Table A: **Evaluation results of chart- and document-specific models** on the validation set of CharXiv.

Model	Reasoning Questions					Descriptive Questions					
	All	Text in Chart	Text in General	Num. in Chart	Num. in General	All	Info. Extr.	Enum.	Patt. Rec.	Cntg.	Comp.
Chart-Specific Models											
UniChart-CQA* R3[2]	5.70	3.41	6.06	3.45	12.23	19.32	9.91	**38.26**	12.23	19.08	0.45
TinyChart R1[7]	8.30	5.00	13.13	6.47	14.41	16.15	13.82	14.61	24.67	28.50	3.12
ChartInst.-Llama2 R1[8]	8.80	4.09	23.23	7.76	12.66	**21.40**	23.31	15.50	33.19	27.48	4.91
ChartInst.-FlanT5 R1[8]	11.70	7.95	32.32	9.48	12.23	15.47	11.68	17.59	15.94	**29.52**	**6.70**
ChartAssistant* R1[1]	11.70	9.09	27.27	10.34	11.35	16.93	16.43	16.87	16.59	27.74	2.68
ChartGemma*	12.50	**11.59**	24.24	**16.81**	4.80	21.30	**27.58**	18.97	14.19	19.59	4.46
ChartLlama R1[6]	**14.20**	8.18	**34.34**	9.91	**21.40**	19.23	17.14	12.19	**43.89**	28.75	**6.70**
Document-Specific Models											
TextMonkey R1[4]	3.90	2.50	4.04	3.02	7.42	12.45	12.16	17.92	8.73	6.36	2.68
DocOwl 1.5 Omni R1[5]	9.10	5.45	14.14	9.48	13.54	25.70	34.46	17.92	31.88	17.56	4.46
UReader*	14.30	11.36	18.18	15.52	**17.03**	18.98	10.20	27.60	33.41	20.36	5.36
DocOwl 1.5 Chat R1[5]	17.00	14.32	**34.34**	15.09	16.59	**37.40**	36.83	**49.23**	36.68	22.90	3.12
CogAgent*	**18.80**	**16.82**	32.32	**20.69**	14.85	36.30	**45.14**	26.80	**43.23**	**37.15**	**6.70**

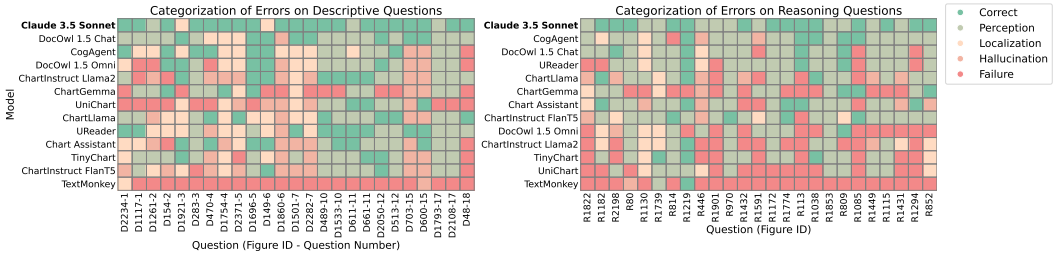

Figure A: **Error categorization** of chart- and document-specific models.

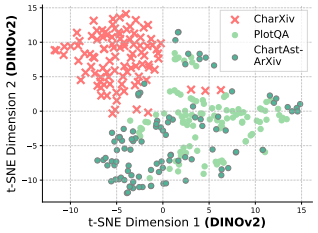

Figure B: **t-SNE visualization of features of chart images *across* chart datasets**. ChartAst R1[1] features highly resemble PlotQA features due to their common synthetic nature.

Table B: **We compare the chart image diversity *within* each dataset** using the **V**endi **S**core from **I**nception and **D**INOv2 features of chart images.

	# Src.	VS_I	VS_D
Figure Dataset/Benchmark			
M-Paper R1[3]	1	17.22	8.81
Chart Dataset/Benchmark			
FigureQA	-	7.08	2.77
DVQA	-	6.17	2.22
PlotQA	-	6.86	2.58
ChartQA	**4**	8.70	2.32
ChartAst R1[1]	-	8.01	3.56
SciGraphQA R1[2]	1	8.83	3.23
OpenCQA R3[1]	1	6.94	2.83
CharXiv (**Ours**)	1	**12.40**	**4.43**

Table C: **Evaluation on automatic grading.** (A): GPT-4o; (B): Claude 3 Sonnet; (C): InternVL Chat V1.5; (D): MGM HD Yi 34B

	(A)	(B)	(C)	(D)
Descriptive Questions				
TP	39	36	28	21
FP	1	0	1	0
FN	0	0	1	0
TN	10	14	20	29
F1	0.99	1.00	0.97	1.00
Reasoning Questions				
TP	24	14	18	14
FP	0	1	0	0
FN	1	1	0	0
TN	25	34	32	36
F1	0.98	0.93	1.00	1.00

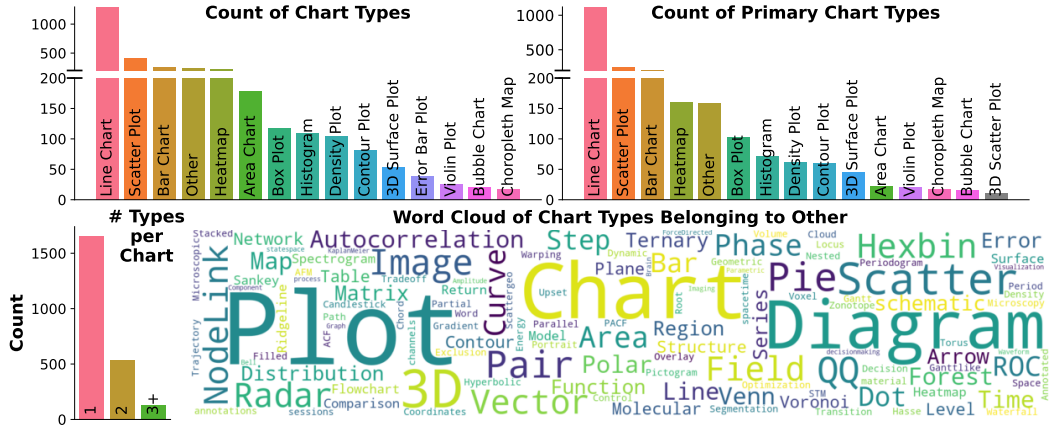

Figure C: **Statistics of chart types.** CharXiv captures a long tail of chart categories in-the-wild.