

Transparency Note: Optical Character Recognition (OCR)



Table of Contents

What is a Transparency Note?	3
Introduction to Optical Character Recognition (OCR)	3
The basics of OCR	. 3
Example use cases	.4
Considerations when choosing other use cases	
Capabilities and limitations of OCR	4
Word-level accuracy measure	. 5
Using a confidence value	. 5
System limitations and best practices to improve system performance	. 6
Learn more about responsible AI	10
Learn more about OCR	10
Contact us	10
About this document	10

What is a Transparency Note?

An Al system includes not only the technology, but also the people who will use it, the people who will be affected by it, and the environment in which it is deployed. Creating a system that is fit for its intended purpose requires an understanding of how the technology works, what its capabilities and limitations are, and how to achieve the best performance. Microsoft's Transparency Notes are intended to help you understand how our Al technology works, the choices system owners can make that influence system performance and behavior, and the importance of thinking about the whole system, including the technology, the people, and the environment. You can use Transparency Notes when developing or deploying your own system, or share them with the people who will use or be affected by your system.

Microsoft's Transparency Notes are part of a broader effort at Microsoft to put our Al Principles into practice. To find out more, see the <u>Microsoft Al principles</u>.

Introduction to Optical Character Recognition (OCR)

Businesses today frequently need to convert text from images, scanned paper documents, and digital files into actionable insights. These insights power knowledge mining, business process automation, and accessibility of content for everyone. Optical Character Recognition (OCR) is an AI service used to extract text from visual content such as images and documents. OCR currently supports several languages for extraction of print text (see OCR supported languages). Handwritten OCR is currently supported exclusively for English.

The basics of OCR

The <u>OCR technology</u> from Microsoft is offered via the Computer Vision Read API. Customers call the Read API with their content to get the extracted text, its location, and other insights in machine readable text output. They process the output within their business applications to implement content intelligence, business process automation, and other scenarios for their users.

Term	Definition					
Asynchronous	Asynchronous means that the service doesn't immediately return the extracted text. Instead, the process starts in the background. The customer application will need to check back at a later time to obtain the extracted text.					
Read	The Read operation is an asynchronous call that accepts images and documents to begin analysis and text extraction, which is returned via another call.					
Get Read Results	While the analysis and extraction process is active, the Get Read Results operation outputs the progress status. When the process is complete, the Get Read Results operation outputs the extracted text (in the form of text lines and words) and confidence values.					
Confidence value	The Get Read Results operation returns confidence values in the range between 0 and 1 for all extracted words. This value represents the service's estimate of how many times it correctly extracts the word out of 100. For example, a word that's estimated to be extracted correctly 82% of the time will result in a confidence value of 0.82.					

Example use cases

The following use cases are popular examples for the OCR technology.

- **Images and documents search and archive** Unstructured documents such as legal contracts, technical documents, and news content contain rich information and metadata that are not available for processes such as automated tagging, categorization, and search. OCR allows the text from these documents to be machine readable for analysis, search, and retrieval.
- **Image content moderation and localization** eCommerce companies, user-generated content publishers, and online gaming and social media communities need to moderate images to be compliant with online safety regulations. In certain cases, they also need to localize content for international audiences. OCR allows you to extract text from images to apply downstream processing.
- **Business process automation** Business process automation requires integrating user-entered data and preferences in documents and application screens with complex business processes. OCR unlocks the text embedded in documents and images and makes it available for use in the steps of the business workflows.
- **Financial and healthcare documents processing** When used in back office processing of financial and insurance application forms, OCR helps save time and effort in document processing. Similarly, OCR applied to medical claim reimbursements and medical information forms speeds up reimbursements and qualification for services and benefits.

Considerations when choosing other use cases

Consider the following factors when you choose a use case.

- **Carefully consider when using for awarding or denying of benefits** Using OCR output directly for awarding or denying benefits can result in errors if based on incorrect or incomplete information. For example, when filling out medical forms, users can make errors or fail to include important information. Additionally, OCR can potentially misread or not detect parts of the form. To ensure fair and high-quality decisions for consumers, combine OCR-based automation with human oversight.
- Avoid use for signature identification When you extract handwritten text, avoid using the OCR results on signatures to identify individuals. Signatures are hard to read for humans and machines alike. A better way to use OCR is to use it for detecting the presence of a signature for further analysis.
- Don't use OCR for decisions that may have serious adverse impacts Examples of such use cases include processing medical prescriptions and dispensing medication. The machine learning models that extract text from prescriptions can result in undetected or incorrect text output. Decisions based on incorrect output could have serious adverse impacts. Additionally, it is advisable to include human review of decisions that have the potential for serious impacts on individuals.

Capabilities and limitations of OCR

In this section, we'll review what accuracy means for OCR and how to assess it for your context.

Word-level accuracy measure

Text is composed of lines, words, and characters. A popular measure of accuracy for OCR is word error rate (WER), or how many words were incorrectly output in the extracted results. The lower the WER, the higher the accuracy.

WER is defined as:

$$WER = rac{S+D+I}{N} = rac{S+D+I}{S+D+C}$$

Term	Definition	Example
S	Count of incorrect words ("substituted") in the output.	"Velvet" gets extracted as "Veivet" because "I" is detected as "i."
D	Count of missing ("deleted") words in the output.	For the text "Company Name: Microsoft," Microsoft isn't extracted because it's handwritten or hard to read.
I	Count of nonexistent ("inserted") words in the output.	"Department" gets incorrectly segmented into three words as "Dep artm ent." In this case, the result is one deleted word and three inserted words.
С	Count of correctly extracted words in the output.	All words that are correctly extracted.

Using a confidence value

As covered in an earlier section, the service provides a confidence value for each predicted word in the OCR output. Customers use this value to calibrate custom thresholds for their content and scenarios to route the content for straight-through processing or forwarding to the human-in-the-loop process. The resulting measurements determine the scenario-specific accuracy.

OCR system performance implications can vary by scenarios where the OCR technology is applied. We'll review a few examples to illustrate that concept.

Medical device compliance: In this first example, a multinational pharmaceutical company with a diverse product portfolio of patents, devices, medications, and treatments needs to analyze FDA-compliant product label information and analysis results documents. The company might prefer a low confidence value threshold for applying human-in-the-loop because the cost of incorrectly extracted data can have significant impact for consumers and fines from regulatory agencies.

Image and documents processing: In this second example, a company performs insurance and loan application processing. The customer using OCR might prefer a medium confidence value threshold because the automated text extraction is combined downstream with other information inputs and human-in-the-loop steps for a holistic review of applications.

Content Moderation: For a large volume of e-commerce catalog data imported from suppliers at scale, the customer might prefer a high confidence value threshold with high accuracy because even a small percentage of falsely flagged content can generate a lot of overhead for their human review teams and suppliers.

System limitations and best practices to improve system performance

The service supports images (JPEG, PNG, and BMP) and documents (PDF and TIFF). The allowable limits for number of pages, image sizes, paper sizes, and file sizes are listed on the OCR overview page.

Document scan quality, resolution, contrast, light conditions, rotation, and text attributes such as size, color, and density can all affect the accuracy of OCR results. For example, we recommend the image be at least 50 X 50 pixels. Customers should refer to the product specifications and test the service on their documents to validate the fit for their situation.

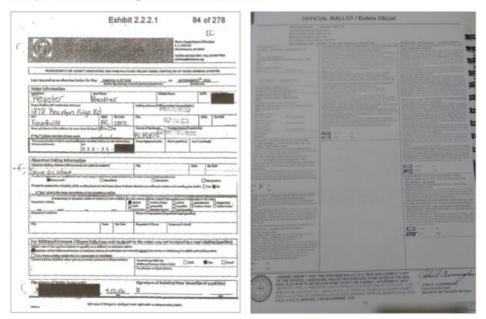
The following example shows a few difficult cases for OCR where you see missed and incorrect text extractions.



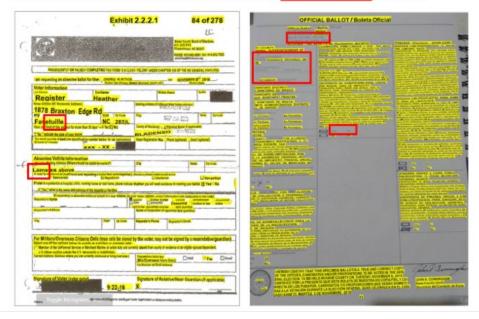
Receipt images with increasing blurring and nissed extractions

Contoso	Contoso	Contoso
Contoso 122 Wain Street Redirent, WA 98052	Contose 121 Main Street Redmond, VA 98052	Contoxo 122 Hain Street Reditiond, NA 98052
123-456-7890 510 2019 13:59 Safes Associate, Paul	122-456-7290 6102019 13-59 Salts Associate Paul	122-456-7890 616/2019 13:59 Sales Associate: Paul
ESurface Pre 6 25008 Lintel Care 15 / IGBRAM (Black) E 999.00	1 Surface Pro 6 2566B lintel Core IS 7 178 RAN (Black) 5999.00	1 Surface Pro 6 25558 Intel Core IS / 058 RAM (Black) 5999.00
1. SurfacePen (\$99.99)	1 SurfacePen \$ 99.99	1 SurfacePen
Sub-Total \$ 1098.99 Tax \$ 1094.40	Side-Total 7 1008.99	Ria Total
Total \$ 1203.30	Talat Bulga	Intel Constanting

Blurry, low resolution images with handwritten text



Blurry, low resolution images with handwritten text with missed or incorrect extraction



Blurry, dense, small text and low-resolution images with handwritten text

Rughed Loca	Variable Trust U.S. Government More art Waterty, 4172 (inge	Wingthest Average UKr 12.00 mays	104	81.17(W.111.081				QC2012004-5	NO. THESAULT
Call AN RUD TO LO AND RUD TO LO AND	Bit House States 21 States States 22 States States 23 States 24 States States 25 States 26 States 27 States 28 States 29 States 20 States 21 States 22 States 23 States 24 States 25 States 26 States	Band Constant of the second of the Second of the second of the Second of the second of the Second of the second of the Constant of the Constant of the Second of	Participal Recent Control 8	and interesting logs W Balance Balance Balance	1,000 1,	Appendix Appendix 30.7.00 30.7.00 30.7.00 30.7.00 40.7.00 30.7.00 40.00 30.7.00 40.00 30.7.00 40.00 30.7.00 40.00 30.7.00 40.00 30.7.00 40.00 30.7.00 40.00 30.7.00 40.00 30.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.00 40.7.00 40.	2019 2019 2019 2019 2019 2019 2019 2019	1 Descentifie Fields of the Technology Braining (Technology) (Technology) (Technology) Braining (Technology) (Technology) (Technology) Braining (Technology) (Technology) (Technology) Braining (Technology) (Technology) (Technology) (Technology) (Technology) (enter design a social contra Marginary Alta contrata contra en an de partenana bench ad editos en an menter han a de diga par falas a
 Marci Unit. 	 J. Conservations Reprint Data Discontrational Agency Data 	FEDERAL HORE LOAVE ANY BRIER FEDERAL HORE LOAVE RIVER FEDERAL HORE LOAVE RIVER	81.424.000 82.204.000 82.004.000 82.04.000 82.04.000 82.04.000 82.04.000	BE 200.000 20.000.000 80.000.007 80.000.074 Bein 2010 Bein 2010 Barrowick	1.40% 1.00% 1.00% 1.40% 1.40% 1.40%	202198 2020-98 2020-98 2020-98 2010-98 2010-98 2010-98	622/16 622/16 5420/16 5420/16 5420/16 242/16 242/16	Prover Product III Scotler 1 III Scotler II III Scotler III	D three Factor 1 D three Factor 1

Blurry, dense, small text and low-resolution images with handwritten text with missed extractions.

	and the second se	Regard Reampt Lie; 72.50 days	A REAL PROPERTY AND A REAL	UNALLINE, DOCKST.					UNE CHECKLIEF
-	Investment Cusepory	heart			Yold to Matura W		furth Data	1 Demographic Profile of the Toocher Brog	and the
POUP 1	U.S. Treasury-Reportment Josef and	P WORDAN CHURCH 6 CD	\$100.000 AND	C 400.483	1.00	2018	2018	Streeting The supported in concess on	ducing a research andbig. "MATHEMATICS
THEFT	0.5. Treesury Report task Agroom with U.S. Treesury Datab	THEASURY days	10.600.000		1786	4919	414.14	ANTIGUTINE OF LAS. MAN	VUEL DISTRICT THROUGH THE USE OF
TORLAS.	or & Treasury Tareet	THEASURY DOLL	\$2,400,000	81,558 125	1275		10110		The construction includes and willing any rests
TRADET	U.S. Secondary Lineary Sale	emetode mac	\$2,440,000	to and THY	120%	-	2011	incompliance for bottom had then be	propriate boass or by filing in the Utarias. You
00-070	U.S. Germanant Agency Date	FRANCE MART	83, 108,000	\$2 and all'S	1.43%	anave.	312118	reasoning will be impled with use	and the second of the second second second
	U.S. fucumment suprov Date	FRANKE MALT	6435 000	5400.574	1100	3 (7 WE	Augure a		
100548-118	U.S. Gavementant Agency Datet	FANNE MLE	BILTON MAN	81 TML 100	1,000	autors a	10110	tere optional child	
104664	U.S. Geveninger Agency Debt	FANNE DECOUNT NOTE	\$7,500,000	1748(875)	1.18%	378-18	Armin .	My School	
100.0018	U.S. Gronmanas Jamery Cabl	FREDERE MAC	\$738.004	\$727.647	1.125	* Dente	STAFTS	My School	
10800118	U.S. Gevenement Spency Debit	RECEIVAL MAIN CREDIT BANK	\$210.000	5216 196	1.00	20010	228718	2 36 31 mm and	
1003473	8-3. Gevennune Agency Debt	FEDERAL FARM CREDIT DAVID FEDERAL FARM CREDIT DAVID	BH 100.000	A1.001 (00)	1100	21112	01010	2 20 25 years old	O 46 SEyembold
100.005	U.S. Generational Reports Debt	PETERAL FARM CREDIT BANK	\$7 400 MM	2 011 545	1479	10710	ACCOUNTS OF	© 26 10 years old	0 S1 85 years old
STRUE.	U.S. Gerennungen Augenry Debt.	FEDERAL FARM CREDIT BANK	\$100.000	Sine pp	1475	42278	10010		
MFECT -	U.S. Geveninget Agency Debt	FEDERAL PARM CREDIT BANK	100.000	4404 TH	1.276	11070	4114-18	@ 31 -35 years old	0 56 dd yware oki
diright in	U.S. Generation Americant	PECENAL NAME CREDIT BANK	\$540,000	\$540 208	1.000	0.0110	6.010	036 til yter eki	
108242	U.S. Gavammant Agency Detail	RECERAL FARM CREDIT DAME	31.400.000	1.900 145	1.475	A1227118	10010	U30 E years and	D 81 years old and above
DCEND+	111 Cavaranaed Bostov Date:	REDICTAL FARM CREDIT BANK	\$3,256,004	\$5,258,577	1.005		208411	4 47 - 45 man old	
011.04	U.S. Government Assecs Debt	REDENAL FARM CREDIT BANK	\$1,400,000	81.009.301	1275	2114	421510-146	performance in the local distance in the loc	
1048072	U.S. Geransmant Agency Dateb	REDERAL FARM CREDITBANK	\$128,010	STOR OF STREET	120%	10.00	924	Sinder ED.11.4	C norse
CIADONA CONTRACTOR	U.S. Geveningert Assesse Debt	TED HONE UNDECOUNTINT	\$150,000	64 708 45 1	1305	22124	20 10 10	Chil Status	
100.01	9.8 Coversmant Asserve Selfe	FED HORE UN DECOUNT NT	10,770,000		100	ATMANT OF		C. B. Balan	
the surgers.	U.S. Growney damery Tall	FED HOME US DECKINT WT	\$400.000	\$100 BOD	1205	2010	1.0414	Q Sirek	Marian
totil as	Or A Severalment Advence Cashi	HED HOME UN DECOUNTINT	\$5,400,000	14 000 070	125	2-019	2010		
Distance and	U.S. Gavement Aparicy Debt	FED HOME IN DECOUNTINT	\$1 N00 000	61 015 814	1775	COMPRESS OF	1218-16	Separated	O Witceniers
THUR WE	E. E. Gaugement & server, Date	PEDERAL HOME LOAN BANK	800.000	210 mm	1.475	ageres .	ALC: NO.	Highest Educational Atlantment.	
0.1419-0.	Mill: Gevenement Assercy Debit	FEDERAL HOMELOAN BANK	8450.000	848.005	1376	21010	2/8-18		
TISNS	U.S. Gavenmant Agency Debt	FEDERAL HOME LOAN BANK	\$100.000	\$162,067	1.545	216118	2816	Beshvior's Degree Graduate	EL OPh & Graduate
HKINGP	U.S. Gevenaturet Againsy Dabit	FEDERAL FARM CREDIT DANK	\$715.006	Exits Adv	1.05	wine	818-18		
4004	U.S. Gavemeant Agency Debt	FEDERAL HOME LOAN DAME	\$190,000	Bå dan sen	100	States and	Writers .	C MAEd Distuite	B Offers specify:
ACUNE	9.5. Gavemented Assency Debt	FEDERAL HOMELOAN DANK	\$3 400-000 97 400-000	21.000.000	158%	1000.00	11000	Present Prantices	
AC UNIT	0.5. Gevenment Assercy Debt 0.5. Gevenment Assercy Debt	PEDERAL HOME LOAN BANK	10,000,000	an annual second	1.475		800448		
	V.B. Geveniment Adamsy Debt	PEDERAL HOMELOAN BANK	\$2,408,004	1,880,067	1.875	ADDRESS .	A100710	Terrator N	D Master Teactor 1
Attents	U.S. Constanting Labors Cable	REDERIAL HOME LOAN BANK	\$100.000	\$750.073	5.525	20445	activated in the		
Statute of the	E.S. Gaugement Supercy Date	PERSONAL NEW COMPENSATION	Walk care	8000.013	1.010	100000	10110	© Teacher III	Martin Tractor 7
EASI-LS	U.S. Geventrant-Agency Debt	FEDERAL HOME LOWB BROKE	\$100.000	Salah and	1505	and the	212.18	O Tenter III	Control register
ATCOD.	U.S. Gevenement Agency Debt	RECEIVAL HOMELOW BANK	\$400.000		1405	21818	2818		and the second sec
GA1081	U.S. Government Agency Debt	REDERAL HOMELOAN BANK	\$276,000	50198.927	1276	21818	31818		

Receipt images with perspective and surrounding white space impact



Receipt images with perspective and surrounding white space impact with missed extractions.



The current version supports handwriting or cursive-style text only for English. This limitation also affects any related features such as print vs. handwriting-style classification (preview) for each text line.

OCR's performance will vary depending on the real-world uses that customers implement. In order to ensure optimal performance in their scenarios, customers should conduct their own evaluations of the solutions they implement using OCR. The service provides a confidence value in the range between 0 and 1 for each detected word included in the OCR output. Customers should scan a sample dataset representing their content to get a sense of the range of confidence scores and the resulting extraction quality. They can then decide on the confidence value thresholds to determine whether the results should be sent for straight-through-processing (STP) or reviewed by a human. For example, the customer may submit results with confidence value greater than or equal to .80 for straight through processing, and apply human review to results with confidence value less than .80.

Learn more about responsible AI

<u>Microsoft AI principles</u> <u>Microsoft responsible AI resources</u> <u>Microsoft Azure Learning courses on responsible AI</u>

Learn more about OCR

What is Optical Character Recognition?

Contact us Give us feedback on this document

About this document

© 2021 Microsoft Corporation. All rights reserved. This document is provided "as-is" and for informational purposes only. Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it. Some examples are for illustration only and are fictitious. No real association is intended or inferred.

Published: 08/20/2021

Last updated: 08/20/2021