is billed along with other Microsoft Azure services.

Users

The primary way to license users is by purchasing the appropriate level Visual Studio subscription for each user who will be participating in software development projects. The software, services, and support included with Visual Studio subscriptions varies by level, so you should consult the Visual Studio subscription comparison to determine the right level for the needs of each team member. The Visual Studio software and other Microsoft software that the individual subscriber can install and run is defined by what is available for that Visual Studio subscription level in Subscriber Downloads while the user's subscription is active.

Visual Studio subscription options:

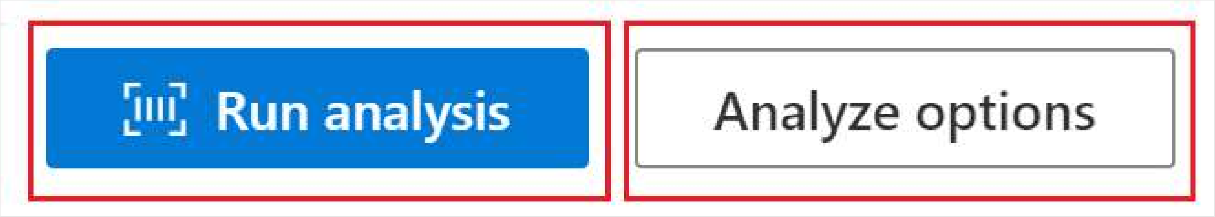
A. Standard subscriptions (sold via the Microsoft Store and Volume Licensing resellers):
- Visual Studio Enterprise Subscription (formerly MSDN)
- Visual Studio Test Professional Subscription (formerly MSDN)
- Visual Studio Professional Subscription (formerly MSDN)
- MSDN Platforms

5 © 2019 Microsoft Corporation. All rights reserved

1. On the Document Intelligence Studio home page, select **Read**

2. You can analyze the sample document or upload your own files.

3. Select the **Run analysis** button and, if necessary, configure the **Analyze options** :



Try Document Intelligence Studio

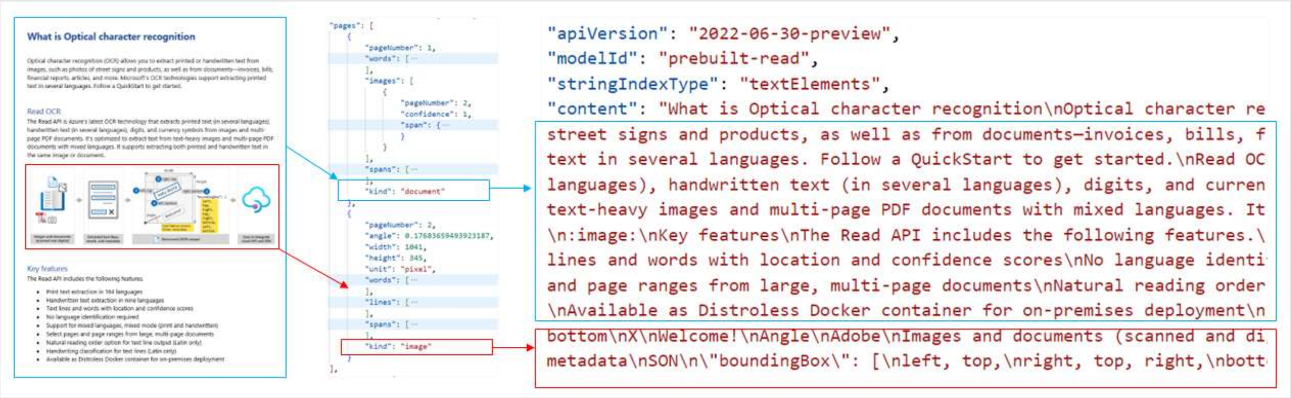
Supported extracted languages and locales

See our Language Support—document analysis models page for a complete list of supported languages.

Microsoft Office and HTML text extraction

Use the parameter `api-version=2023-07-31` when using the REST API or the corresponding SDKs of that API version to extract text from Microsoft Word, Excel, PowerPoint, and HTML files. The following illustration shows extraction of the digital text and text in the Word document by running OCR on the images. Text from embedded images isn't included in the extraction.

> ⓘ Note
>
> - **Add-on capabilities** are not supported for Microsoft Word, Excel, PowerPoint, and HTML file formats.



What is Optical character recognition

Optical character recognition (OCR) allows you to extract printed or handwritten text from images, such as photos of street signs and products, as well as from documents—invoices, bills, financial reports, articles, and more. Microsoft's OCR technologies support extracting printed text in several languages. Follow a QuickStart to get started.

Read OCR

The Read API is Azure's latest OCR technology that extracts printed text (in several languages), handwritten text (in several languages), digits, and currency symbols from images and multi-page PDF documents. It's optimized to extract text from text-heavy images and multi-page PDF documents with mixed languages. It supports extracting both printed and handwritten text in the same image or document.

Key features

The Read API includes the following features.

- Print text extraction in 164 languages
- Handwritten text extraction in nine languages
- Text lines and words with location and confidence scores
- No language identification required
- Support for mixed languages, mixed mode (print and handwritten)
- Select pages and page ranges from large, multi-page documents
- Natural reading order option for text line output (Latin only)
- Handwriting classification for text lines (Latin only)
- Available as Distroless Docker container for on-premises deployment

```
"pages": [
    {
        "pageNumber": 1,
        "words": [⋯
        ],
        "images": [
            {
                "pageNumber": 2,
                "confidence": 1,
                "span": {⋯
                }
            }
        ],
        "spans": [⋯
        ],
        "kind": "document"
    },
    {
        "pageNumber": 2,
        "angle": 0.17683659493923187,
        "width": 1041,
        "height": 345,
        "unit": "pixel",
        "words": [⋯
        ],
        "lines": [⋯
        ],
        "spans": [⋯
        ],
        "kind": "image"
    }
],
```

```
"apiVersion": "2022-06-30-preview",
"modelId": "prebuilt-read",
"stringIndexType": "textElements",
"content": "What is Optical character recognition\nOptical character re
street signs and products, as well as from documents—invoices, bills, f
text in several languages. Follow a QuickStart to get started.\nRead OC
languages), handwritten text (in several languages), digits, and curren
text-heavy images and multi-page PDF documents with mixed languages. It
\n:image:\nKey features\nThe Read API includes the following features.\
lines and words with location and confidence scores\nNo language identi
and page ranges from large, multi-page documents\nNatural reading order
\nAvailable as Distroless Docker container for on-premises deployment\n
bottom\nX\nWelcome!\nAngle\nAdobe\nImages and documents (scanned and di
metadata\nSON\n\"boundingBox\": [\nleft, top,\nright, top, right,\nbott
```

The page units in the model output are computed as shown:

⛶ Expand table

File format	Computed page unit	Total pages
Word	Up to 3,000 characters = 1 page unit, embedded or linked images not supported	Total pages of up to 3,000 characters each
Excel	Each worksheet = 1 page unit, embedded or linked images not supported	Total worksheets
PowerPoint	Each slide = 1 page unit, embedded or linked images not supported	Total slides
HTML	Up to 3,000 characters = 1 page unit, embedded or linked images not supported	Total pages of up to 3,000 characters each

Paragraphs extraction

The Read OCR model in Document Intelligence extracts all identified blocks of text in the `paragraphs` collection as a top level object under `analyzeResults`. Each entry in this collection represents a text block and includes the extracted text as `content` and the bounding `polygon` coordinates. The `span` information points to the text fragment within the top-level `content` property that contains the full text from the document.

JSON

```json
"paragraphs": [
    {
        "spans": [],
        "boundingRegions": [],
        "content": "While healthcare is still in the early stages of its Al
journey, we are seeing pharmaceutical and other life sciences organizations
making major investments in Al and related technologies.\" TOM LAWRY | National
Director for Al, Health and Life Sciences | Microsoft"
    }
]
```

The page units in the model output are computed as shown:

Expand table

File format	Computed page unit	Total pages
Images	Each image = 1 page unit	Total images
PDF	Each page in the PDF = 1 page unit	Total pages in the PDF
TIFF	Each image in the TIFF = 1 page unit	Total images in the PDF

JSON

```json
"pages": [
    {
        "pageNumber": 1,
        "angle": 0,
        "width": 915,
        "height": 1190,
        "unit": "pixel",
        "words": [],
        "lines": [],
        "spans": [],
        "kind": "document"
    }
]
```

Text lines and words extraction

The Read OCR model extracts print and handwritten style text as `lines` and `words`. The model outputs bounding `polygon` coordinates and `confidence` for the extracted words. The `styles` collection includes any handwritten style for lines if detected along with the spans pointing to the associated text. This feature applies to supported handwritten languages.

For the preview of Microsoft Word, Excel, PowerPoint, and HTML file support, Read extracts all embedded text as is. For any embedded images, it runs OCR on the images to extract text and append the text from each image as an added entry to the `pages` collection. These added entries include the extracted text lines and words, their bounding polygons, confidences, and the spans pointing to the associated text.

JSON

```json
"words": [
    {
        "content": "While",
        "polygon": [],
        "confidence": 0.997,
        "span": {}
    },
],
"lines": [
    {
        "content": "While healthcare is still in the early stages of its Al
journey, we",
        "polygon": [],
        "spans": [],
    }
]
```

Select page(s) for text extraction

For large multi-page PDF documents, use the `pages` query parameter to indicate specific page numbers or page ranges for text extraction.

> ⓘ Note
>
> For the Microsoft Word, Excel, PowerPoint, and HTML file support, the Read API ignores the pages parameter and extracts all pages by default.

Handwritten style for text lines

The response includes classifying whether each text line is of handwriting style or not, along with a confidence score. For more information, *see* handwritten language support. The following example shows an example JSON snippet.

JSON

```json
"styles": [
{
    "confidence": 0.95,
    "spans": [
    {
        "offset": 509,
        "length": 24
    }
    "isHandwritten": true
    ]
}
```

Next steps

Complete a Document Intelligence quickstart:

- ✓ REST API
- ✓ C# SDK
- ✓ Python SDK
- ✓ Java SDK
- ✓ JavaScript

Explore our REST API:

Document Intelligence API v3.1