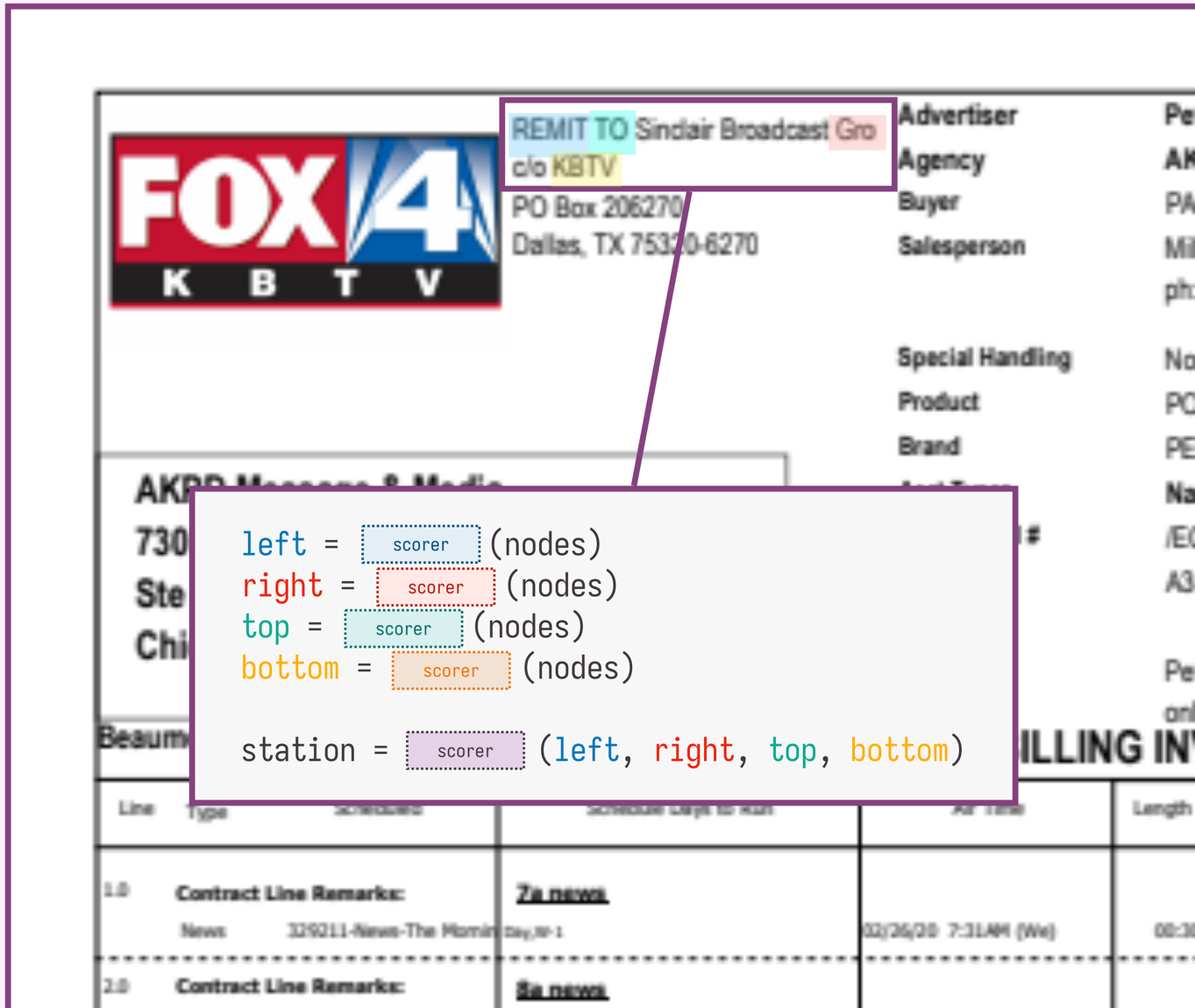
Program Sketches with Scorer Holes

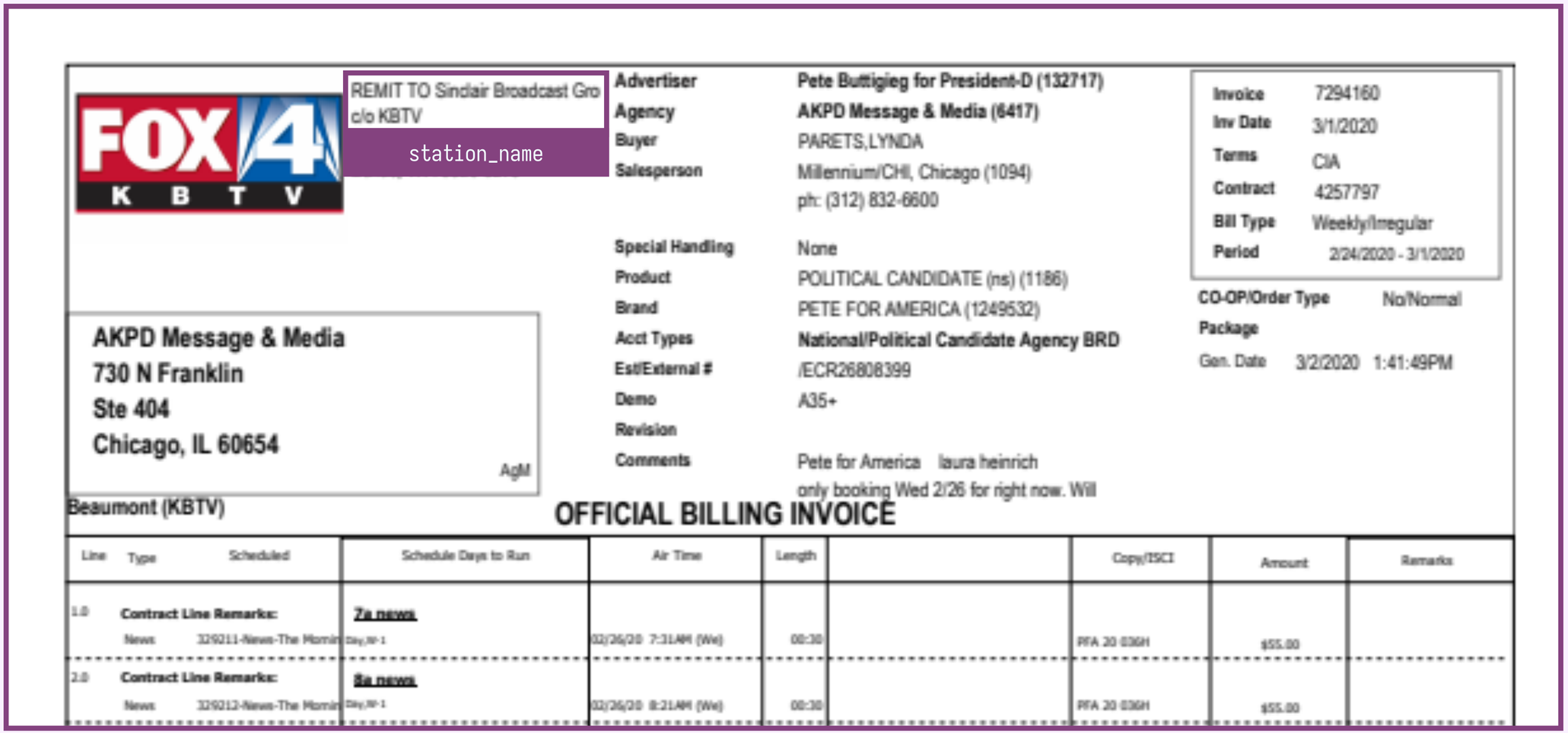
David Minh-Duy Cao, Sarah E. Chasins

April 16, 2025 • EPIC Retreat



Motivation:

Folks are often charged with getting information out of **templated** PDF documents



FOX 4 KBTV

REMIT TO Sinclair Broadcast Gro
c/o KBTV

station_name

Advertiser	Pete Buttigieg for President-D (132717)
Agency	AKPD Message & Media (6417)
Buyer	PARETS,LYNDA
Salesperson	Millennium/CHI, Chicago (1094) ph: (312) 832-6600
Special Handling	None
Product	POLITICAL CANDIDATE (ns) (1186)
Brand	PETE FOR AMERICA (1249532)
Acct Types	National/Political Candidate Agency BRD
Est/External #	/ECR26808399
Demo	A35+
Revision	
Comments	Pete for America laura heinrich only booking Wed 2/26 for right now. Will

Invoice	7294160
Inv Date	3/1/2020
Terms	CIA
Contract	4257797
Bill Type	Weekly/Irregular
Period	2/24/2020 - 3/1/2020
CO-OP/Order Type	No/Normal
Package	
Gen. Date	3/2/2020 1:41:49PM

AKPD Message & Media
730 N Franklin
Ste 404
Chicago, IL 60654
AgM

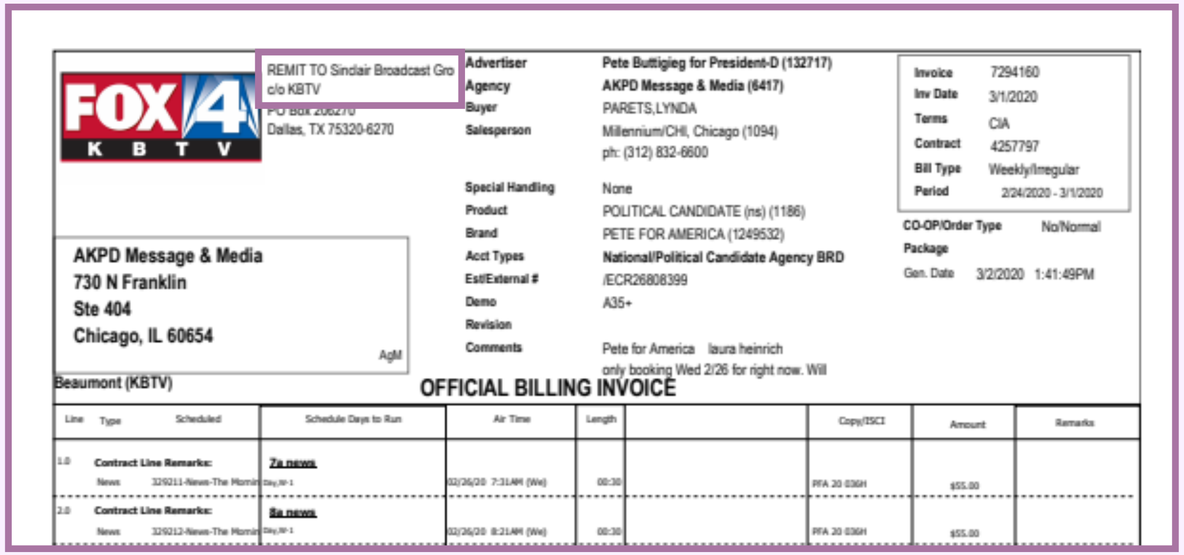
Beaumont (KBTV)

OFFICIAL BILLING INVOICE

Line	Type	Scheduled	Schedule Days to Run	Air Time	Length		Copy/ISCI	Amount	Remarks
1.0	Contract Line Remarks:		7a news						
	News	329211-News-The Mornin	Day;M-1	02/26/20 7:31AM (We)	00:30		PFA 20 030H	$55.00	
2.0	Contract Line Remarks:		8a news						
	News	329212-News-The Mornin	Day;M-1	02/26/20 8:21AM (We)	00:30		PFA 20 030H	$55.00	

Motivation:

Folks are often charged with getting information out of **templated** PDF documents that are **heterogeneous** (differences in formatting or alignment, PDF artifacts, ...).



FOX 4 KBTV

REMIT TO Sinclair Broadcast Gro
c/o KBTV
PO Box 206270
Dallas, TX 75320-6270

AKPD Message & Media
730 N Franklin
Ste 404
Chicago, IL 60654
AgM

Advertiser: Pete Buttigieg for President-D (132717)
Agency: AKPD Message & Media (6417)
Buyer: PARETS,LYNDA
Salesperson: Millennium/CHI, Chicago (1094)
ph: (312) 832-6600

Special Handling: None
Product: POLITICAL CANDIDATE (ns) (1186)
Brand: PETE FOR AMERICA (1249532)
Acct Types: National/Political Candidate Agency BRD
Est/External #: /ECR26808399
Demo: A35+
Revision:
Comments: Pete for America laura heinrich
only booking Wed 2/26 for right now. Will

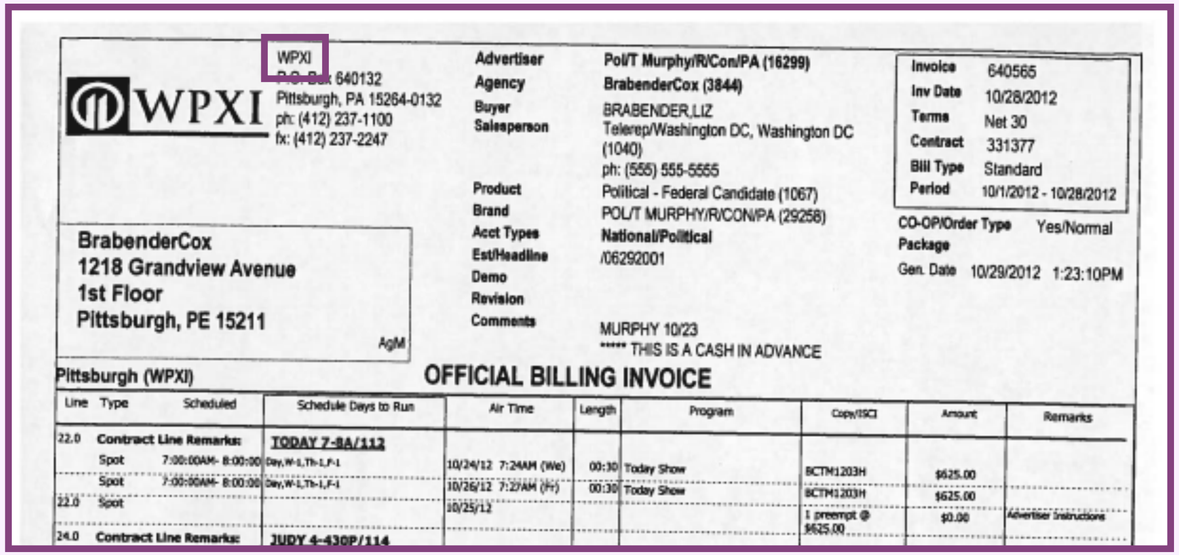
Invoice 7294160
Inv Date 3/1/2020
Terms CIA
Contract 4257797
Bill Type Weekly/Irregular
Period 2/24/2020 - 3/1/2020

CO-OP/Order Type No/Normal
Package
Gen. Date 3/2/2020 1:41:49PM

Beaumont (KBTV)

OFFICIAL BILLING INVOICE

Line	Type	Scheduled	Schedule Days to Run	Air Time	Length		Copy/ISCI	Amount	Remarks
1.0	Contract Line Remarks:		7a news						
	News	329211-News-The Morni	Day,M-1	02/26/20 7:31AM (We)	00:30		PFA 20 036H	$55.00	
2.0	Contract Line Remarks:		8a news						
	News	329212-News-The Morni	Day,M-1	02/26/20 8:21AM (We)	00:30		PFA 20 036H	$55.00	



WPXI
P.O. Box 640132
Pittsburgh, PA 15264-0132
ph: (412) 237-1100
fx: (412) 237-2247

BrabenderCox
1218 Grandview Avenue
1st Floor
Pittsburgh, PE 15211
AgM

Advertiser: Pol/T Murphy/R/Con/PA (16299)
Agency: BrabenderCox (3844)
Buyer: BRABENDER,LIZ
Salesperson: Telerep/Washington DC, Washington DC (1040)
ph: (555) 555-5555
Product: Political - Federal Candidate (1067)
Brand: POL/T MURPHY/R/CON/PA (29258)
Acct Types: National/Political
Est/Headline: /06292001
Demo:
Revision:
Comments: MURPHY 10/23
***** THIS IS A CASH IN ADVANCE

Invoice 640565
Inv Date 10/28/2012
Terms Net 30
Contract 331377
Bill Type Standard
Period 10/1/2012 - 10/28/2012

CO-OP/Order Type Yes/Normal
Package
Gen. Date 10/29/2012 1:23:10PM

Pittsburgh (WPXI)

OFFICIAL BILLING INVOICE

Line	Type	Scheduled	Schedule Days to Run	Air Time	Length	Program	Copy/ISCI	Amount	Remarks
22.0	Contract Line Remarks:		TODAY 7-8A/112						
	Spot	7:00:00AM- 8:00:00	Day,W-1,Th-1,F-1	10/24/12 7:24AM (We)	00:30	Today Show	BCTM1203H	$625.00	
	Spot	7:00:00AM- 8:00:00	Day,W-1,Th-1,F-1	10/26/12 7:27AM (Fr)	00:30	Today Show	BCTM1203H	$625.00	
22.0	Spot			10/25/12			1 preempt @ $625.00	$0.00	Advertiser Instructions
24.0	Contract Line Remarks:		JUDY 4-430P/114						

Motivation:

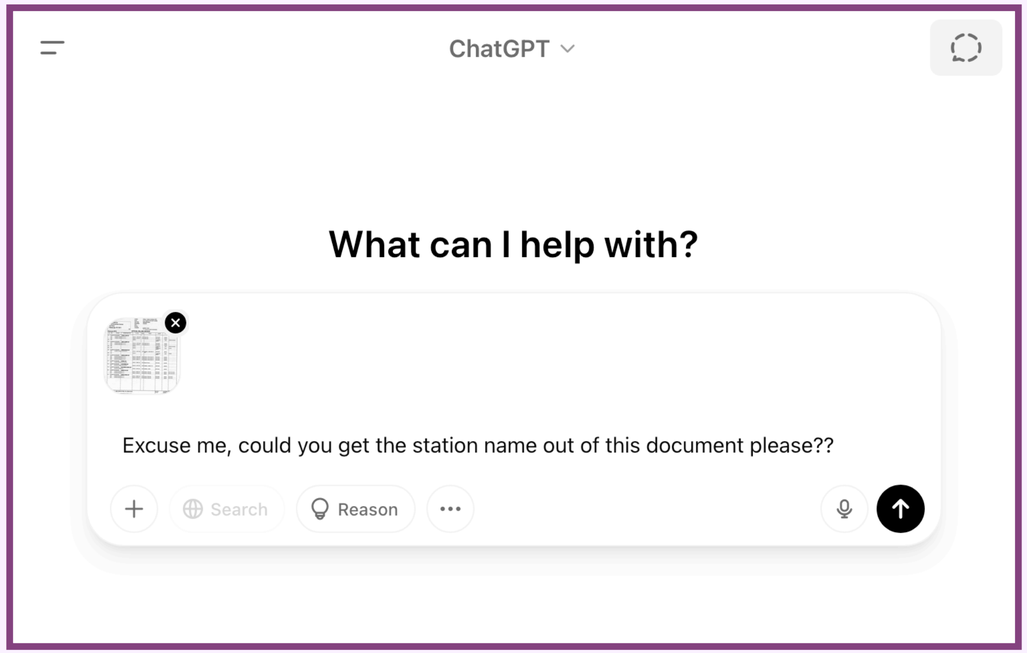
Folks are often charged with getting information out of **templated** PDF documents that are **heterogeneous** (differences in formatting or alignment, PDF artifacts, ...). and **numerous** (hundreds or thousands of documents).

and 1,225 more

Under-resourced folks lack the tools to process messy documents at scale.

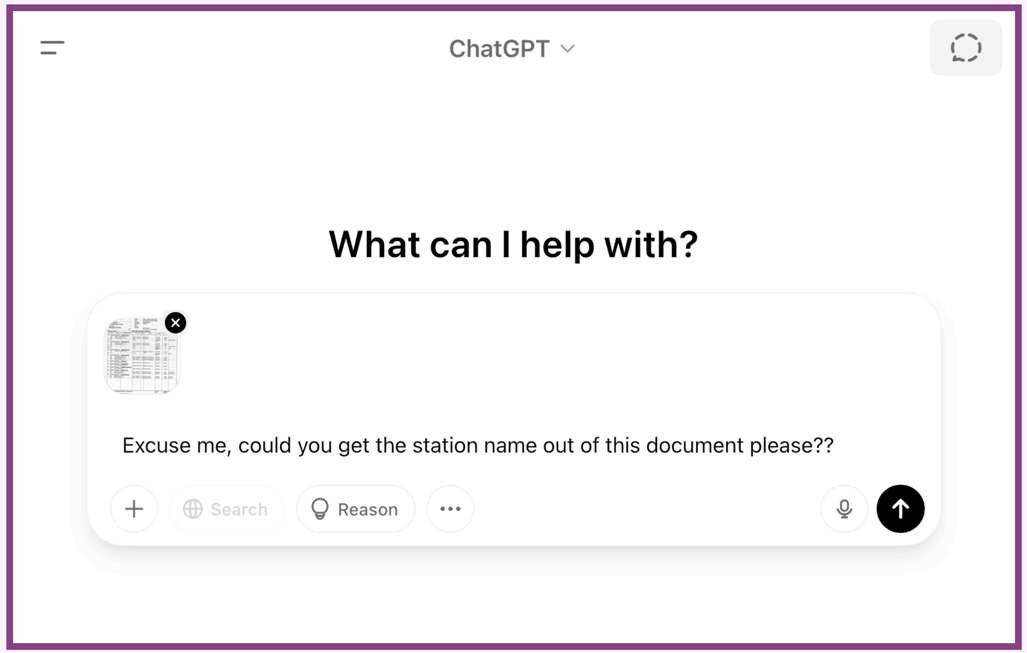
Under-resourced folks lack the tools to process messy documents at scale.

LLMs (e.g., GPT-4o)?



Under-resourced folks lack the tools to process messy documents at scale.

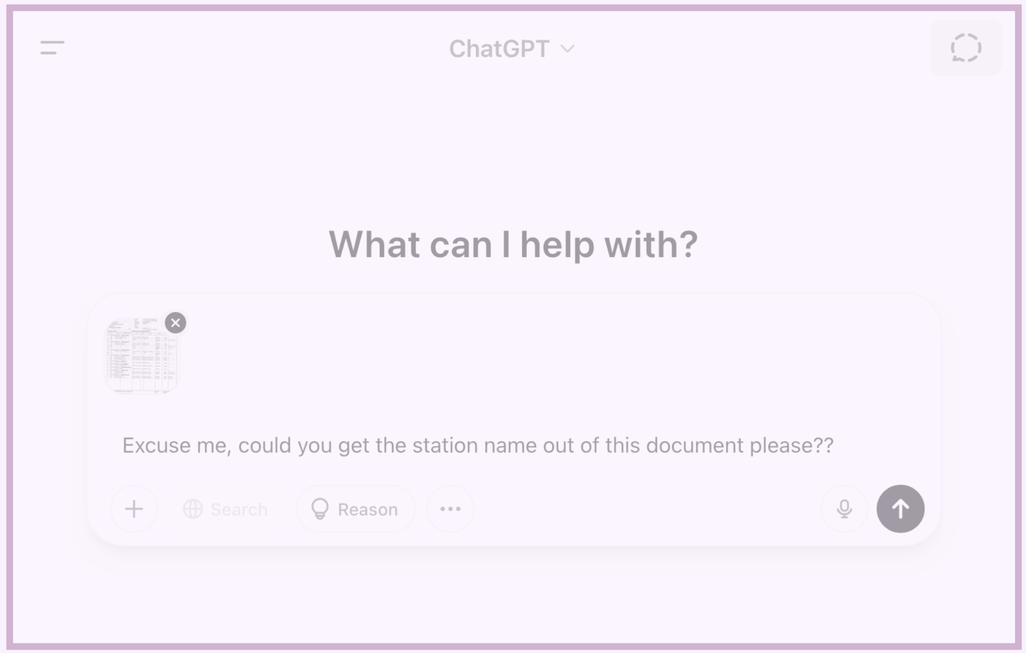
LLMs (e.g., GPT-4o)?



Running our benchmark suite of **1,227 documents** took **≥14.5 hours** and cost us **$1,675.05.**

Under-resourced folks lack the tools to process messy documents at scale.

LLMs (e.g., GPT-4o)?



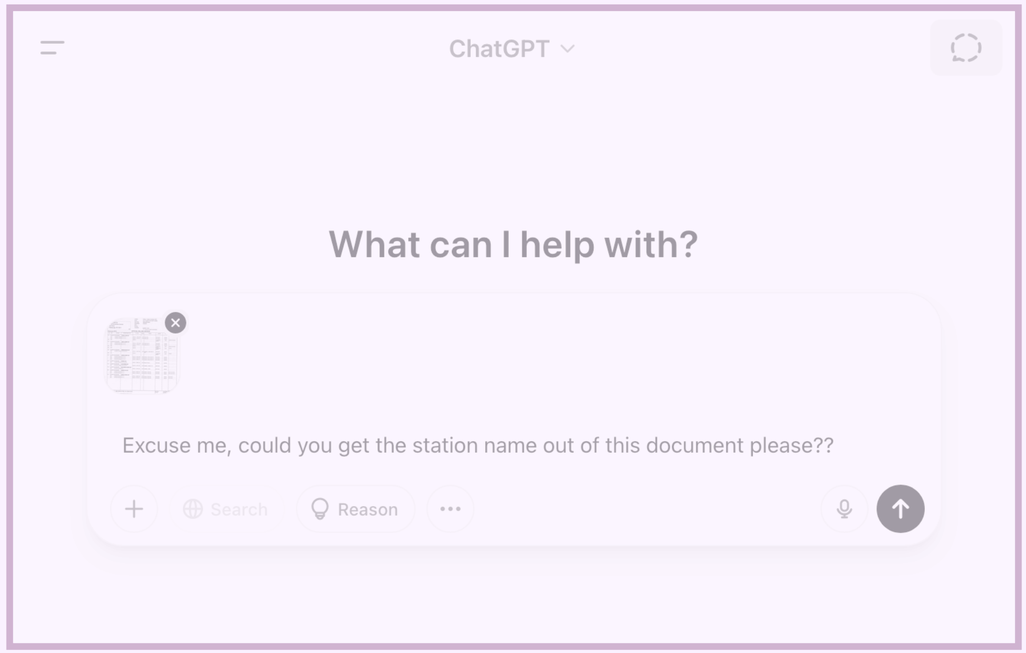
Traditional program synthesis?

```
def extract_info(doc):
    line = get_line(4, doc)
    label = search_for("Station:", line)
    return rest_of_line(label)
```

Running our benchmark suite of **1,227 documents** took **≥14.5 hours** and cost us **$1,675.05.**

Under-resourced folks lack the tools to process messy documents at scale.

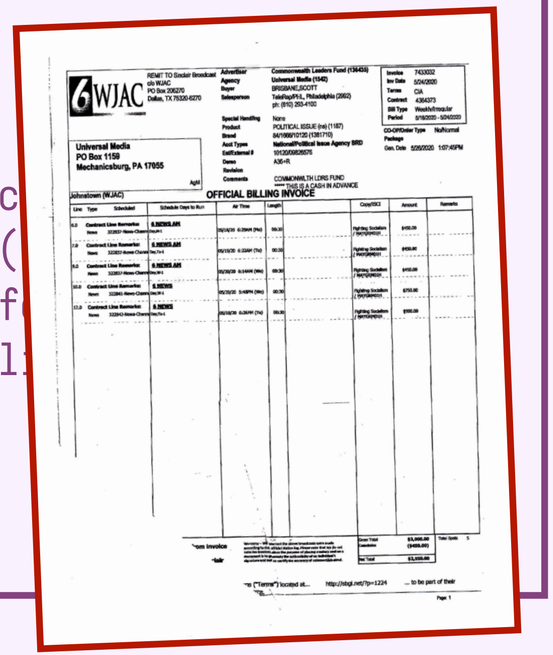
LLMs (e.g., GPT-4o)?



Running our benchmark suite of **1,227 documents** took **≥14.5 hours** and cost us **$1,675.05.**

Traditional program synthesis?

```
def extract_info(doc
    line = get_line(
    label = search_f
    return rest_of_l
```

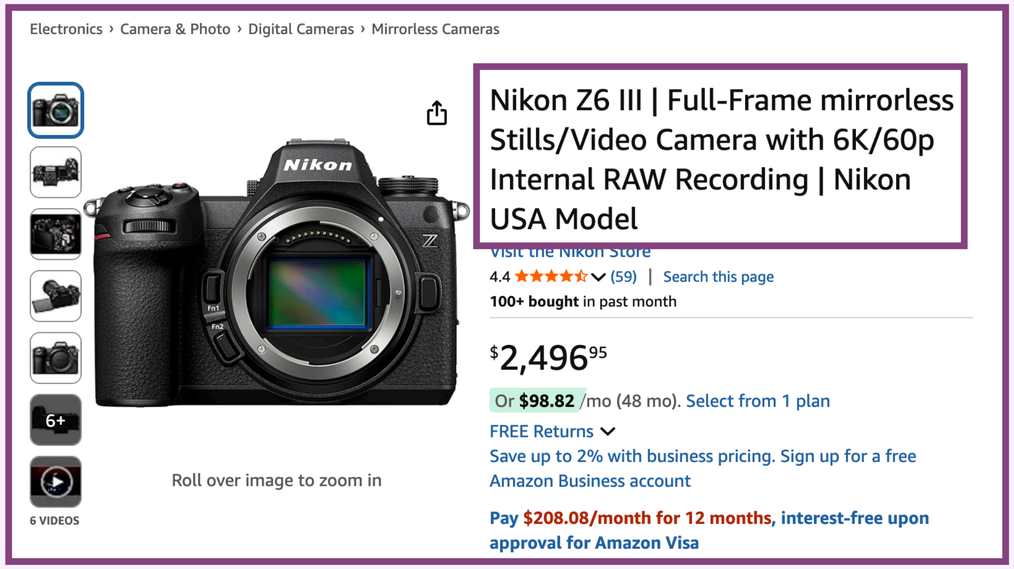


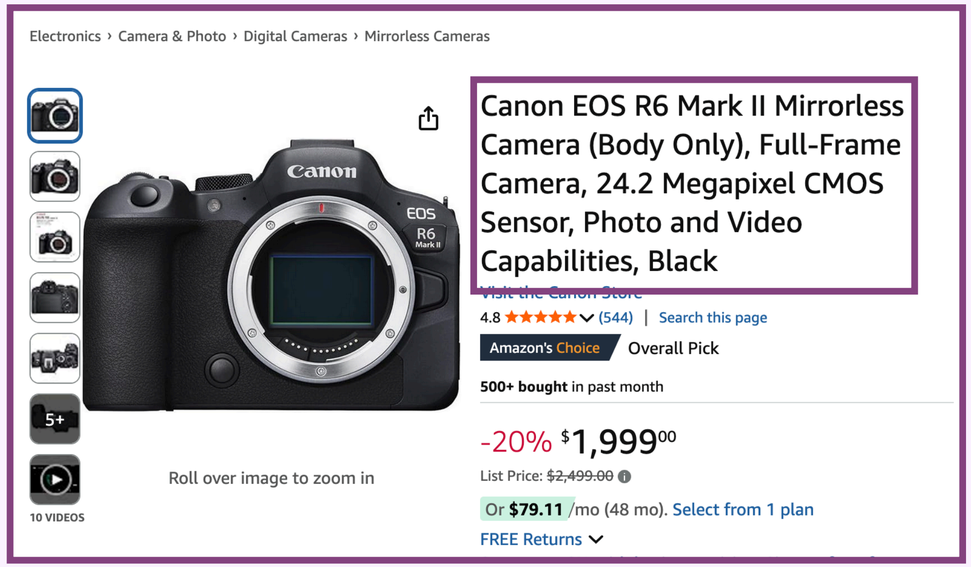
Traditional synthesis techniques **don't work** with messy PDF data.

Can we adapt traditional synthesis techniques for messy data domains?

Can we adapt traditional synthesis techniques for messy data domains?

Let's use insights from Ringer and web automation.

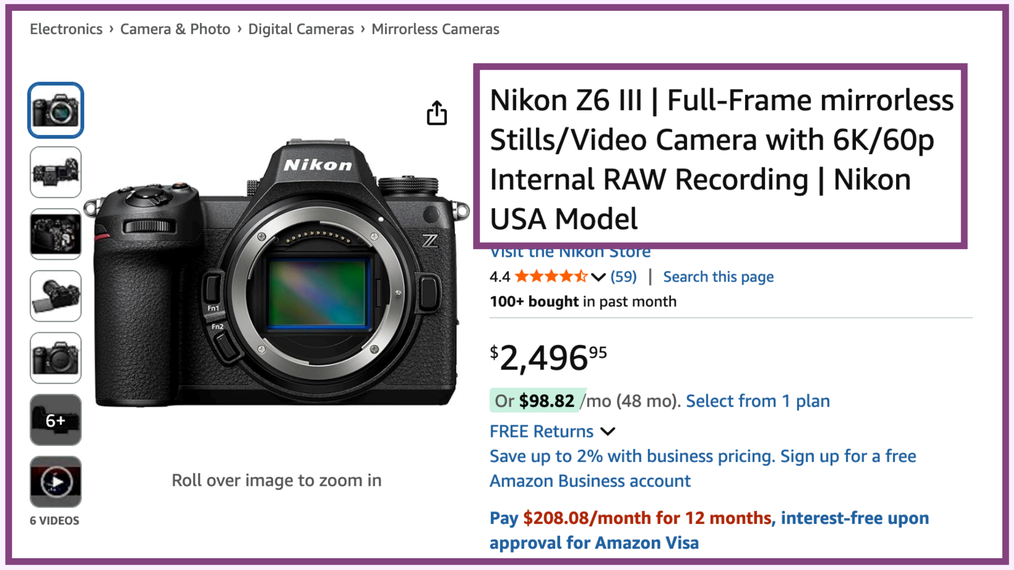


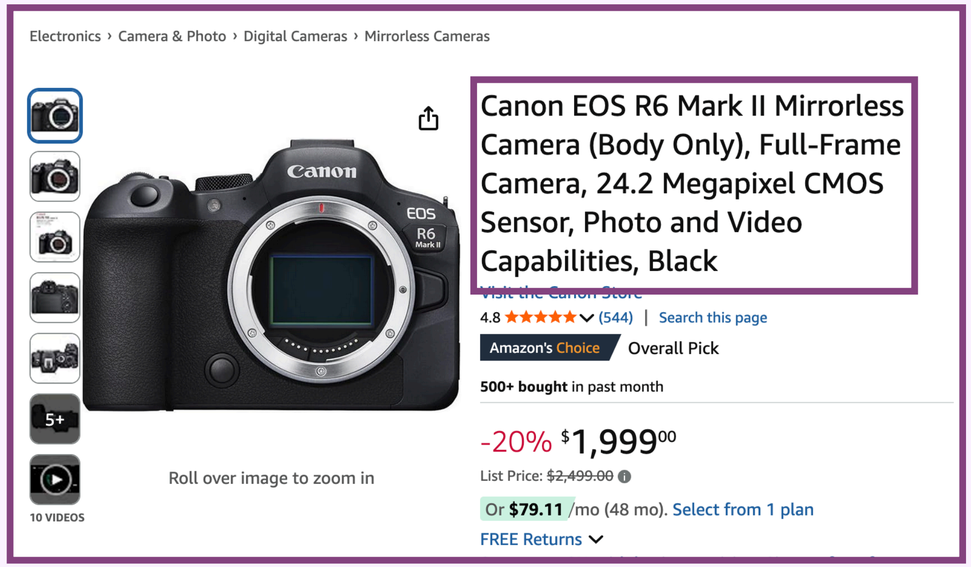


Can we adapt traditional synthesis techniques for messy data domains?

Let's use insights from Ringer and web automation.

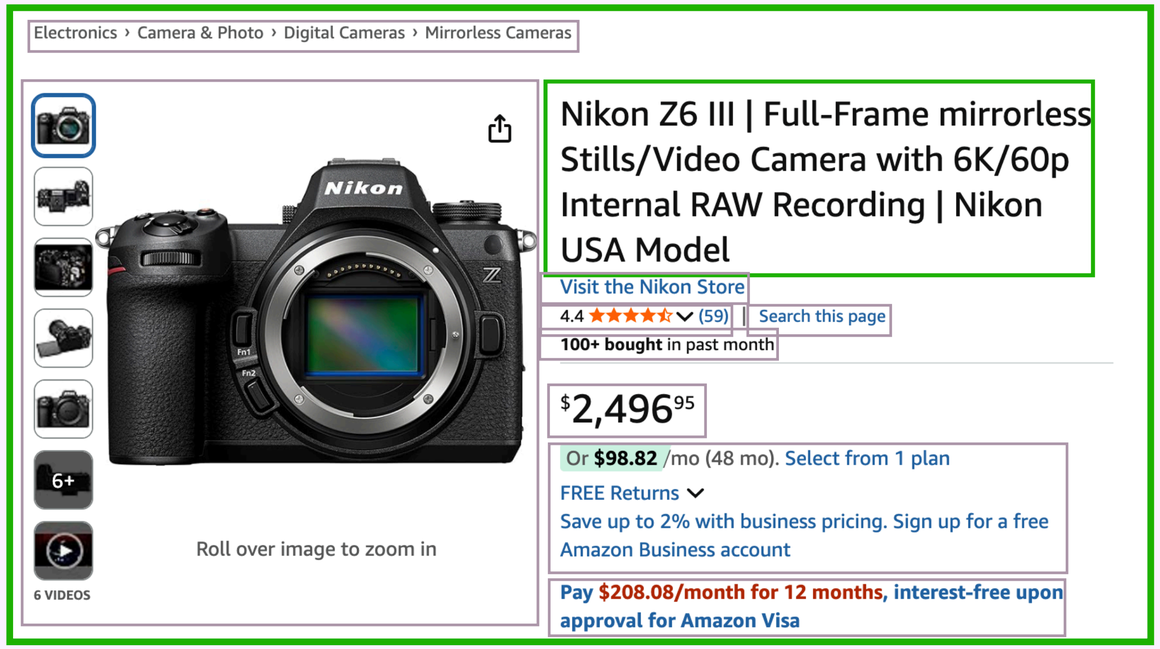
Ringer's **scorers** contend with webpage heterogeneity.



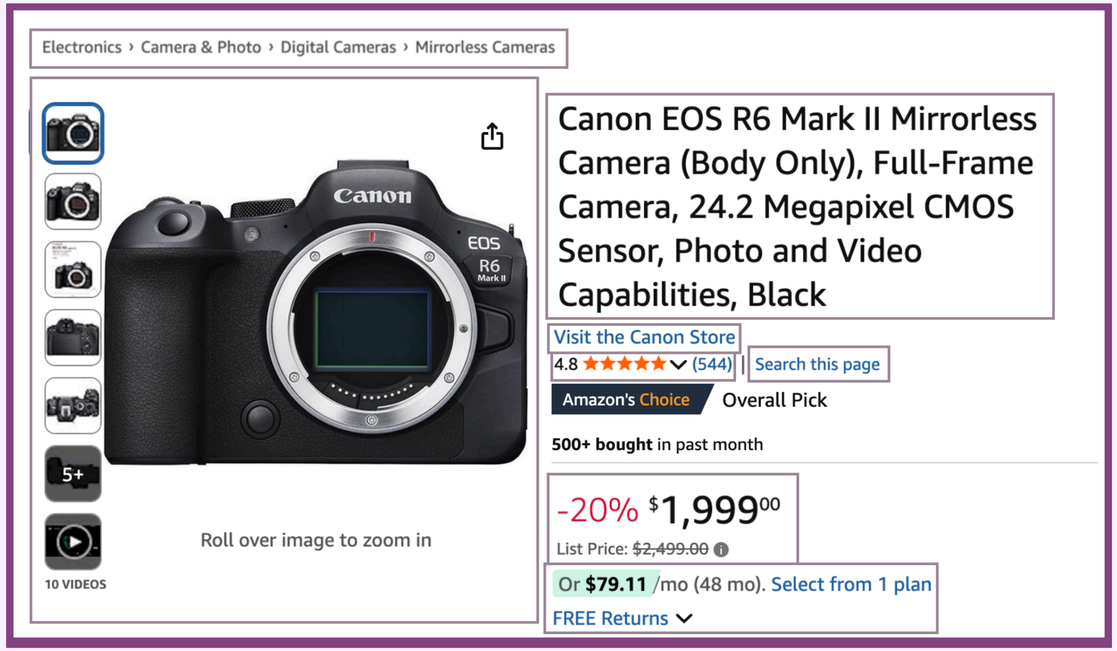


Ringer's scorers contend with webpage heterogeneity.

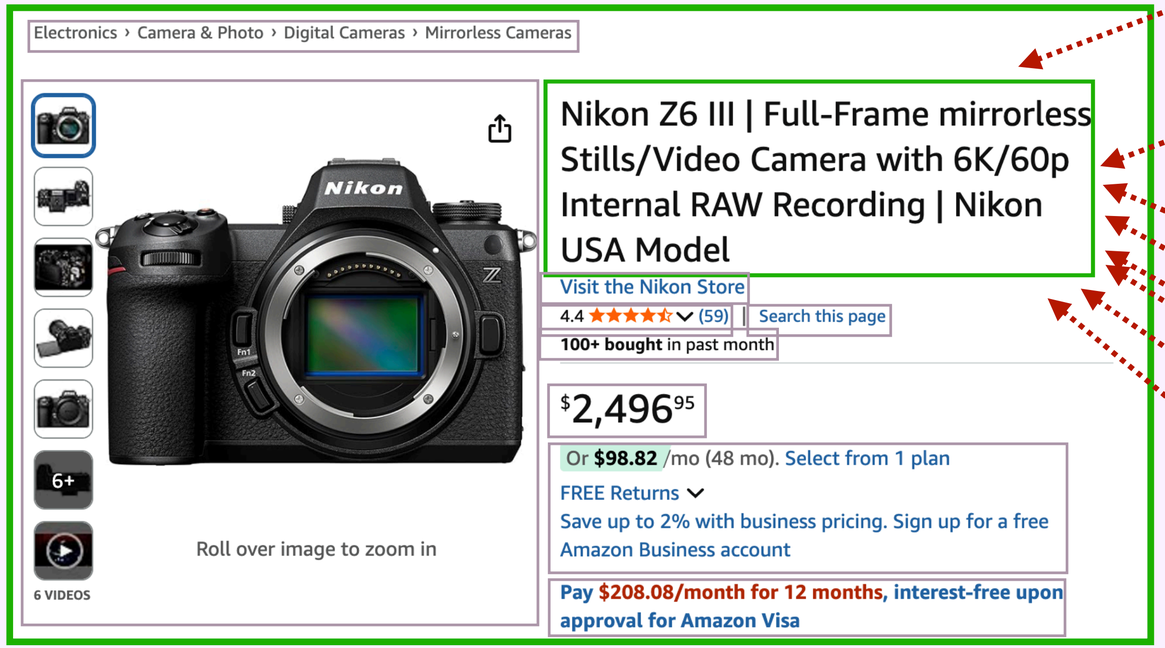
Unseen webpage



User-provided example



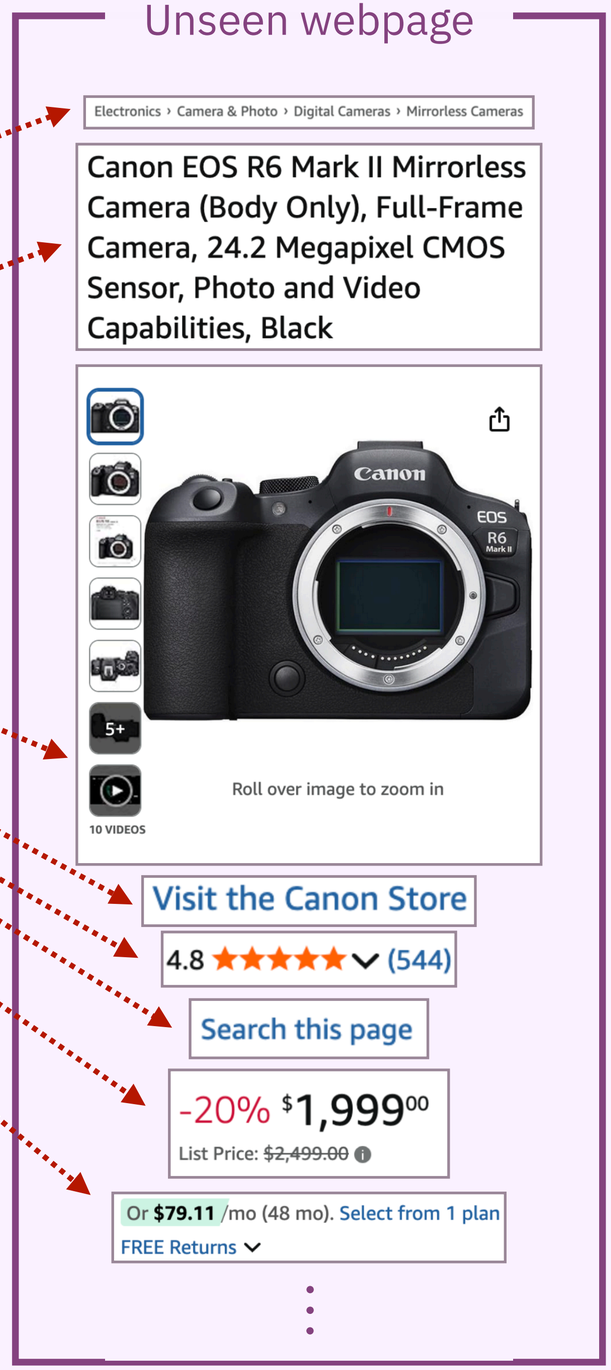
Ringer's scorers contend with webpage heterogeneity.



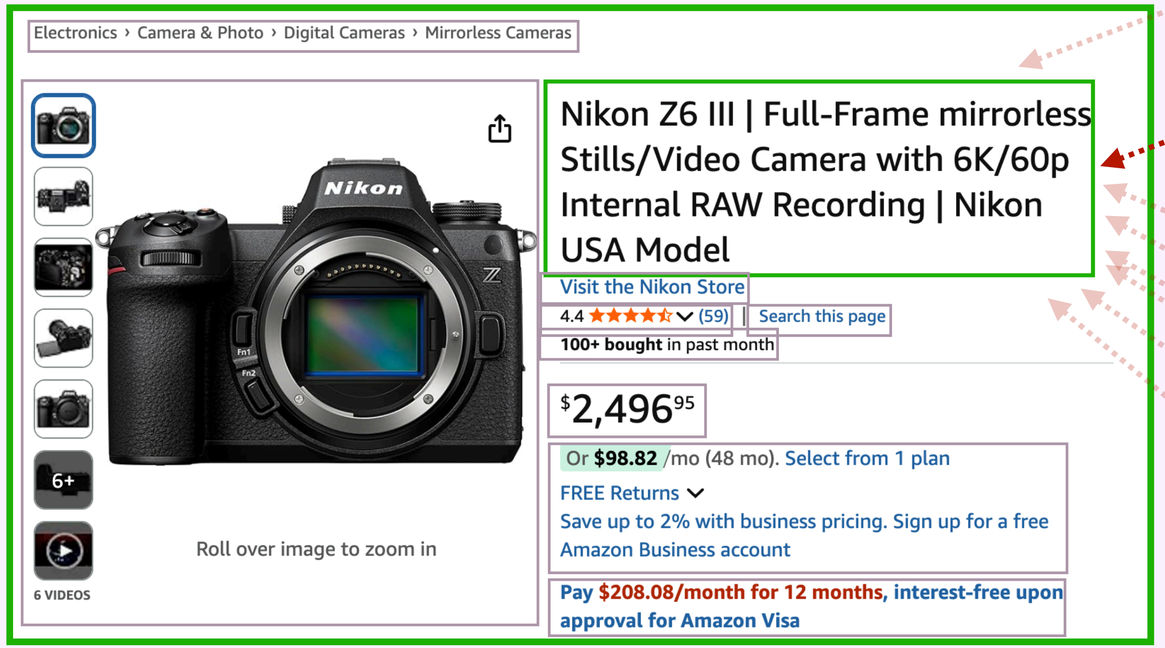
User-provided example

Attribute comparisons

Unseen webpage



Ringer's scorers contend with webpage heterogeneity.

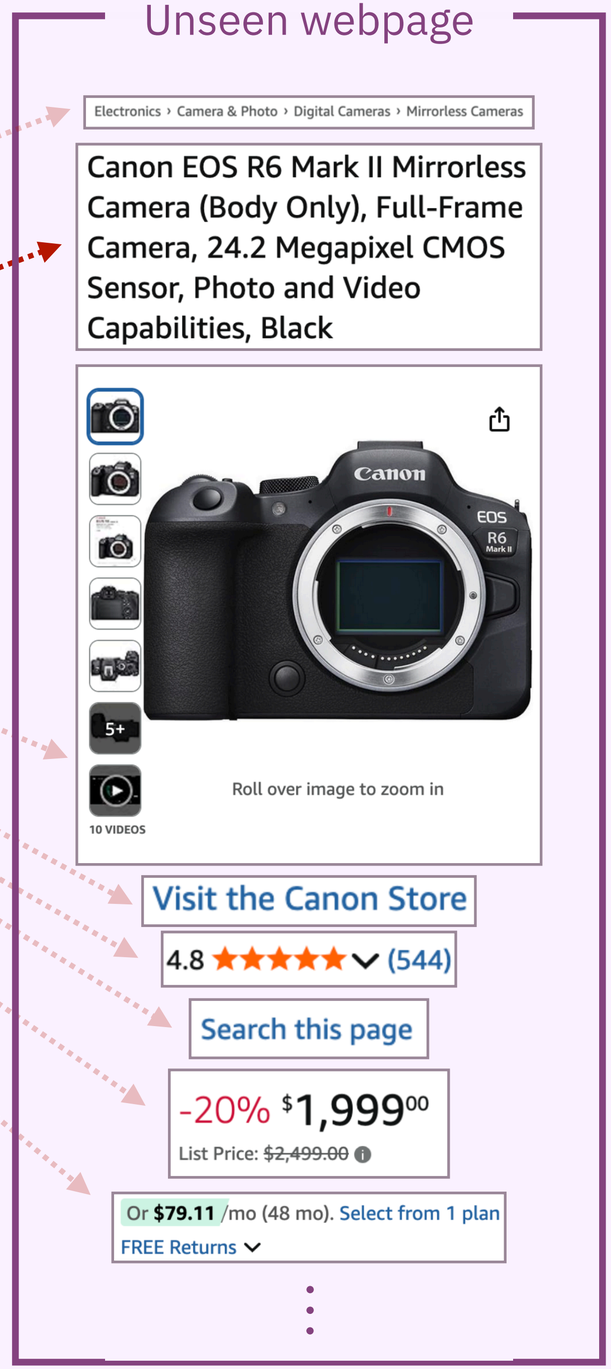


User-provided example

Same tag? ✗
Same ID? ✓
Same classes? ✓
Common attrs: 2

Attribute comparisons

Unseen webpage

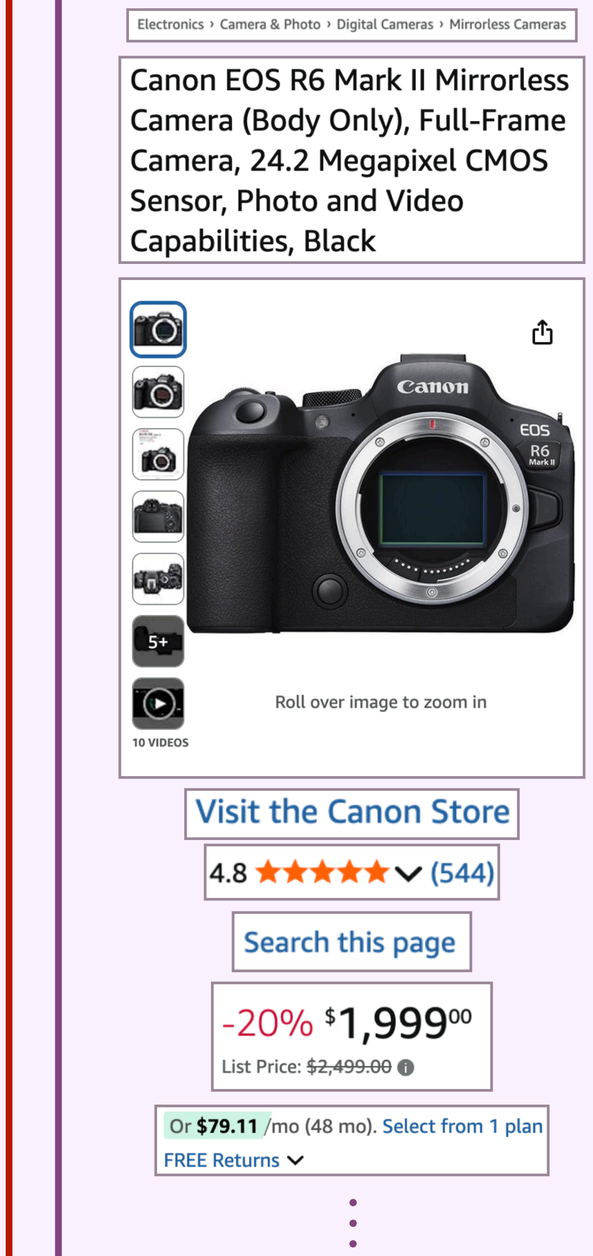


Ringer's scorers contend with webpage heterogeneity.

But they require structured documents.

Hundreds or thousands of nodes

Unseen webpage



Ringer's scorers require structured documents.

REMIT TO Sinclair Broadcast
dlb WXLV
PO Box 206270
Dallas, TX 75320-6270

Advertiser
Agency
Buyer
Salesperson

Citizens for Responsible Energy (84874)
Matson Media LLC (19039)
MediaAssistant,
Millennium DC, Washington DC (1108)
ph: (202) 955-5342, fx: (202) 955-5348x

Invoice 7245066
Inv Date 2/23/2020
Terms CIA
Contract 4248847
Bill Type Weekly Irregular
Period 2/17/2020 - 3/23/2020

Special Handling None
Product POLITICAL ISSUE (ns) (1187)
Brand NCHD29 (1340882)
Acct Types National Political Issue Agency BRD
Est/External # 1163 ECR26803967
Demo
Revision
Comments WXLV NCHD29

CG DR Order Type Nc Normal
Package
Gen. Date 2/25/2020 8:23:41 AM

Matson Media LLC
1201 Hampton Street
Ste 3B
Columbia, SC 29201

Greensboro (WXLV)

OFFICIAL BILLING INVOICE

Line	Type	Scheduled	Schedule	Days	to	Run	Air Time	Length	Program	Copy/ISCI	Amount	Remarks
1.0	Contract Line Remarks											
1.0	News	83209 ABC America					02/19/20 5:15AM (We)	00:30	ABC America This Morning	SEFTV021820H	$20.00	
1.0	News	83209 ABC America					02/19/20 5:59AM (We)	00:30	ABC America This Morning	SEFTV021820H	$20.00	
1.0	News	83209 ABC America					02/20/20 5:23AM (Th)	00:30	ABC America This Morning	SEFTV021820H	$20.00	
1.0	News	83209 ABC America					02/20/20 5:55AM (Th)	00:30	ABC America This Morning	SEFTV021820H	$20.00	
1.0	News	83209 ABC America					02/21/20 5:10AM (Fr)	00:30	ABC America This Morning	SEFTV021820H	$20.00	
1.0	News	83209 ABC America					02/21/20 5:53AM (Fr)	00:30	ABC America This Morning	SEFTV021820H	$20.00	
4.0	SPOT	20714 ABC Good Mor					02/19/20 7:55AM (We)	00:30	ABC Good Morning America	SEFTV021820H	$90.00	
4.0	SPOT	20716 ABC Good Mor					02/19/20 8:55AM (We)	00:30	ABC Good Morning America	SEFTV021820H	$90.00	
4.0	SPOT	20716 ABC Good Mor					02/20/20 7:55AM (Th)	00:30	ABC Good Morning America	SEFTV021820H	$90.00	
4.0	SPOT	20716 ABC Good Mor					02/20/20 8:39AM (Th)	00:30	ABC Good Morning America	SEFTV021820H	$90.00	
4.0	SPOT	20716 ABC Good Mor					02/21/20 7:25AM (Fr)	00:30	ABC Good Morning America	SEFTV021820H	$90.00	
4.0	SPOT	20716 ABC Good Mor					02/21/20 8:39AM (Fr)	00:30	ABC Good Morning America	SEFTV021820H	$90.00	
7.0	SPOT	20374 ABC Good Mor					02/22/20 8:28AM (Sa)	00:30	ABC Good Morning American Saturday	SEFTV021820H	$30.00	
9.0	SPOT	20374 ABC Good Mor					02/22/20 9:25AM (Sa)	00:30	ABC Good Morning American Saturday	SEFTV021820H	$30.00	
11.0	SPOT	14347 ABC Good Mo					02/23/20 8:25AM (Su)	00:30	ABC Good Morning America Sunday	SEFTV021820H	$40.00	

Net 30 Payment is due within 30 days from invoice date.
Station owned or provided services by the Sinclair Broadcast Group.

Warranty. I/We warrant the above broadcast event made according with the official station log. Please note that we do not notarize invoices since we feel a personal check or electronic acceptance document is adequate. The authenticity of an individual's signature (and not its clarity) is the accuracy of commercial credit

Gross Total $760.00
Commission ($114.00)
Net Total $646.00

Total Spots 15

Officially printed on 02/27/20 11:44 AM

Page 1

Hundreds or thousands of nodes

Unseen webpage

Electronics › Camera & Photo › Digital Cameras › Mirrorless Cameras

Canon EOS R6 Mark II Mirrorless Camera (Body Only), Full-Frame Camera, 24.2 Megapixel CMOS Sensor, Photo and Video Capabilities, Black

5+

10 VIDEOS

Roll over image to zoom in

Visit the Canon Store

4.8 ★★★★★ (544)

Search this page

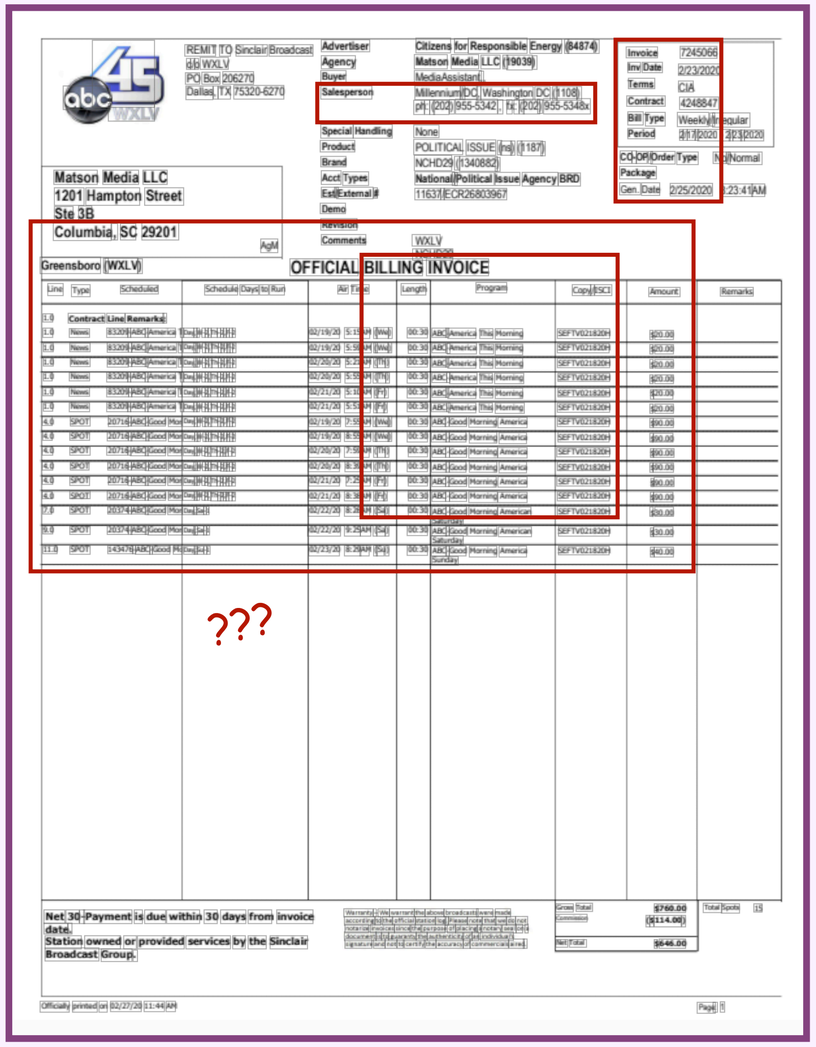
-20% $1,999.00

List Price: $2,499.00

Or $79.11/mo (48 mo). Select from 1 plan

FREE Returns

Ringer's scorers require structured documents.



Hundreds or thousands of nodes

Unseen webpage

Electronics › Camera & Photo › Digital Cameras › Mirrorless Cameras

Canon EOS R6 Mark II Mirrorless Camera (Body Only), Full-Frame Camera, 24.2 Megapixel CMOS Sensor, Photo and Video Capabilities, Black

5+

10 VIDEOS

Roll over image to zoom in

Visit the Canon Store

4.8 ★★★★★ (544)

Search this page

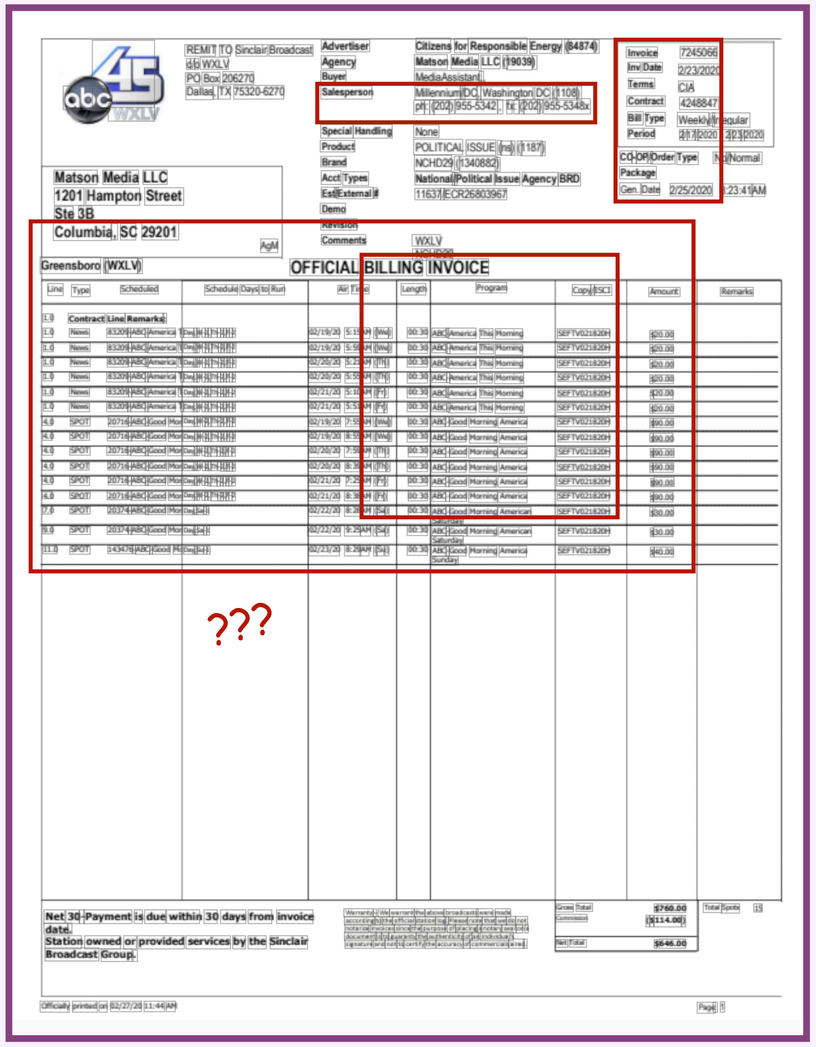
-20% $1,999.00

List Price: $2,499.00

Or $79.11 /mo (48 mo). Select from 1 plan

FREE Returns

Ringer's scorers require structured documents.



~17 billion OCR node subsets

Hundreds or thousands of nodes

Unseen webpage

Electronics › Camera & Photo › Digital Cameras › Mirrorless Cameras

Canon EOS R6 Mark II Mirrorless Camera (Body Only), Full-Frame Camera, 24.2 Megapixel CMOS Sensor, Photo and Video Capabilities, Black

5+

10 VIDEOS

Roll over image to zoom in

Visit the Canon Store

4.8 ★★★★★ (544)

Search this page

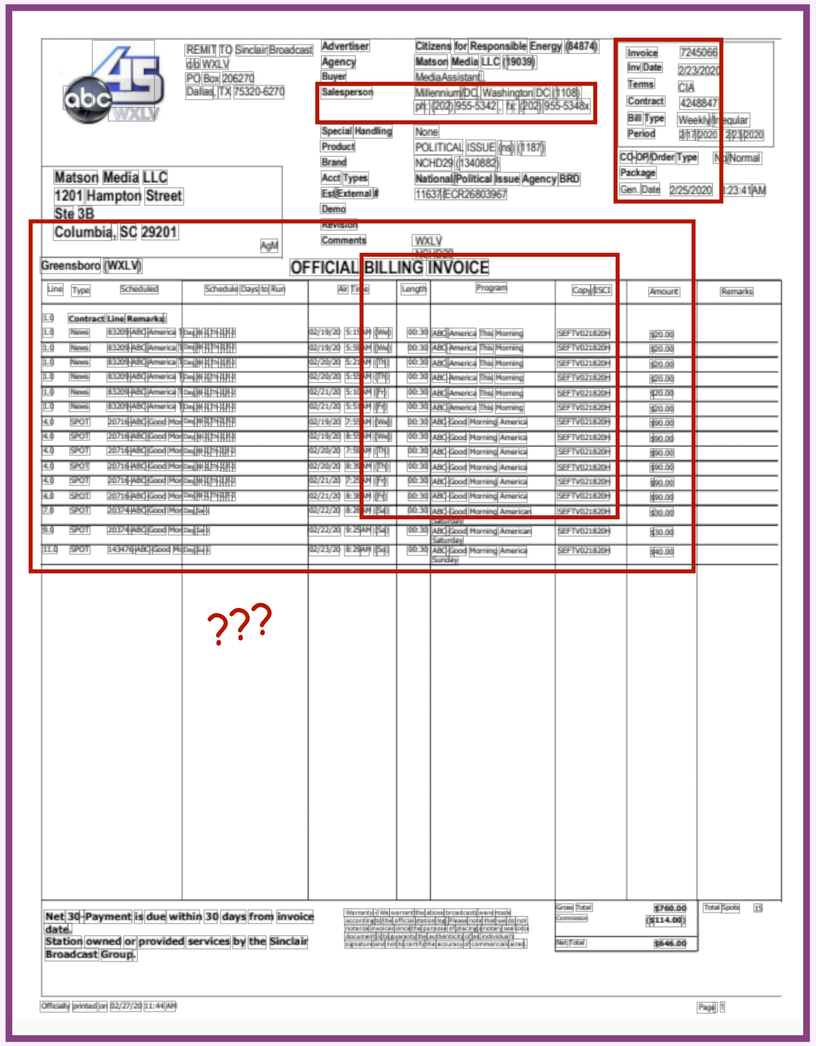
-20% $1,999.00

List Price: $2,499.00

Or $79.11 /mo (48 mo). Select from 1 plan

FREE Returns

Ringer's scorers require structured documents.



~17 billion OCR node subsets

Hundreds or thousands of nodes

Projected Ringer benchmark completion time:
the year 4105

Unseen webpage

Electronics › Camera & Photo › Digital Cameras › Mirrorless Cameras

Canon EOS R6 Mark II Mirrorless Camera (Body Only), Full-Frame Camera, 24.2 Megapixel CMOS Sensor, Photo and Video Capabilities, Black

Roll over image to zoom in

Visit the Canon Store

4.8 ★★★★★ (544)

Search this page

-20% $1,999.00

List Price: $2,499.00

Or $79.11/mo (48 mo). Select from 1 plan

FREE Returns

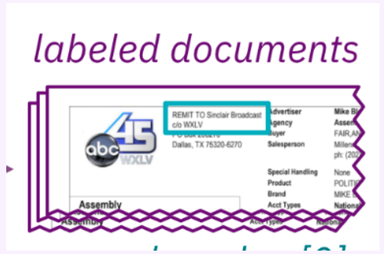
Ringer's scorers require structured documents.

Program sketches with scorer holes adapt traditional synthesis for messy data.

Our intervention

```
left = scorer (nodes)
right = scorer (nodes)
top = scorer (nodes)
bottom = scorer (nodes)

station = scorer (left, right, top, bottom)
```



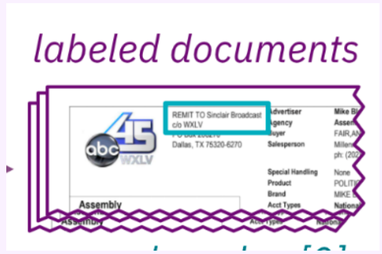
Ringer's scorers require structured documents.

Program sketches with scorer holes adapt traditional synthesis for messy data.

Our intervention

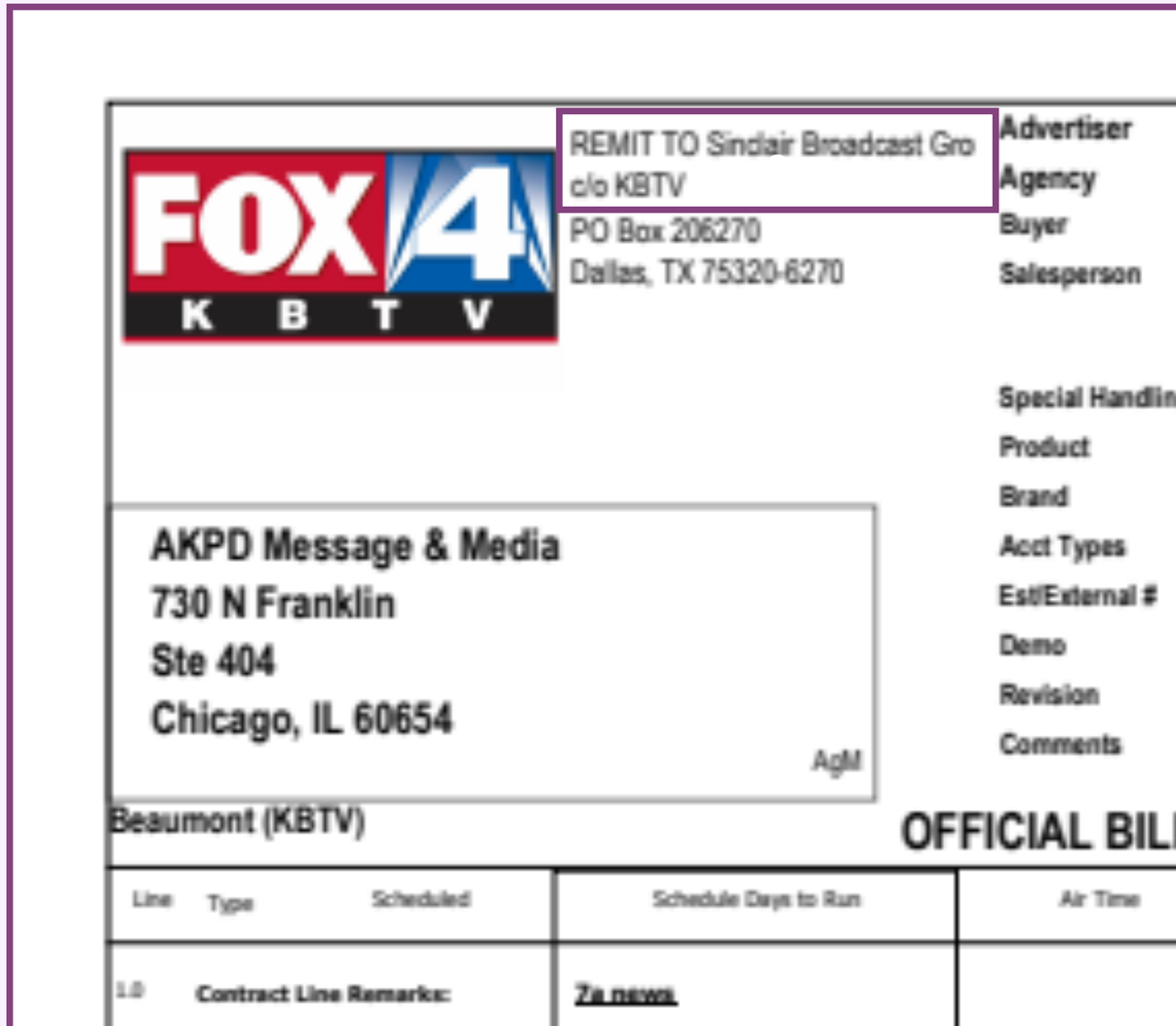
```
left = scorer(nodes)
right = scorer(nodes)
top = scorer(nodes)
bottom = scorer(nodes)

station = scorer(left, right, top, bottom)
```



Program sketches with scorer holes adapt traditional synthesis for messy data.

In particular, we can synthesize multiple scorers to identify information!



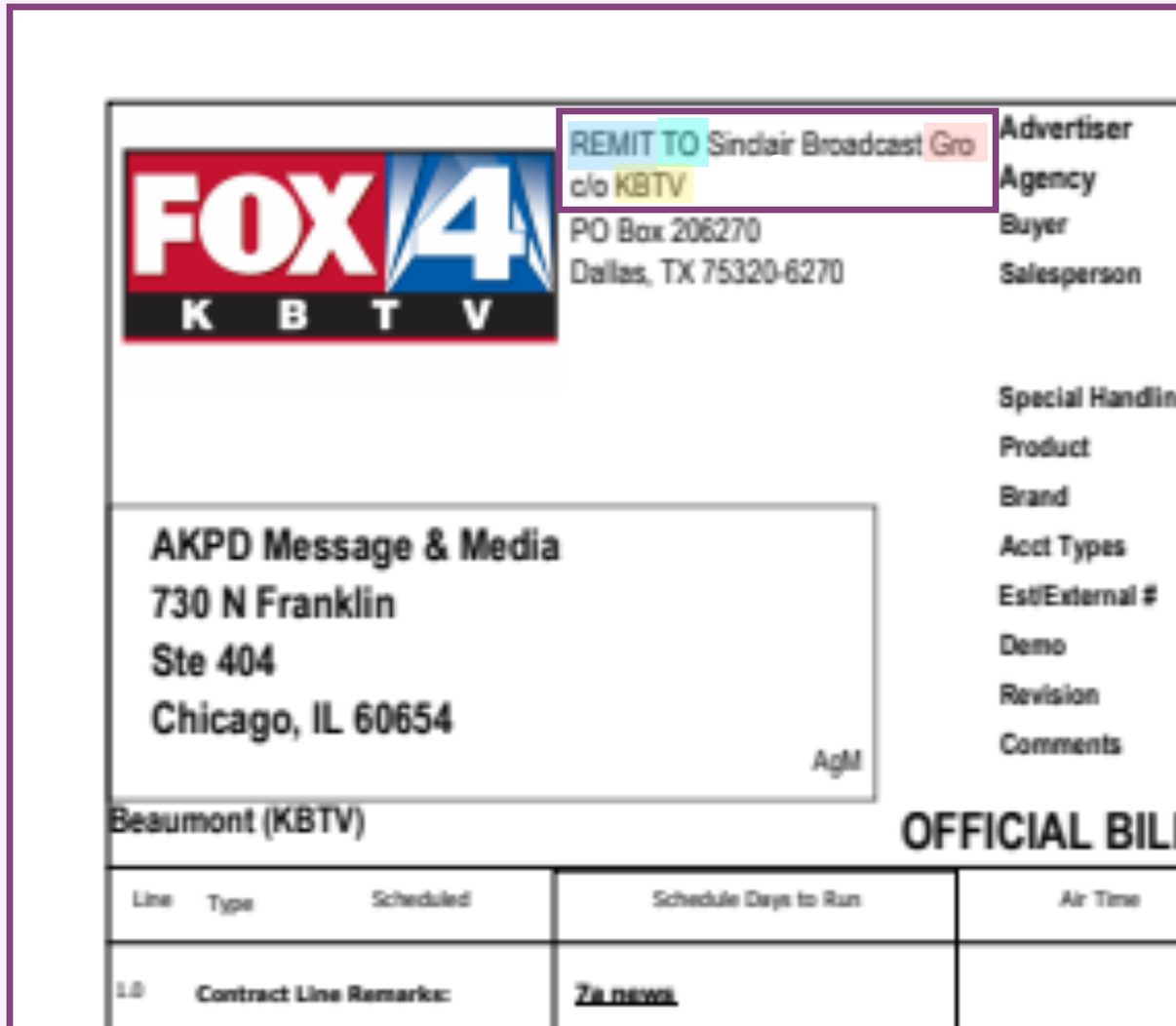
Program sketches with scorer holes adapt traditional synthesis for messy data.

In particular, we can synthesize multiple scorers to identify information!

```
left = scorer (nodes)
right = scorer (nodes)
top = scorer (nodes)
bottom = scorer (nodes)

station = scorer (left, right, top, bottom)
```

Synthesize scorers to find the edges of the field, then use those results to find the field itself.



Program sketches with scorer holes adapt traditional synthesis for messy data.

In particular, we can synthesize multiple scorers to identify information!

Program sketch with scorer holes

```
left = scorer (nodes)
right = scorer (nodes)
top = scorer (nodes)
bottom = scorer (nodes)

station = scorer (left, right, top, bottom)
```

Marshall
(our tool)

```
left = scorer_l(nodes)
right = scorer_r(nodes)
top = scorer_t(nodes)
bottom = scorer_b(nodes)

station = scorer_s(left, right, top, bottom)
```

Concrete scorer program

Program sketches with scorer holes adapt **sketch synthesis** for messy data.

In particular, we can synthesize multiple scorers to identify information!

Program sketch with scorer holes

```
left = scorer(nodes)
right = scorer(nodes)
top = scorer(nodes)
bottom = scorer(nodes)

station = scorer(left, right, top, bottom)
```

Marshall (our tool)

```
left = scorer_l(nodes)
right = scorer_r(nodes)
top = scorer_t(nodes)
bottom = scorer_b(nodes)

station = scorer_s(left, right, top, bottom)
```

Concrete scorer program

Program sketch

```
def two_times(num):
    result = num << ??
    return result
```

sketch synthesizer

```
def two_times(num):
    result = num << 1
    return result
```

Concrete program

Program sketches with scorer holes adapt traditional synthesis for messy data.

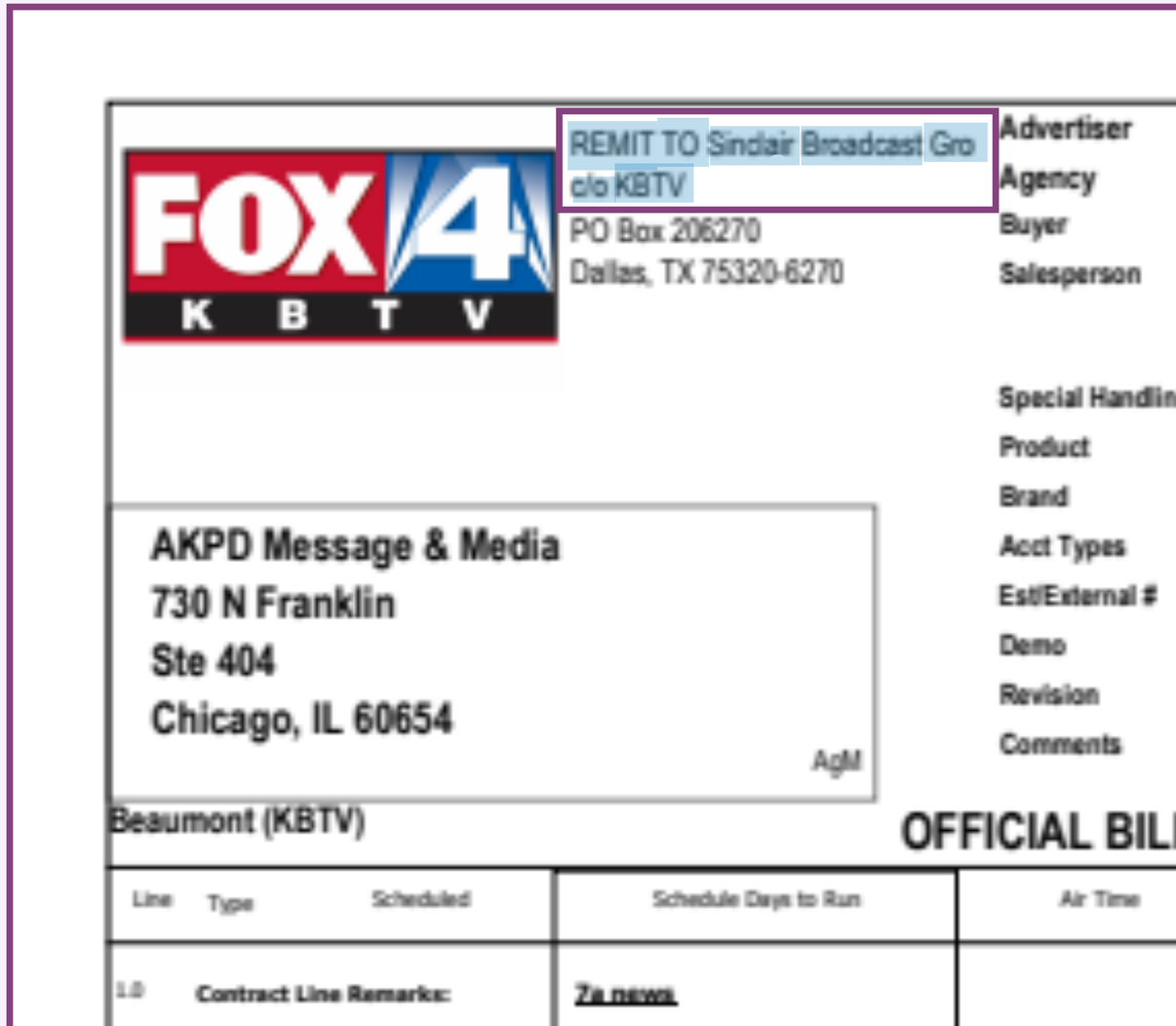
In particular, we can synthesize multiple scorers to identify information!

```
left = scorer(nodes)
right = scorer(nodes)
top = scorer(nodes)
bottom = scorer(nodes)

station = scorer(left, right, top, bottom)
```

```
inner = scorer(nodes)
station = scorer(inner)
```

Synthesize one scorer to rank OCR boxes by likelihood of field inclusion, then synthesize another scorer to find the field itself.



Program sketches with scorer holes adapt traditional synthesis for messy data.

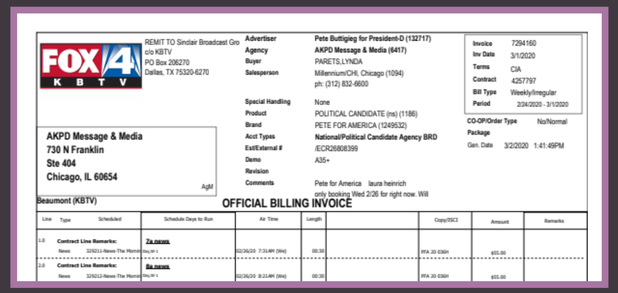
Different program sketches with scorer holes...

...yield different extraction tools for many domains.

```
left = scorer(nodes)
right = scorer(nodes)
top = scorer(nodes)
bottom = scorer(nodes)

station = scorer(left, right, top, bottom)
```

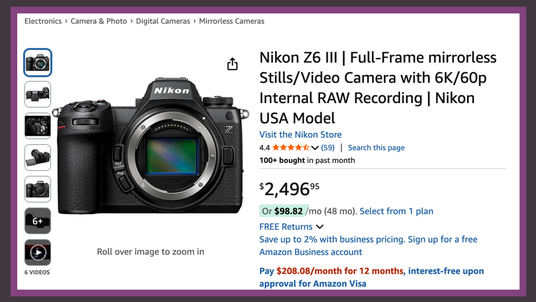
```
inner = scorer(nodes)
station = scorer(inner)
```



*Tools for **PDFs***

```
landmark = scorer(nodes)
target = scorer(landmark)
```

```
DOM node = scorer(nodes)
```

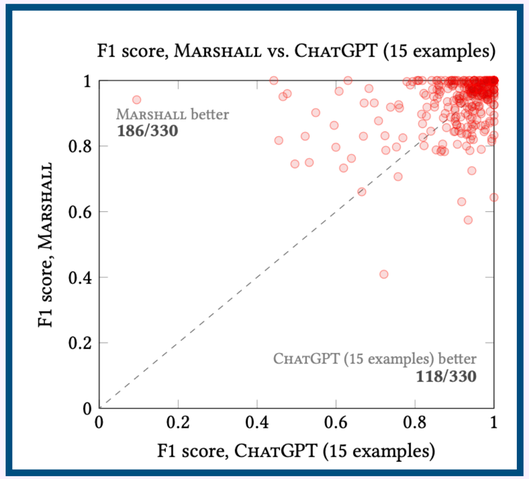


*Tools for **webpages***

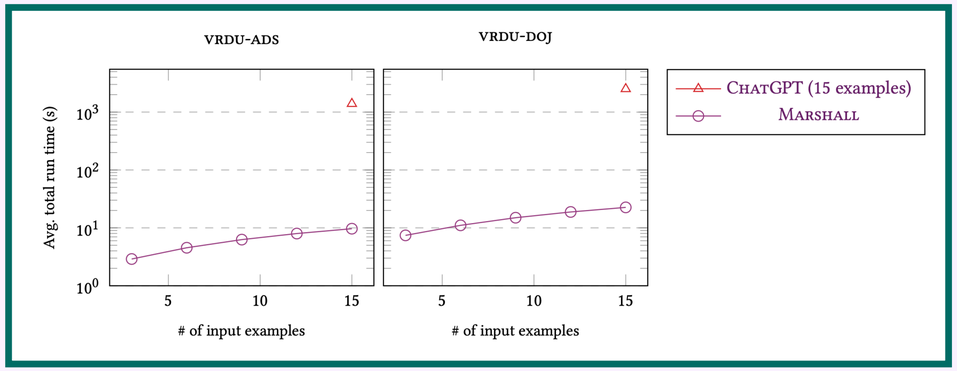
Does this work for actual datasets?

Does this work for actual datasets?

versus LLMs



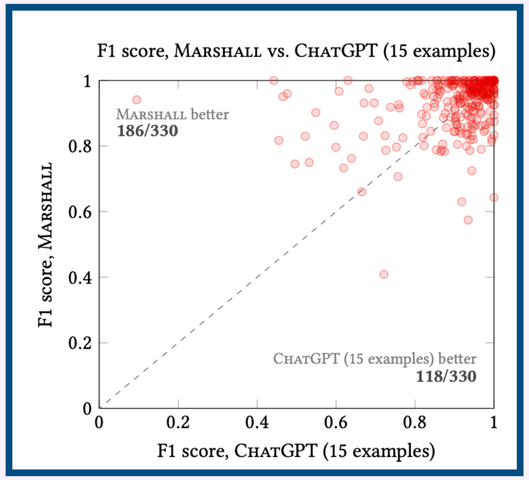
Our approach achieves precision and accuracy parity with GPT-4o...



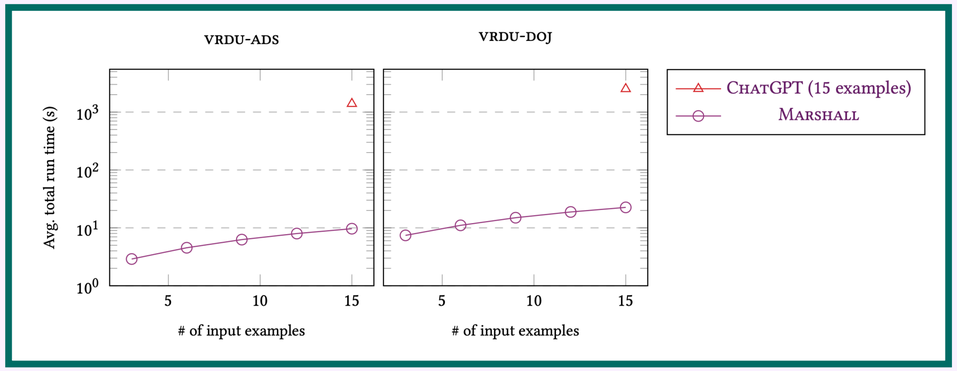
while running ≈120× faster and ≈95,700× cheaper.

Does this work for actual datasets?

versus LLMs



Our approach achieves precision and accuracy parity with GPT-4o...



while running ≈120× faster and ≈95,700× cheaper.

versus traditional synthesis

MARSHALL is 4,216,020.35× faster than RINGER.

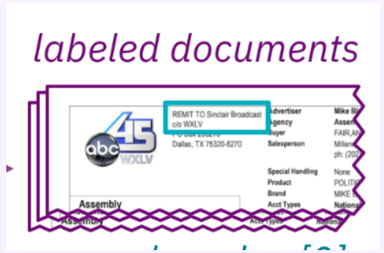
Program Sketches with Scorer Holes

an approach for retrieving info from heterogenous data by synthesizing multiple scorers.

Our intervention

```
left = scorer (nodes)
right = scorer (nodes)
top = scorer (nodes)
bottom = scorer (nodes)

station = scorer (left, right, top, bottom)
```



David Minh-Duy Cao, Sarah E. Chasins
dmcao@berkeley.edu