Structuring Clinical Notes into Analyzable Data with R and LLMs

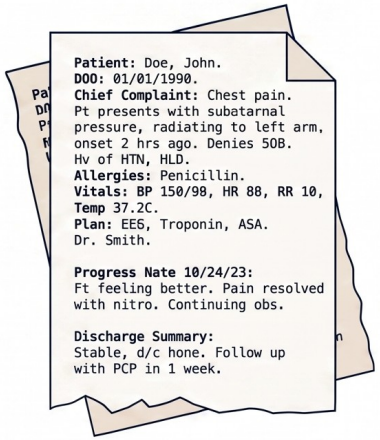
Dr. Wan Nor Arifin

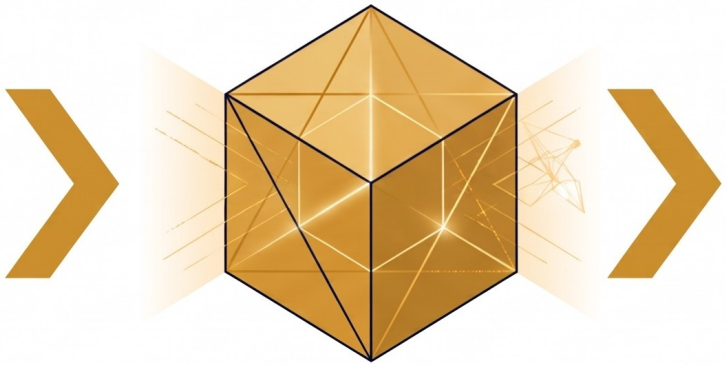
Biostatistics and Research Methodology Unit
Universiti Sains Malaysia
wnarifin@usm.my / wnarifin.github.io

Last update: Jul 21, 2026

From Text To Structured Data

Transforming unstructured clinical narratives into statistically ready datasets





PATIENT_ID	ENCOUNTER_DATE	CHIEF_COMPLAINT	DIAGNOSIS	MEDICATION	ALLERGY	VITALS_BP	VITALS_HR	PROVIDER
P12345	2023-10-24	Chest pain	Angina HTN, HLD	ASA, Nitro	Penicillin	150/90	88	Dr. Smith
P12345	2023-10-25	Follow-up	Resolved	ASA	Penicillin	–	–	Dr. Smith
P12345	2023-10-26	Discharge	Stable	–	Penicillin	–	–	–

Unstructured Text

Free-text clinical narratives, progress notes, and discharge summaries.

Large Language Model (LLM)

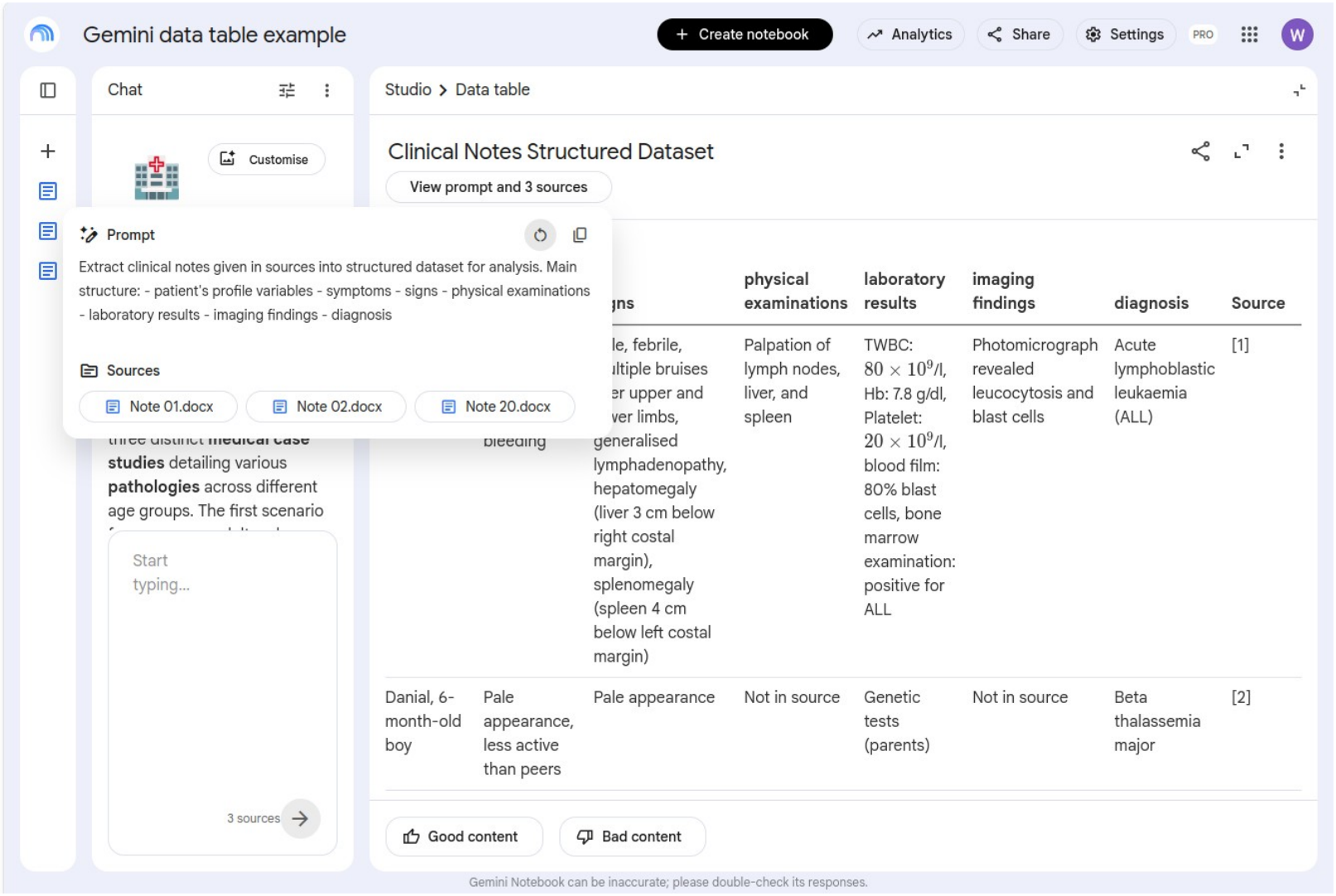
Acts as the processing engine to parse, interpret, and extract clinical entities.

Structured Data

Clean, tabular datasets ready for statistical analysis and modeling.

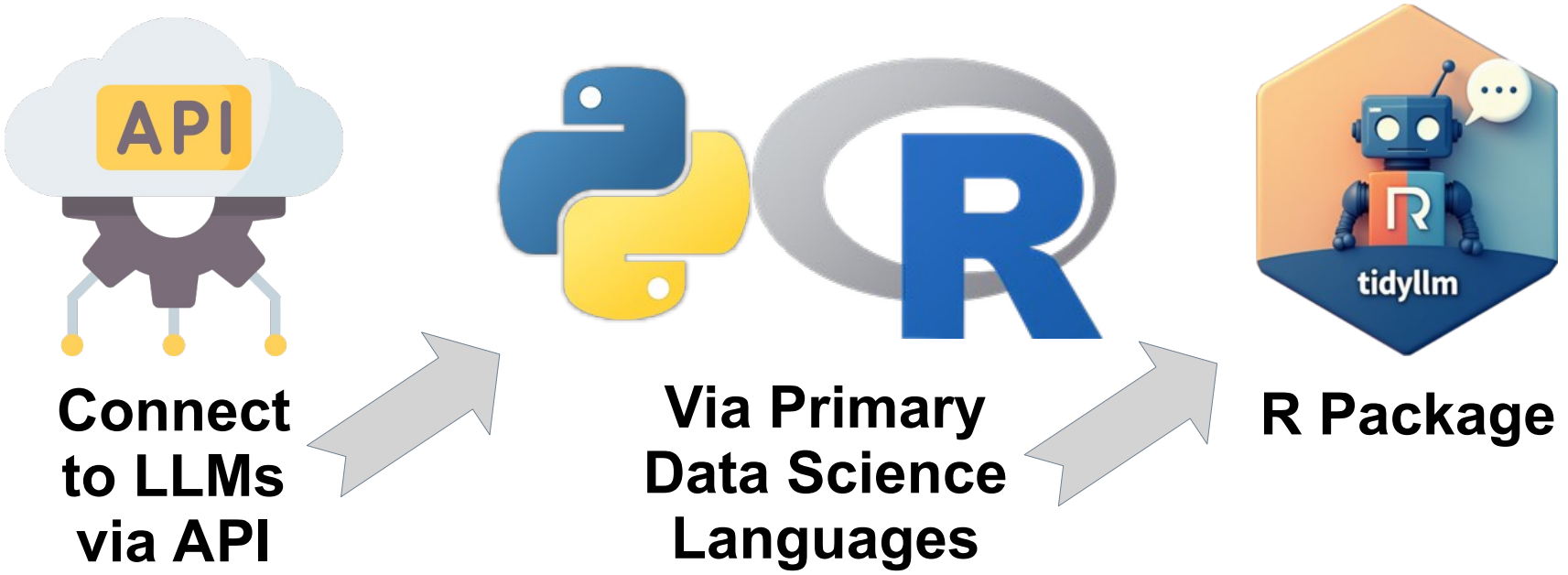
NotebookLM

Gemini Notebook's Data Table



		signs	physical examinations	laboratory results	imaging findings	diagnosis	Source
	bleeding	le, febrile, ltiple bruises er upper and ver limbs, generalised lymphadenopathy, hepatomegaly (liver 3 cm below right costal margin), splenomegaly (spleen 4 cm below left costal margin)	Palpation of lymph nodes, liver, and spleen	TWBC: 80×10^9/l, Hb: 7.8 g/dl, Platelet: 20×10^9/l, blood film: 80% blast cells, bone marrow examination: positive for ALL	Photomicrograph revealed leucocytosis and blast cells	Acute lymphoblastic leukaemia (ALL)	[1]
Danial, 6-month-old boy	Pale appearance, less active than peers	Pale appearance	Not in source	Genetic tests (parents)	Not in source	Beta thalassemia major	[2]

Programmatic Approach In Research & Practice



Images: flaticon.com, r-project.org, python.org, edubruell.github.io/tidyllm

tidyllm

edubruell.github.io/tidyllm

*"**tidyllm** is an R package for working with **large language model APIs** in data analysis workflows. It supports Anthropic Claude, OpenAI, Google Gemini, Mistral, Groq, Perplexity, DeepSeek, OpenRouter, local models via Ollama and llama.cpp, and more — all through a single consistent interface."*

tidyllm

```r
# 1. Define the schema
patient_schema <- tidyllm_schema(
  name = "PatientHistory",
  age = "numeric",
  gender = "character",
  occupation = "character",
  symptoms = "character[]"
)

# 2. Raw unstructured text
raw_text <- "Mrs. Amy, a 35-year-old ward attendant came to the outpatient
             clinic with the complaint of fever for three days and cough."

# 3. Process using a tidy pipeline
full_prompt <- paste("Extract patient note according to the schema.", raw_text)
result <- llm_message(full_prompt) |>
  chat(
    .provider = gemini(), # Gemini as LLM provider
    .model = "gemini-3.1-flash-lite", # Gemini model
    .json_schema = patient_schema # Schema
  )
result

# 4. Structure as JSON
json_result <- get_reply(result) |> minify()
json_result
```

Gemini API

- LLM: Gemini 3.1 Flash Lite
- Get API from Google AI Studio

  aistudio.google.com » Projects » + Create a new project

  aistudio.google.com » API Keys » 🔑 Create API key
- Save API in “api” or “api.txt” in a working folder for our R session

Time for practical session

wnarifin.github.io/workshop.html