Understanding Dataset Collection: (Problem-2)

এখন যেহেতু আমরা Problem 1 (Vocabulary Design) এর একটা ক্লিয়ার সলিউশন পেয়েছি, এবার আসা যাক Problem 2 তে—কিভাবে ডেটাসেট ডিজাইন করবো?

ডেটাসেট ডিজাইন পুরোপুরি নির্ভর করে আমাদের মডেল ঠিক কী কাজ করবে তার উপর।আমরা যেহেতু Bangla Handwritten OCR বানাতে যাচ্ছি, তাই প্রথমেই আমার মাথায় যা এসেছে, সেটা হলো—

👉 এই রিলেটেড কোনো পাবলিকলি অ্যাভেইলেবল ডেটাসেট আগে থেকেই আছে কি না? আর থাকলেও তারা কীভাবে তাদের ডেটাসেট ডিজাইন করেছে?

এই প্রশ্নগুলোর উত্তর খুঁজতে গিয়ে যেটা বুঝলাম, সেটা হলো—

আমাদের স্পেসিফিক নিড (Grapheme-level, End-to-End, Word/Sentence OCR) অনুযায়ী কোনো এক্সাক্টলি ম্যাচিং ডেটাসেট এখনো নেই।

আমি নিচে কিছু এক্সিস্টিং ডেটাসেটের লিস্ট দিচ্ছি যেগুলো আমাদের ইন্সপিরেশন দিতে পারে—

Dataset Name	Content Type	Scale (No. of Samples)	Annotation Style	Key Limitation for SOTA ViT Model	Source(s)
NumtaDB	Digits	85,000+	Class label	Digits only; insufficient for general text	[1]

				recognition.	
BanglaLekha-Isolated	Isolated Characters/ Numerals/Compound	166,105	Class label + Metadata	Lacks natural word/line context; model cannot learn inter-character spacing.	[2]
Bengali.AI Grapheme	Graphemes (multi-target)	411,000+	Multi-target class (root, vowel, consonant)	Isolated graphemes, not full words or lines. Requires pre-segmentation.	[3]
BanglaWriting	Word-level	21,234 words	Word BBox + Unicode	Excellent annotation style but limited scale for training a large ViT model from scratch.	[4]
MatriVasha	Compound Characters	2,552	Class label	Focuses only on a subset of compound characters, not comprehensive.	[5]

সবধরনের বাংলা OCR রিলেটেড ডেটাসেট এক এক করে রিভিউ করার পর আমার যেটা মনে হয়েছে—

“BanglaWriting” ডেটাসেট এর ডিজাইনটা আমাদের ফলো করা উচিত।কারণ, এই ডেটাসেটের স্ট্রাকচার আমাদের প্রোজেক্টের গোলের সাথে অনেকটাই মিল খায়।আমরা যেটা বানাতে চাচ্ছি, সেটা হলো—ওয়ার্ড-লেভেল annotated Bangla handwritten OCR system।

→ BanglaWriting এই জিনিসটা অনেক সুন্দরভাবে করেছে।

তারা প্রতিটা ইমেজে ওয়ার্ড লেভেল এনোটেশন দিয়েছে (bounding box সহ)। আর সবচেয়ে বড় কথা, তারা আউটপুট Grapheme-level এ টোকেনাইজ করেছে—যেটা আমরা আগেই ঠিক করে ফেলেছি আমাদের vocabulary টোকেনাইজার হিসেবেও ব্যবহার করবো।

JSON


```
{
  "image_path": "data/images/writer042_form001_20250615.png",
  "image_dimensions": {
    "width": 3500,
    "height": 4950
  },
  "writer_metadata": {
    "id": "writer042",
    "age": 28,
    "gender": "female",
    "handedness": "right"
  },
  "annotations": [
    {
      "bounding_box": ,
      "text": "আমার",
      "graphemes": ["আ", "মা", "র"],
      "is_valid": true,
      "is_overwritten": false
    },
    {
      "bounding_box": ,
      "text": "সোনার",
      "graphemes": ["সো", "না", "র"],
      "is_valid": true,
      "is_overwritten": false
    },
    {
      "bounding_box": ,
```

```
    "text": "বাংলা",
    "graphemes": ["বাং", "লা"],
    "is_valid": true,
    "is_overwritten": false
   },
   {
    "bounding_box": ,
    "text": "দিয়েছেন",
    "graphemes": ["দি", "য়ে", "ছেন"],
    "is_valid": false,
    "is_overwritten": true
   }
  ]
}
```


আমরা শুধু BanglaWriting কে ফলো করবো না,

আমরা আরও এক ধাপ এগিয়ে নিজের মতো করে একটা কাস্টম ডেটাসেট ডিজাইন করবো—যেটা থাকবে:

✅ ওয়ার্ড লেভেল এনোটেশন

✅ লাইন লেভেল এনোটেশন

✅ আর ইমেজ লেভেল তো থাকছেই—but here's the twist:

আমরা ইমেজ লেভেল এনোটেশন ম্যানুয়ালি করবো না!আমরা যে ওয়েবসাইট বানাচ্ছি, সেখানে ইউজার ইমেজ আপলোড করবে এবং সেই ইমেজের মধ্যে লেখা টেক্সটটা টাইপ করে সাবমিট করবে। এতে অটোমেটিকভাবে আমাদের কাছে থাকবে:

→ ইমেজ ফাইল

→ ইমেজের GT (ground truth) টেক্সট

মানে, ইমেজ লেভেল এনোটেশন একপ্রকার অটো জেনারেটেড হয়ে যাবে—যেখানে ম্যানুয়ালি কিছুই করতে হবে না, কিন্তু পারফেক্ট লেবেল পাওয়া যাবে!

1. NumtaDB: Bengali Handwritten Digits - Kaggle, accessed July 10, 2025, https://www.kaggle.com/datasets/BengaliAI/numta
2. (PDF) BanglaLekha-Isolated: A multi-purpose comprehensive ..., accessed July 10, 2025, https://www.researchgate.net/publication/315945090_BanglaLekha-Isolated_A_multi-purpose_comprehensive_dataset_of_Handwritten_Bangla_Isolated_characters
3. (PDF) A Large Multi-target Dataset of Common Bengali Handwritten ..., accessed July 10, 2025, https://www.researchgate.net/publication/354358208_A_Large_Multi-target_Dataset_of_Common_Bengali_Handwritten_Graphemes
4. BanglaWriting: A multi-purpose offline Bangla handwriting dataset ..., accessed July 10, 2025, https://data.mendeley.com/datasets/r43wkvdk4w/1
5. MatriVasha: A Multipurpose Comprehensive Database for Bangla Handwritten Compound Characters This Report Presented in Partial Fu - arXiv, accessed July 10, 2025, https://arxiv.org/pdf/2005.02155