bigdocs.github.io

BigDocs

An Open Dataset for Training Multimodal Models on Document and Code Tasks

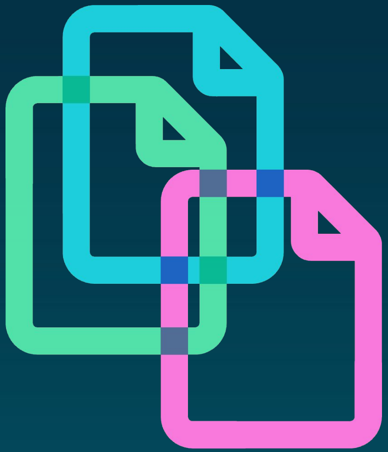
bigdocs.github.io

Juan A. Rodriguez, Xiangru Jian, et. al.
ServiceNow Research

juan.rodriguez@mila.quebec
xiangru.jian@waterloo.ca

servicenow Mila Université de Montréal ÉTS Le génie pour l'industrie McGill

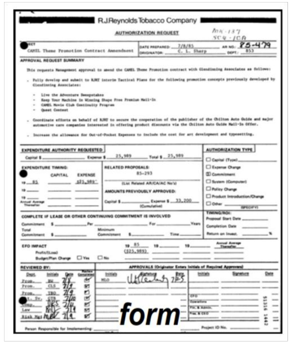
Multimodal Document Understanding



servicenow

Why Multimodal Document Understanding?

Existing demand of models to deal with:



[redacted] R.J.Reynolds Tobacco Company [redacted]

AUTHORIZATION REQUEST

MK-137
SCQ-ICA

SUBJECT: CAMEL Theme Promotion Contract Amendment

DATE PREPARED: 7/8/85 AR NO.: 85-479
ORIGINATOR: C. L. Sharp DEPT.: 853

APPROVAL REQUEST SUMMARY

This requests Management approval to amend the CAMEL Theme Promotion contract with Glendinning Associates as follows:

- Fully develop and submit to RJRT interim Tactical Plans for the following promotion concepts previously developed by Glendinning Associates:
  - Live the Adventure Sweepstakes
  - Keep Your Machine in Winning Shape Free Premium Mail-In
  - CAMEL Movie Club Continuity Program
  - Quest Contest
- Coordinate efforts on behalf of RJRT to secure the cooperation of the publisher of the Chilton Auto Guide and major automotive care companies interested in offering product discounts via the Chilton Auto Guide Mail-In Offer.
- Increase the allowance for Out-of-Pocket Expenses to include the cost for art development and typesetting.

EXPENDITURE AUTHORITY REQUESTED

Capital $ ______ Expense $ 25,989 Total $ 25,989

AUTHORIZATION TYPE

□ Capital (Type) ______
□ Expense Change
☒ Commitment
□ System (Computer)
□ Policy Change
□ Product Introduction/Change
□ Other ______ (SPECIFY)

EXPENDITURE TIMING:

	CAPITAL	EXPENSE
19 85		$25,989
19		
19		
Annual Average Thereafter		

RELATED PROPOSALS: 85-293
(List Related AR/CA/AC No's)

AMOUNTS PREVIOUSLY APPROVED:
Capital $ ______ Expense $ 33,200
(Cumulative)

COMPLETE IF LEASE OR OTHER CONTINUING COMMITMENT IS INVOLVED

Commitment $ ______ Per ______ For ______ Years
Total Commitment $ ______ Minimum Commitment $ ______ Time ______

TIMING/ROI:
Proposal Start Date ______
Completion Date ______
Return on Invest. ______ %

EFO IMPACT

	19 85	19	19	Annual Average Thereafter
Profit/(Loss)	($25,989)			
Budget/Plan Change □ Yes □ No				

REVIEWED BY:

Dept.	Initials	Date	Review Completed
Prom.	BW	7/8	✓
Prom.	CLS	7/9	✓
Prom.	TBO	7/9	✓
[illegible]	GTS	7/10	✓
Comp.	DPS	7/11	✓
Law	[illegible]	7/11	✓
Risk Mgt	MBV	7/9	✓

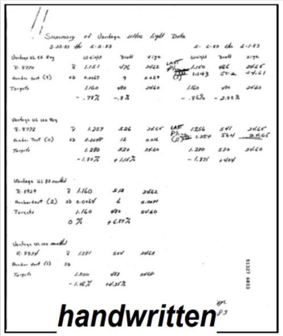
APPROVALS (Originator Enters Initials of Required Approvers)

Initials	Signature	Date	Initials	Signature	Date
MLO	[illegible]	7/25			
			EFO		
			Operations		
			Fin. & Admin.		
			Pres. & CEO		

Person Responsible for Implementing: ______ Project ID No. ______

51316 1442

form



Summary of Vantage ultra light Data
3-25-83 thr 6-5-83 6-6-83 thr 6-1-83

Vantage UL 85 Reg — Weight Draft Size
R-8770 x̄ 1.151 [illegible] [illegible]
Number test (9) SD 0.0067 9 0.029
Targets 1.160 [illegible] [illegible]
-.78% -.8%
LAST Ph III (7) 1.150 [illegible] [illegible] / 1.143 [illegible] [illegible]
1.160 [illegible] [illegible]
-.86% -2.92%

Vantage UL 100 Reg
R-8772 x̄ 1.257 526 [illegible]
Number Test (n) SD [illegible] 12 [illegible]
Targets 1.280 520 [illegible]
-1.80% +1.15%
LAST Ph (2) 1.256 541 [illegible] / 1.254 544 [illegible]
1.280 520 [illegible]
-1.875 [illegible]

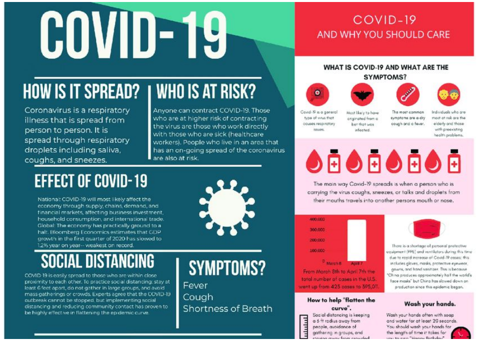
Vantage UL 85 menthol
R-8929 x̄ 1.160 [illegible] [illegible]
Number test (2) SD [illegible] 6 [illegible]
Targets 1.160 [illegible] [illegible]
0% +6.88%

Vantage UL 100 menthol
R-8934 x̄ 1.281 [illegible] [illegible]
Number test (1) SD
Targets 1.300 [illegible] [illegible]
-1.46% [illegible]

51327 6933

handwritten

Complex Paper Documents



COVID-19

HOW IS IT SPREAD?

Coronavirus is a respiratory illness that is spread from person to person. It is spread through respiratory droplets including saliva, coughs, and sneezes.

WHO IS AT RISK?

Anyone can contract COVID-19. Those who are at higher risk of contracting the virus are those who work directly with those who are sick (healthcare workers). People who live in an area that has an on-going spread of the coronavirus are also at risk.

EFFECT OF COVID-19

National: COVID-19 will most likely affect the economy through supply, chains, demand, and financial markets, affecting business investment, household consumption, and international trade. Global: The economy has practically ground to a halt. Bloomberg Economics estimates that GDP growth in the first quarter of 2020 has slowed to 1.2% year on year-- weakest on record.

SOCIAL DISTANCING

COVID-19 is easily spread to those who are within close proximity to each other. To practice social distancing: stay at least 6 feet apart, do not gather in large groups, and avoid mass gatherings or crowds. Experts agree that the COVID-19 outbreak cannot be stopped, but implementing social distancing and reducing community contact has proven to be highly effective in flattening the epidemic curve.

SYMPTOMS?

Fever
Cough
Shortness of Breath

COVID-19 AND WHY YOU SHOULD CARE

WHAT IS COVID-19 AND WHAT ARE THE SYMPTOMS?

Covid-19 is a general type of virus that causes respiratory issues.

Most likely to have originated from a bat that was infected.

The most common symptoms are a dry cough and a fever.

Individuals who are most at risk are the elderly and those with preexisting health problems.

The main way Covid-19 spreads is when a person who is carrying the virus coughs, sneezes, or talks and droplets from their mouths travels into another persons mouth or nose.

From March 8th to April 7th the total number of cases in the U.S. went up from 423 cases to 395,011.

There is a shortage of personal protective equipment (PPE) and ventilators during this time due to rapid increase of Covid-19 cases; this includes gloves, masks, protective eyewear, gowns, and hand sanitizer. This is because "China produces approximately half the world's face masks" but China has slowed down on production since this epidemic began.

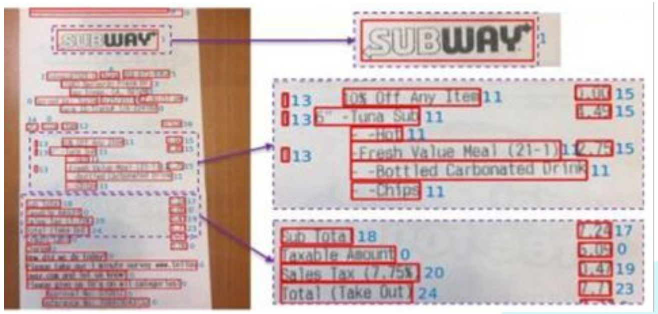
How to help "flatten the curve".

Social distancing is keeping a 6 ft radius away from people, avoidance of gathering in groups, and staying away from crowded

Wash your hands.

Wash your hands often with soap and water for at least 20 seconds. You should wash your hands for the length of time it takes for you to sing "Happy Birthday"

Text-intensive Infographics



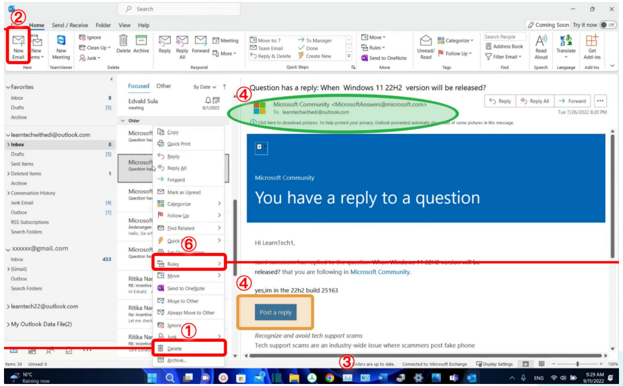
SUBWAY 1
13 10% Off Any Item 11 1.10 15
13 6" -Tuna Sub 11 5.49 15
-Hot 11
13 -Fresh Value Meal (21-1) 11 2.75 15
-Bottled Carbonated Drink 11
-Chips 11
Sub Total 18 7.24 17
Taxable Amount 0 5.06 0
Sales Tax (7.75%) 20 0.47 19
Total (Take Out) 24 7.71 23

Robust Perception & Grounding

Why Multimodal Document Understanding?

New tasks for Document AI

Multimodal Code Generation



Question has a reply: When Windows 11 22H2 version will be released?

Microsoft Community <MicrosoftAnswers@microsoft.com>

You have a reply to a question

Hi LearnTech1,

yes,im in the 22h2 build 25163

Recognize and avoid tech support scams

Tech support scams are an industry-wide issue where scammers post fake phone

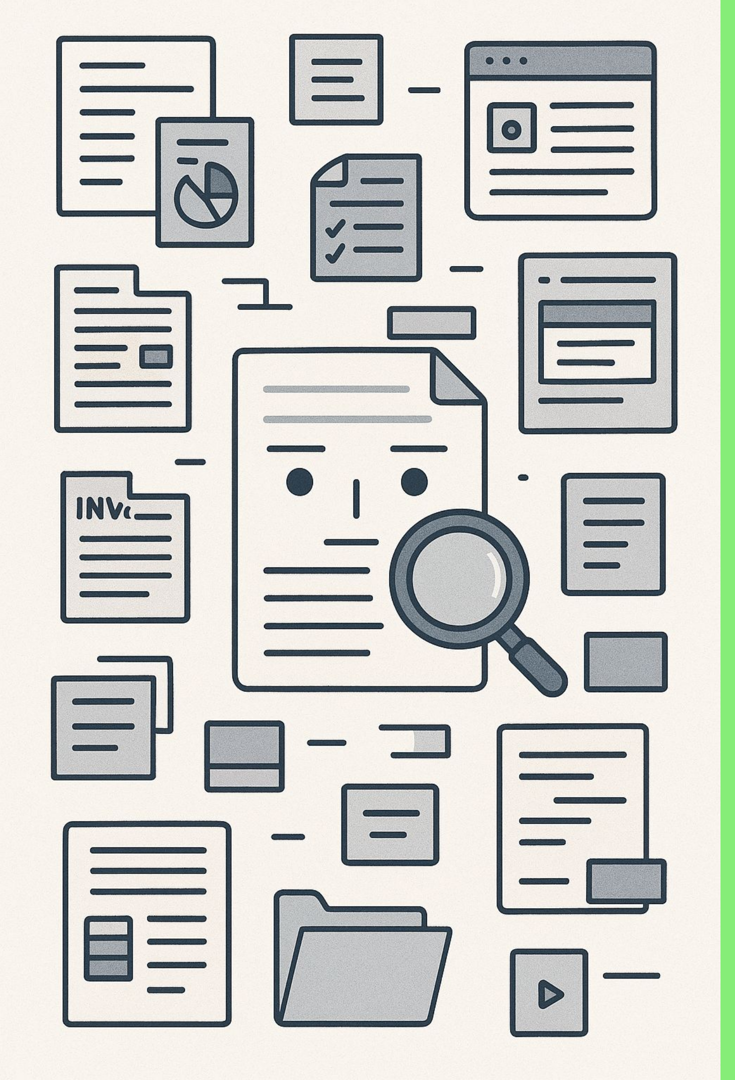
User Interface Understanding

Current Limitations:

1. **Models** need ***huge*** *and* ***diverse*** *datasets* to generalize well.

2. **Data is limited**, scattered, poorly licensed, or simplistic.

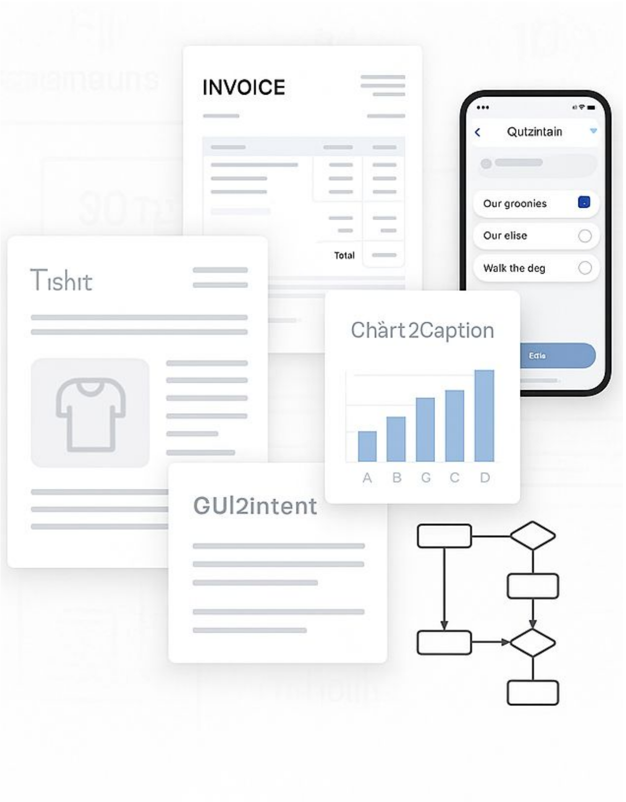
3. **Modern/real world tasks** are underexplored



> “
>
> *We need **Scalable**, **Open**, and **High-Quality** Multimodal Document Datasets*

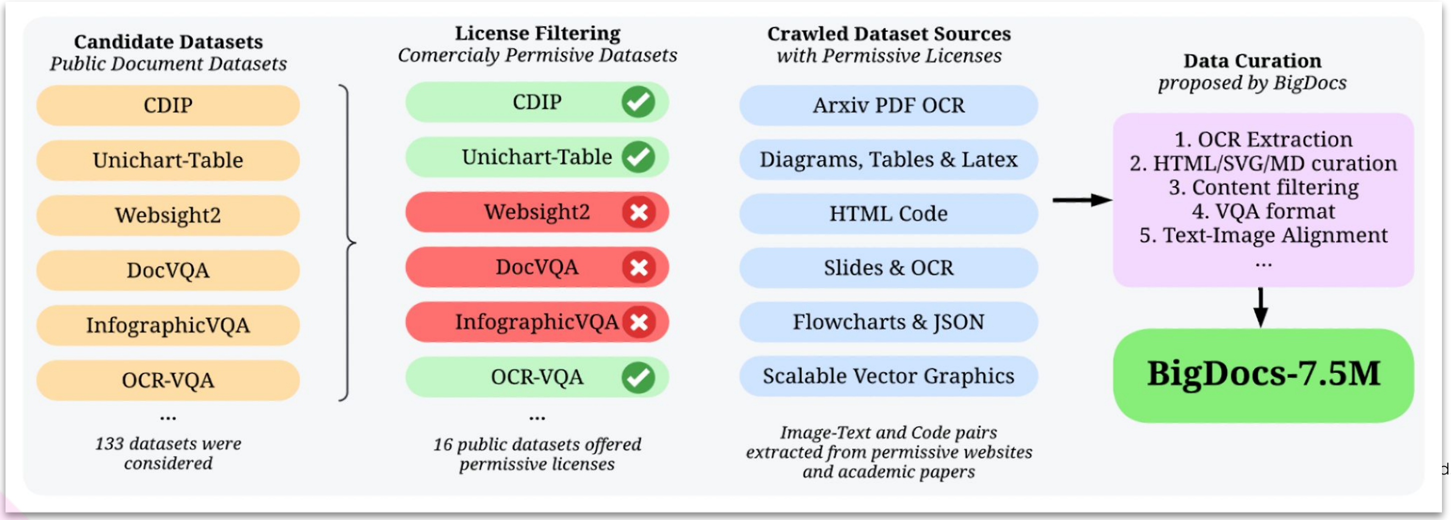
BigDocs Contributions Summary

BigDocs 7.5M	Large-scale pre-training dataset with clearly licensing and transparent
BigDocs-Bench	10 innovative tasks for multimodal code & GUI Understanding
BigDocs-Toolkit	Crawling, dataset compound management, safety filtering, document AI training
BigDocs-Models	License-permissive multimodal document models Phi-3 QwenVL2 LLaVA



Dataset Construction

- Merged and **filtered 130+ datasets → kept only 16** with **permissive licenses**
- Added **new data:** UIs, HTML screenshots, LaTeX tables, …
- **Strict content**, **license**, and **quality checks** throughout



Candidate Datasets
Public Document Datasets

- CDIP
- Unichart-Table
- Websight2
- DocVQA
- InfographicVQA
- OCR-VQA
- …

133 datasets were considered

License Filtering
Comercialy Permisive Datasets

- CDIP ✓
- Unichart-Table ✓
- Websight2 ✗
- DocVQA ✗
- InfographicVQA ✗
- OCR-VQA ✓
- …

16 public datasets offered permissive licenses

Crawled Dataset Sources
with Permissive Licenses

- Arxiv PDF OCR
- Diagrams, Tables & Latex
- HTML Code
- Slides & OCR
- Flowcharts & JSON
- Scalable Vector Graphics

Image-Text and Code pairs extracted from permissive websites and academic papers

Data Curation
proposed by BigDocs

1. OCR Extraction
2. HTML/SVG/MD curation
3. Content filtering
4. VQA format
5. Text-Image Alignment

…

BigDocs-7.5M

What’s in BigDocs-7.5M?

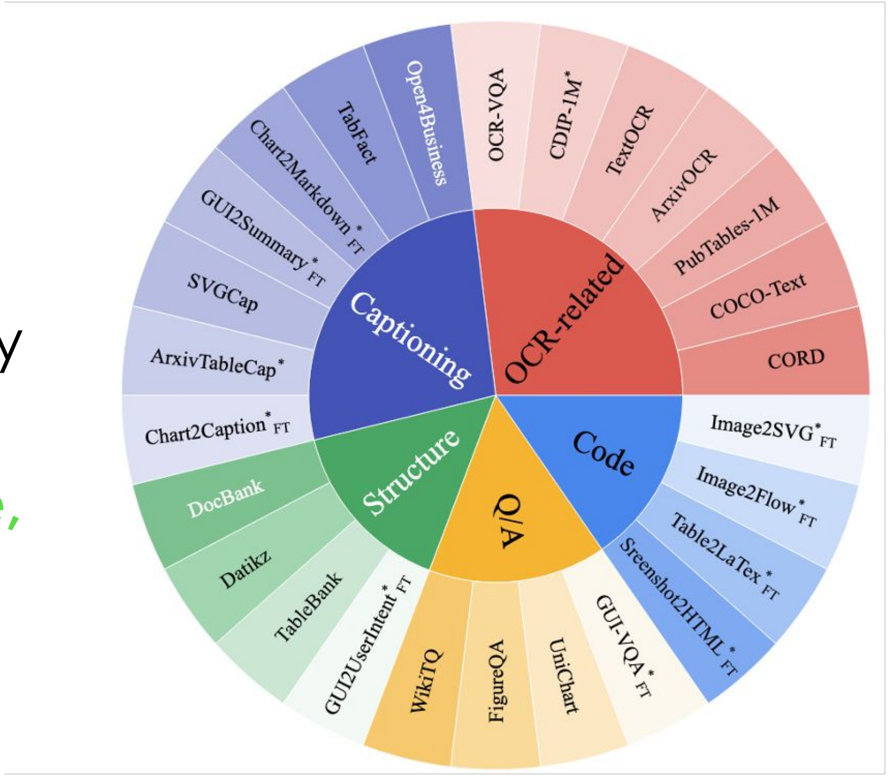
OCR, Layout, Tables, Forms, Charts, UIs, Scientific Diagrams

Text, Images, Structured Outputs (HTML, LaTeX, JSON, SVG)

Balanced across input-output modalities: vision → code, vision → text

What’s in BigDocs-7.5M?

We **curated multimodal datasets** in an **open** and **transparent** way specifically **for training multimodal models document, website, coding domains**



“

We created ***new Benchmarks*** *for* ***Modern Document AI***

Introducing **BigDocs-Bench**

- **Proposing 10 new tasks** for challenging real-world use cases
- Tests **visual *reasoning***, **layout understanding**, ***structured code generation***, ***UI comprehension***

Multimodal Code Generation

Image2HTML, Image2LaTeX, Image2SVG, Flow2Code

User Interface Reasoning

GUI2Intent, GUI2Summary, GUI-VQA

Chart Understanding

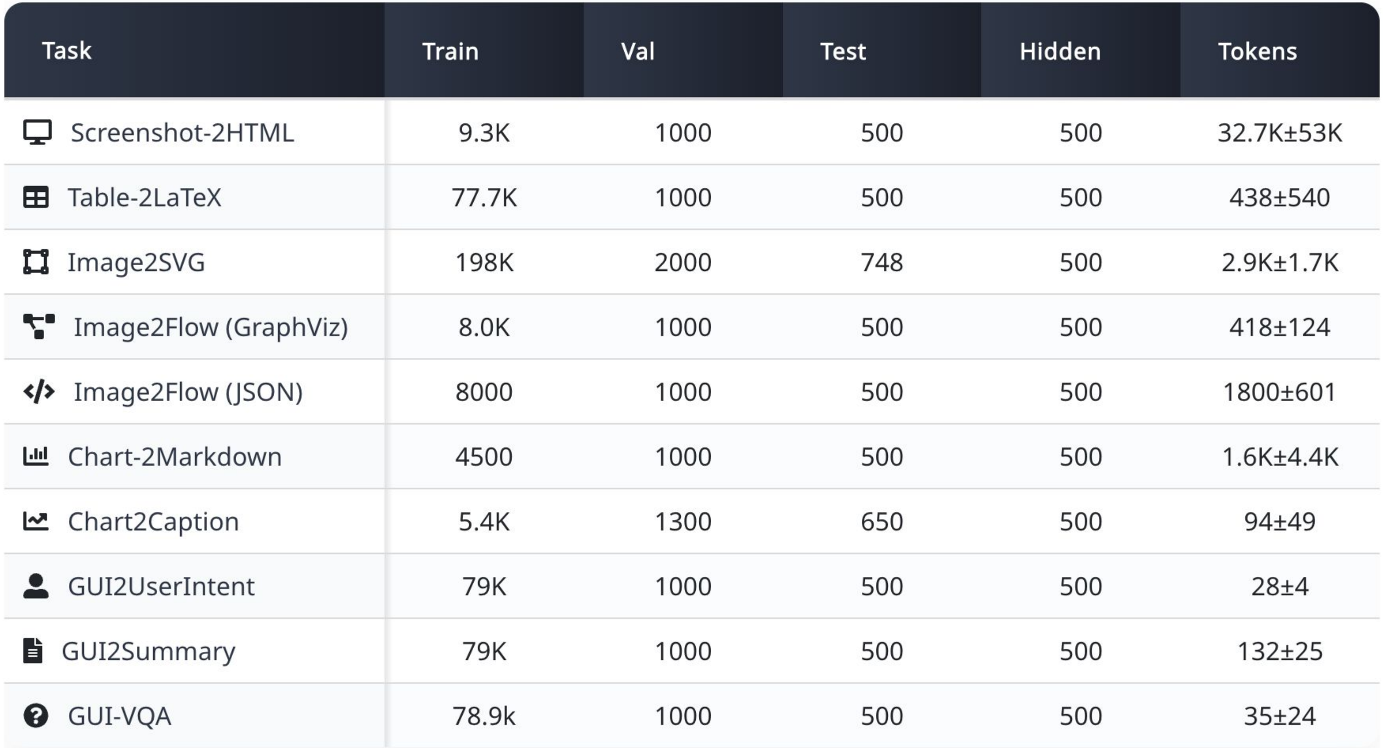
Chart2Caption

“

It’s not just OCR anymore.

Models must generate complex outputs from visuals.

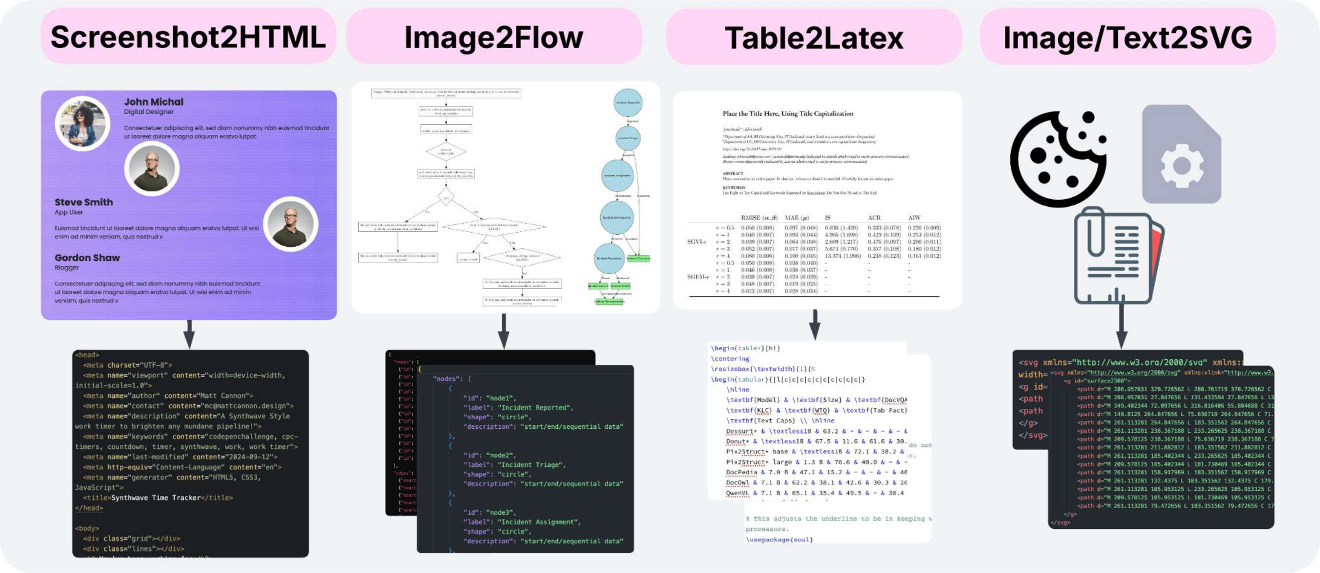
Novel Tasks in BigDocs-Bench



Task	Train	Val	Test	Hidden	Tokens
Screenshot-2HTML	9.3K	1000	500	500	32.7K±53K
Table-2LaTeX	77.7K	1000	500	500	438±540
Image2SVG	198K	2000	748	500	2.9K±1.7K
Image2Flow (GraphViz)	8.0K	1000	500	500	418±124
Image2Flow (JSON)	8000	1000	500	500	1800±601
Chart-2Markdown	4500	1000	500	500	1.6K±4.4K
Chart2Caption	5.4K	1300	650	500	94±49
GUI2UserIntent	79K	1000	500	500	28±4
GUI2Summary	79K	1000	500	500	132±25
GUI-VQA	78.9k	1000	500	500	35±24

Novel Tasks: Image2Code

- Input: Screenshot of web page, table, workflow, logotype
- Output: HTML / LaTeX / SVG
- Requires: understanding ***layout***, ***content***, and **generating *structured* code**



Screenshot2HTML

John Michal
Digital Designer

Consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam eratvo lutpat.

Steve Smith
App User

Euismod tincidunt ut laoreet dolore magna aliquam eratvo lutpat. Ut wisi enim ad minim veniam, quis nostrud v

Gordon Shaw
Blogger

Consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam eratvo lutpat. Ut wisi enim ad minim veniam, quis nostrud v

```html
<head>
  <meta charset="UTF-8">
  <meta name="viewport" content="width=device-width,
initial-scale=1.0">
  <meta name="author" content="Matt Cannon">
  <meta name="contact" content="mc@mattcannon.design">
  <meta name="description" content="A Synthwave Style
work timer to brighten any mundane pipeline!">
  <meta name="keywords" content="codepenchallenge, cpc-
timers, countdown, timer, synthwave, work, work timer">
  <meta name="last-modified" content="2024-09-12">
  <meta http-equiv="Content-Language" content="en">
  <meta name="generator" content="HTML5, CSS3,
JavaScript">
  <title>Synthwave Time Tracker</title>
</head>

<body>
  <div class="grid"></div>
  <div class="lines"></div>
```

Image2Flow

```json
{
  "nodes": [
    {"id":
    {"id":
    {"id":
    {"id":
    {"id":
    {"id":
    {"id":
    {"id":
    {"id":
    {"id":
    {"id":
    {"id":
    {"id":
  ],
  "edges":
    {"sourc
    {"sourc
    {"sourc
    {"sourc
    {"sourc
```

```json
{
    "nodes": [
        {
            "id": "node1",
            "label": "Incident Reported",
            "shape": "circle",
            "description": "start/end/sequential data"
        },
        {
            "id": "node2",
            "label": "Incident Triage",
            "shape": "circle",
            "description": "start/end/sequential data"
        },
        {
            "id": "node3",
            "label": "Incident Assignment",
            "shape": "circle",
            "description": "start/end/sequential data"
        },
```

Table2Latex

Place the Title Here, Using Title Capitalization

[illegible]

ABSTRACT
Please summarize an entire paper. In abstract, references should be avoided. Carefully format an entire paper.

KEYWORDS
List Eight to Ten Capitalized Keywords Separated by Semicolons; Do Not Use Period at The End

		RMISE (α, β)	MAE (μ)	IS	ACR	AIW
SGVI-c	r = 0.5	0.050 (0.008)	0.097 (0.048)	6.030 (1.420)	0.333 (0.078)	0.220 (0.009)
	r = 1	0.040 (0.007)	0.093 (0.044)	4.905 (1.698)	0.429 (0.139)	0.213 (0.012)
	r = 2	0.039 (0.007)	0.064 (0.038)	2.609 (1.217)	0.476 (0.097)	0.200 (0.011)
	r = 3	0.052 (0.007)	0.077 (0.037)	5.674 (0.770)	0.357 (0.108)	0.180 (0.012)
	r = 4	0.080 (0.006)	0.100 (0.045)	13.374 (1.996)	0.238 (0.123)	0.161 (0.012)
SGEM-c	r = 0.5	0.050 (0.008)	0.038 (0.040)	-	-	-
	r = 1	0.046 (0.008)	0.038 (0.037)	-	-	-
	r = 2	0.039 (0.007)	0.024 (0.029)	-	-	-
	r = 3	0.048 (0.007)	0.019 (0.025)	-	-	-
	r = 4	0.073 (0.007)	0.028 (0.034)	-	-	-

```latex
\begin{table*}[h!]
\centering
\resizebox{\textwidth}{!}{%
\begin{tabular}{|l|c|c|c|c|c|c|c|c|c|c|}
    \hline
    \textbf{Model} & \textbf{Size} & \textbf{DocVQA
    \textbf{KLC} & \textbf{WTQ} & \textbf{Tab Fact}
    \textbf{Text Caps} \\ \hline
    Dessurt* & \textless1B & 63.2 & - & - & - & - &
    Donut* & \textless1B & 67.5 & 11.6 & 61.6 & 30.
    Pix2Struct* base & \textless1B & 72.1 & 38.2 &
    Pix2Struct* large & 1.3 B & 76.6 & 40.0 & - & -
    DocPedia & 7.0 B & 47.1 & 15.2 & - & - & - & 46
    DocOwl & 7.1 B & 62.2 & 38.1 & 42.6 & 30.3 & 26
    QwenVL & 7.1 B & 65.1 & 35.4 & 49.5 & - & 30.4
```

```latex
do not
it.

% This adjusts the underline to be in keeping w
processors.
\usepackage{soul}
```

Image/Text2SVG

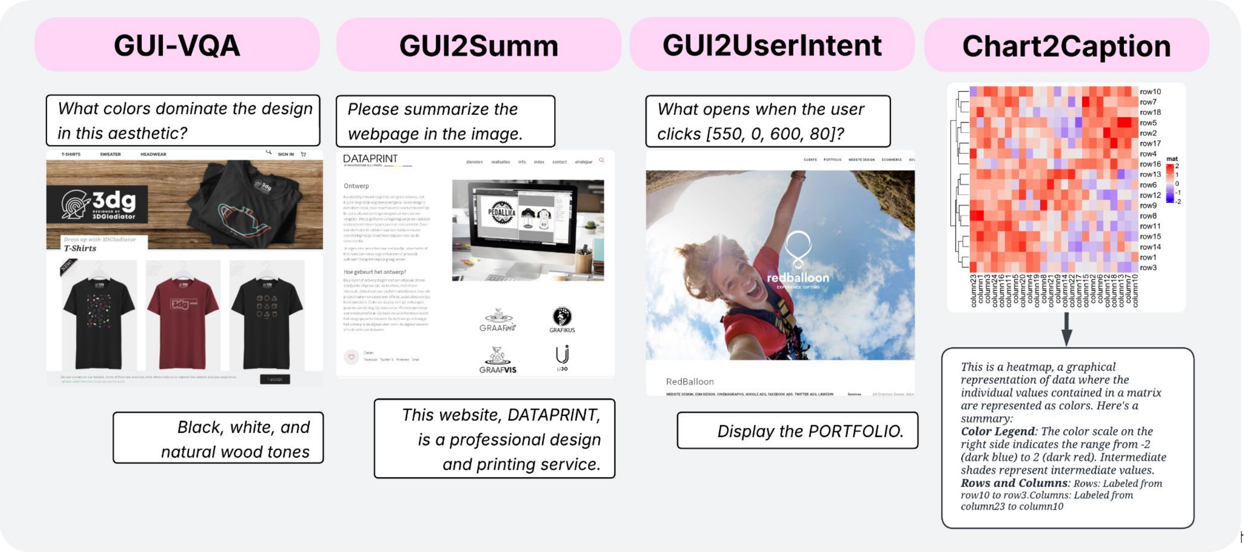
```xml
<svg xmlns="http://www.w3.org/2000/svg" xmlns:
width=
<g id=
<path
<path
</g>
</svg>
```

```xml
<svg xmlns="http://www.w3.org/2000/svg" xmlns:xlink="http://www.w3.
    <g id="surface2308">
        <path d="M 286.957031 370.726562 L 280.761719 370.726562 C
        <path d="M 286.957031 27.847656 L 131.433594 27.847656 L 13
        <path d="M 349.402344 72.097656 L 316.816406 55.804688 C 31
        <path d="M 149.8125 264.847656 L 75.636719 264.847656 C 71.
        <path d="M 261.113281 264.847656 L 183.351562 264.847656 C
        <path d="M 261.113281 238.367188 L 233.265625 238.367188 C
        <path d="M 209.578125 238.367188 L 75.636719 238.367188 C 7
        <path d="M 261.113281 211.882812 L 183.351562 211.882812 C
        <path d="M 261.113281 185.402344 L 233.265625 185.402344 C
        <path d="M 209.578125 185.402344 L 181.730469 185.402344 C
        <path d="M 261.113281 158.917969 L 183.351562 158.917969 C
        <path d="M 261.113281 132.4375 L 183.351562 132.4375 C 179.
        <path d="M 261.113281 105.953125 L 233.265625 105.953125 C
        <path d="M 209.578125 105.953125 L 181.730469 105.953125 C
        <path d="M 261.113281 79.472656 L 183.351562 79.472656 C 17
    </g>
</svg>
```

Novel Tasks: GUI & Chart Understanding

BigDocs

- Answer questions & describe UI screenshots.
- Predict user’s action from a UI screenshot
- Interpret data visualizations



GUI-VQA

What colors dominate the design in this aesthetic?

Black, white, and natural wood tones

GUI2Summ

Please summarize the webpage in the image.

This website, DATAPRINT, is a professional design and printing service.

GUI2UserIntent

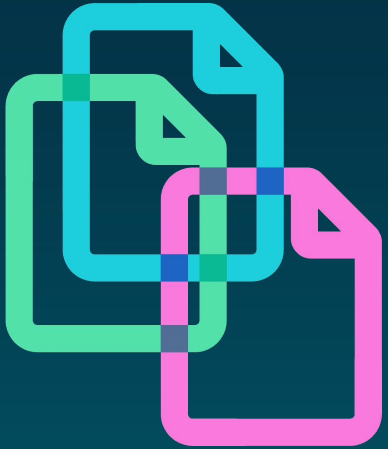
What opens when the user clicks [550, 0, 600, 80]?

Display the PORTFOLIO.

Chart2Caption

This is a heatmap, a graphical representation of data where the individual values contained in a matrix are represented as colors. Here's a summary:
Color Legend*: The color scale on the right side indicates the range from -2 (dark blue) to 2 (dark red). Intermediate shades represent intermediate values.*
Rows and Columns*: Rows: Labeled from row10 to row3.Columns: Labeled from column23 to column10*

Why use **BigDocs**?



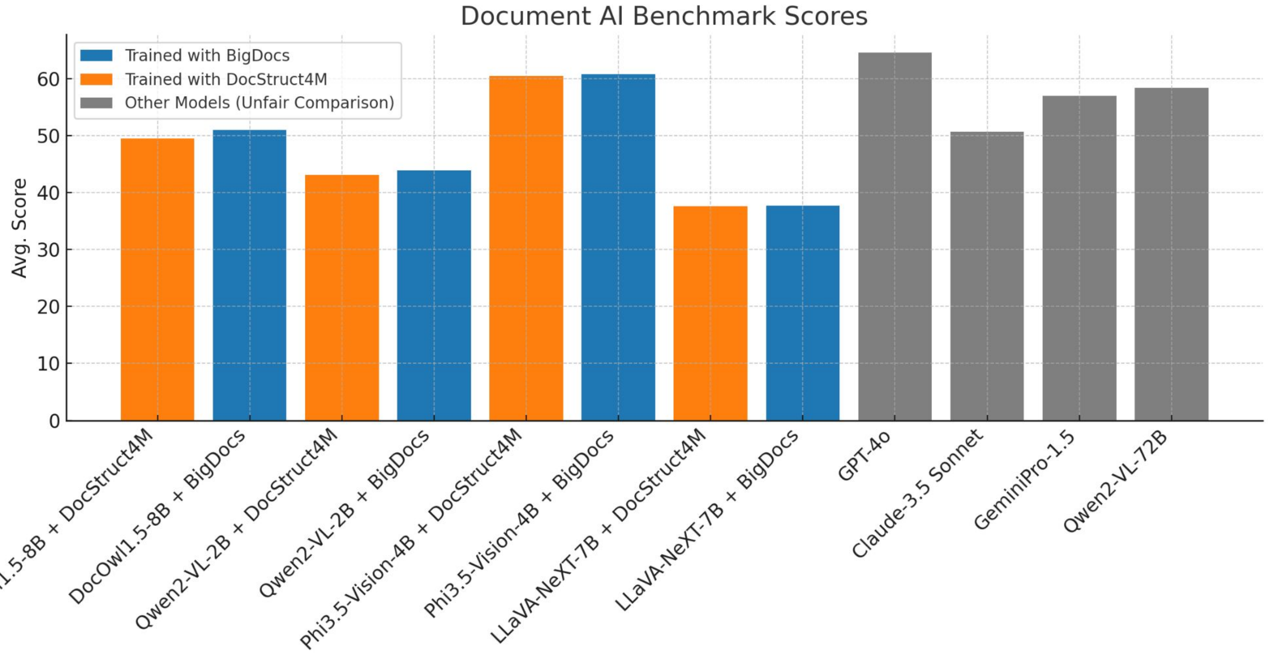
Training & Models

- We **Continually Pre-Trained** open-source VLMs to perform **Document AI**
- We **instruction-tuned** the models for multiple downstream tasks
- We released full **training/inference** pipelines and recipes

Qwen

Results on Document AI Benchmarks

- **Pretraining on BigDocs** surpasses other pre-training approaches
    - No benchmark contamination, transparent, license clarity
- Our best model **matches/outperforms SOTA closed models**



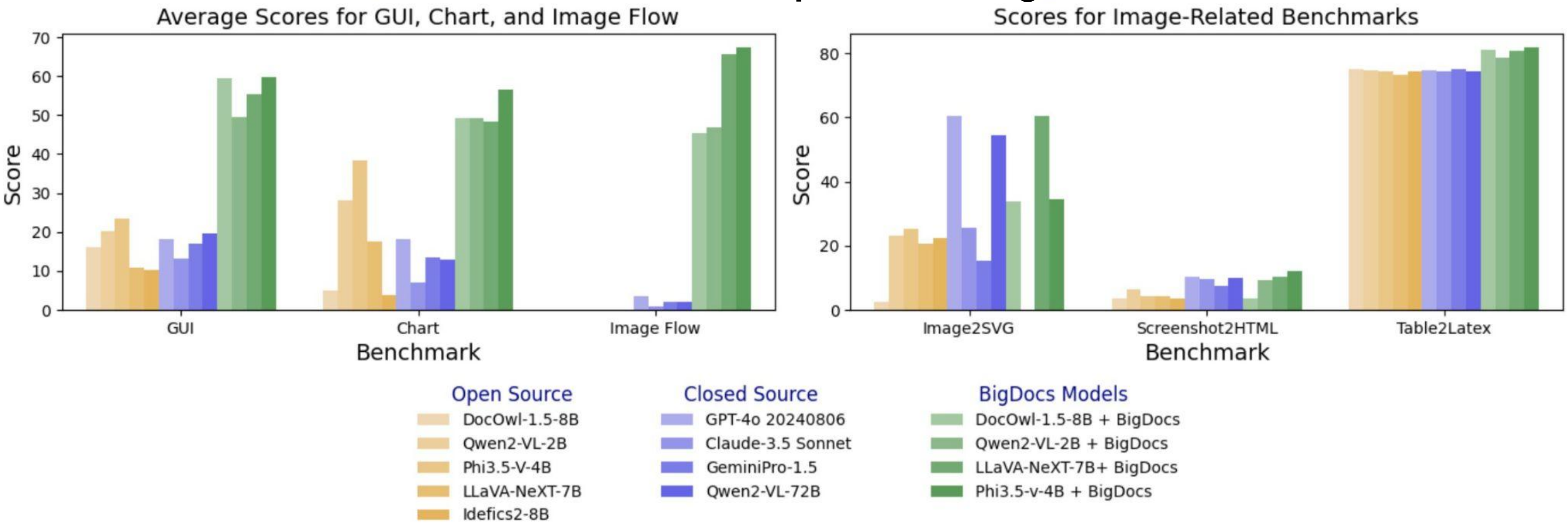
Results on BigDocs-Bench

BigDocs

Models struggle at Multimodal Coding and GUI Reasoning

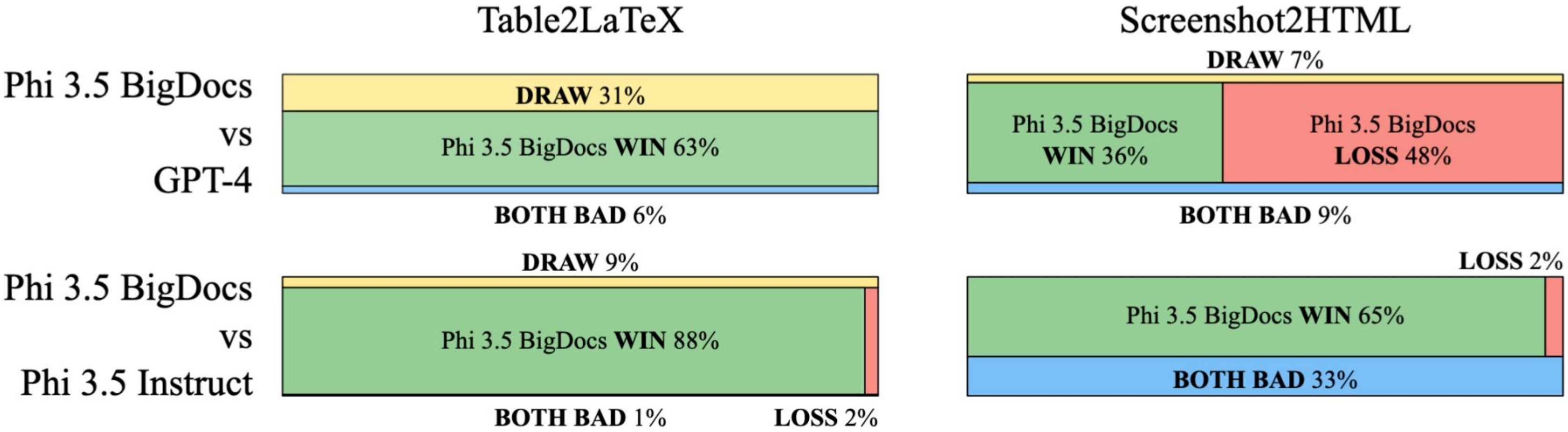
BigDocs teaches new skills, surpassing SOTA models

Model Performance Comparison on BigDocs-Bench



Results - Human Evaluation

- Human evaluation **reveals a clear preference**
- BigDocs **outperforms** GPT4 on **Latex** and **HTML** generation

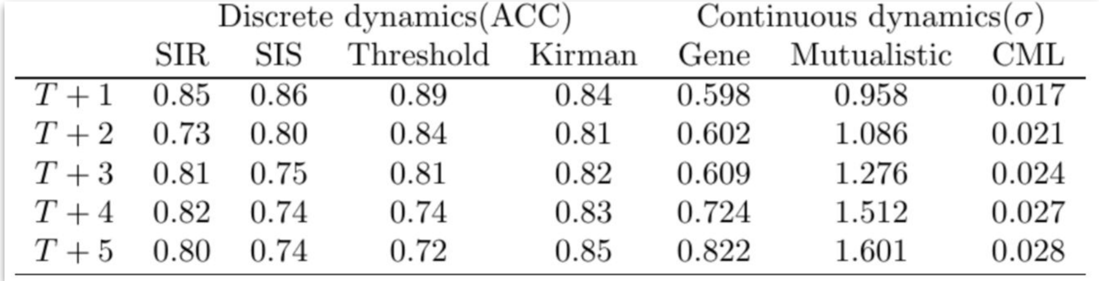


Qualitative Example - Latex Generation

BigDocs

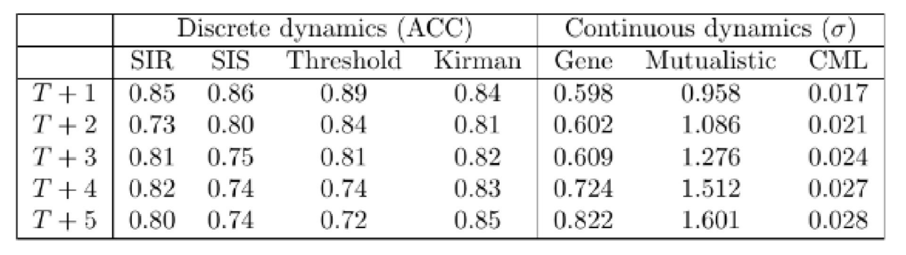
- BigDocs achieves a better table conversion than GPT4

Input Image



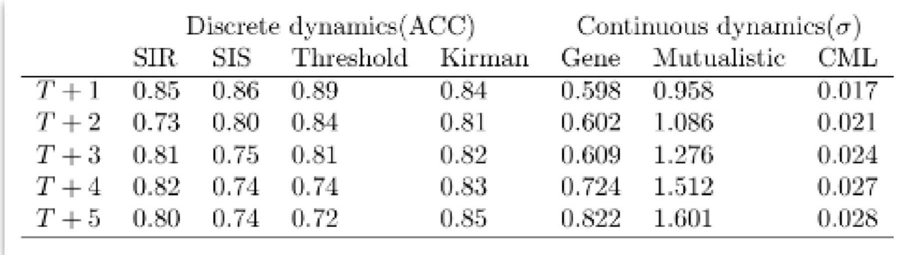
	Discrete dynamics(ACC)				Continuous dynamics(σ)		
	SIR	SIS	Threshold	Kirman	Gene	Mutualistic	CML
$T+1$	0.85	0.86	0.89	0.84	0.598	0.958	0.017
$T+2$	0.73	0.80	0.84	0.81	0.602	1.086	0.021
$T+3$	0.81	0.75	0.81	0.82	0.609	1.276	0.024
$T+4$	0.82	0.74	0.74	0.83	0.724	1.512	0.027
$T+5$	0.80	0.74	0.72	0.85	0.822	1.601	0.028

Output from GPT4o



	Discrete dynamics (ACC)				Continuous dynamics (σ)		
	SIR	SIS	Threshold	Kirman	Gene	Mutualistic	CML
$T+1$	0.85	0.86	0.89	0.84	0.598	0.958	0.017
$T+2$	0.73	0.80	0.84	0.81	0.602	1.086	0.021
$T+3$	0.81	0.75	0.81	0.82	0.609	1.276	0.024
$T+4$	0.82	0.74	0.74	0.83	0.724	1.512	0.027
$T+5$	0.80	0.74	0.72	0.85	0.822	1.601	0.028

Output from BigDocs



	Discrete dynamics(ACC)				Continuous dynamics(σ)		
	SIR	SIS	Threshold	Kirman	Gene	Mutualistic	CML
$T+1$	0.85	0.86	0.89	0.84	0.598	0.958	0.017
$T+2$	0.73	0.80	0.84	0.81	0.602	1.086	0.021
$T+3$	0.81	0.75	0.81	0.82	0.609	1.276	0.024
$T+4$	0.82	0.74	0.74	0.83	0.724	1.512	0.027
$T+5$	0.80	0.74	0.72	0.85	0.822	1.601	0.028

Results – Does BigDocs Help?

💪 BigDocs **adapts eneral multimodal models** to reach **SOTA performance in Document AI**.

🏆 Delivers **+34.5% gains** on **code generation** and **GUI understanding** tasks.

🚀 **Outperforms GPT-4** on **LaTeX** generation and intent prediction from **GUIs**.

Summary & Takeaways

- 📦 **BigDocs-7.5M**: Large open multimodal document dataset
- 🧠 **BigDocs-Bench**: 10 hard tasks, real-world relevance
- 🚀 **Strong gains** – even outperform GPT-4 in key areas
- 🔓 **Fully open** – data, tools, models, benchmarks

bigdocs.github.io

ICLR

BigDocs

An Open Dataset for Training Multimodal Models on Document and Code Tasks

bigdocs.github.io

Juan A. Rodriguez, Xiangru Jian, et. al.
ServiceNow Research

Thanks! Contact us

juan.rodriguez@mila.quebec
xiangru.jian@waterloo.ca

servicenow Mila Université de Montréal ÉTS Le génie pour l'industrie McGill