

Synthetic Data Generation & Fine-Tuning Specialist

Author: Santhi Maddipudi

Synthetic Data Generator

Generate realistic datasets using OpenAI or Gemini.

Input

Prompt (natural language)

Hospital patient records with diagnosis and age

JSON Spec (optional — paste or edit)

```
{
  "entity": "employee",
  "count": 20,
  "output_format": "csv",
  "fields": [
    { "name": "employee_id", "type": "integer", "description": "unique ID starting from 1001", "unique": true },
    { "name": "name", "type": "string", "description": "full name" },
    { "name": "department", "type": "string", "values": ["Engineering", "Marketing", "Sales", "HR"] },
    { "name": "salary", "type": "float", "min": 50000, "max": 200000 },
    { "name": "email", "type": "email", "description": "work email" },
    { "name": "hire_date", "type": "date", "format": "YYYY-MM-DD", "min": "2015-01-01", "max": "2024-12-31" }
  ]
}
```

Provider **Format**

openai ▼ csv ▼

Count (overrides spec) **Model override**

20 gpt-4o-mini / gemini-2.0-flash

Output filename (no extension)

hospital_data

Generate

Output

Status / Error

Done — hospital_data generated.

Download

hospital_data.csv 15 KB

Preview

patient_id	name	age	diagnosis	admission_date	email	is
1	James Smith	34	Hypertension	2023-06-14	james.smith@example.com	tri
2	Mary Johnson	28	Flu	2023-08-22	mary.johnson@example.com	fa
3	Robert Williams	45	Heart Disease	2023-09-10	robert.williams@example.com	tri
4	Patricia Brown	62	Diabetes	2023-07-15	patricia.brown@example.com	tri
5	Michael Jones	58	Asthma	2023-04-01	michael.jones@example.com	fa
6	Linda Garcia	39	COVID-19	2023-01-30	linda.garcia@example.com	tri
7	William Miller	24	Flu	2023-09-25	william.miller@example.com	fa
8	Elizabeth Davis	74	Heart Disease	2023-03-13	elizabeth.davis@example.com	tri
9	David Rodriguez	38	Hypertension	2023-02-17	david.rodriguez@example.com	tri
10	Barbara Martinez	56	Diabetes	2023-06-05	barbara.martinez@example.com	fa
11	Richard Hernandez	67	Asthma	2023-03-22	richard.hernandez@example.com	tri
12	Susan Lopez	21	COVID-19	2023-08-05	susan.lopez@example.com	fa

Problem we are solving :

In today's data-driven world, developers, testers, and data scientists frequently face a common challenge: obtaining realistic datasets without exposing sensitive production data. Synthetic Data Generator solves this problem by creating high-quality, customizable datasets for testing, development, analytics, and AI experimentation while preserving privacy and compliance. Synthetic data has become increasingly important for organizations looking to innovate without risking exposure of personally identifiable information (PII). Generate realistic synthetic datasets using LLMs (OpenAI or Gemini). Describe your data in plain English or define it precisely with a JSON spec — output to CSV, JSON, or JSONL.

Features

- **Prompt-driven:** describe your data in natural language and let the LLM build the schema automatically
- **Spec-driven:** define fields, types, constraints, and enum values explicitly via a JSON file
- **Combined mode:** use a spec for structure and a prompt for flavor/context
- **Output formats:** CSV, JSON, JSONL
- **Providers:** OpenAI (default) or Google Gemini
- **Batching:** large record counts are split into batches automatically

- **Web UI:** browser-based Gradio interface for no-code usage

Gihub repo: <https://github.com/santhimaddipudi/syntheticdatagenerator>