



Applications of generative AI to Dataverse and research projects

AI @ FAS Symposium - May 1st, 2024

Stefano M. Iacus

Senior Research Scientist

Director of Data Science and Product Research, IQSS

Affiliate faculty of the Kempner Institute for the Study of Natural and Artificial Intelligence @ Harvard University

“Ask the data” DV chatbot

HARVARD
Dataverse

Add Data - Search - About - User Guide - Support - Sign Up - Log In

Harvard Dataverse >

Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends

Version 1.0

[Access Dataset -](#)
Contact Owner | Share

Description
Research into the causes of violence against civilians has increased significantly in recent years, yet the mechanisms governing spatial patterns of victimization remain poorly understood. My investigation explores if and why one specific locality, capital cities, experiences a higher frequency of violence against civilians perpetrated by armed insurgent organizations. I argue that the political value associated with capitals allows these groups to asymmetrically impose higher costs on the regime by targeting civilians in these localities. I lay out and validate three specific mechanisms to explain this pattern: elite coercion, popular intimidation, and international persuasion. In the first scenario insurgents aim to influence domestic elites directly. In the second, they aim to affect domestic civilians' resolve. In the third, they seek to influence international audiences. Using new geolocated global atrocities data for the years 1996-2009, I evaluate this linkage by employing different methodological approaches and accounting for potential reporting biases. Finally, I show that ethnic and secessionist wars are more likely to experience atrocities in the capital compared with other conflicts. The findings illustrate potential benefits from explaining the temporal and spatial variation in violence by insurgents, with a focus on strategic conditions and power asymmetries. (2017-09-21)

Subject Social Sciences
Keyword Political violence; Civilian victimization; Conflict; Spatial analysis
License/Data Use Agreement CC0 1.0

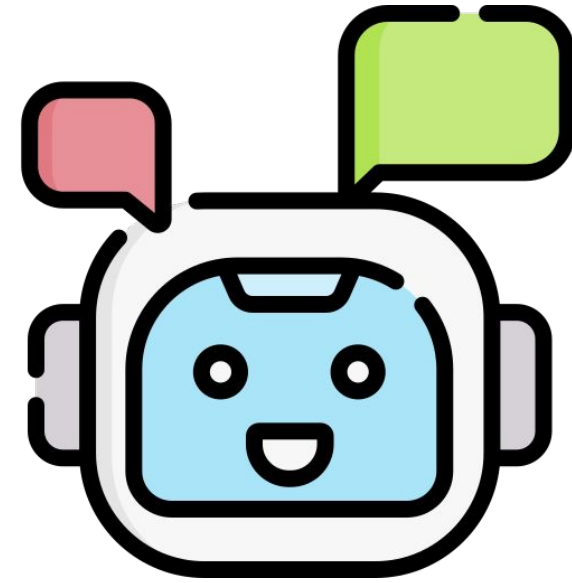
Files | Metadata | Terms | Versions

Change View | Table | Tree

Search this dataset...

Filter by | File Type: All | Access: All | Sort

File Name	Size	Published	Downloads	MDS	Actions
A.8. Forecasting and Cross Validation Exercises.R	4.9 KB	Sep 21, 2017	39	9f8...7a5	
A.9. Capital Atrocities and Regime Survival.R	11.1 KB	Sep 21, 2017	39	7c1...990	
eth_sec_war_dat.tab	50.5 MB	Sep 21, 2017	46	UNF:6:URZ8...MBw===	



This chatbot only sees the tabular data but is clueless about the metadata

Q: Tell me what is this data about

HARVARD
Dataverse

Harvard Dataverse > Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends >

afginc.tab

This file is part of "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends".

Version 1.0

File Citation

Koren, Ore, 2017, "afginc.tab", *Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends*, <https://doi.org/10.7910/DVN/PHHZI7/GVJOH3>, Harvard Dataverse, V1, UNF:6:nUddO4V8XEZZCnhl+9rxYg== [fileUNF]

Cite Data File - Learn about [Data Citation Standards](#).

Dataset Citation

Koren, Ore, 2017, "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends", <https://doi.org/10.7910/DVN/PHHZI7>, Harvard Dataverse, V1, UNF:6:NDQ2oRK6QGGrA+OZgUMdDA== [fileUNF]

Cite Dataset - Learn about [Data Citation Standards](#).

File Tools Metadata Versions

File Tools - Open in New Window

Tell me what you want to know

What is this data about?

Answer please

This data is about latitude, longitude, and event information.

Unnamed: 0	latitude	longitude	event
1	30.9666666667	61.8833333333	2
2	31	64	3
3	31.0077777778	66.4002777778	3
4	31.0744444444	53.3191666667	1

Cool but poor



This chatbot only sees the tabular data but is clueless about the metadata

Q: tell me the range of latitudes and longitudes with the highest number of events

HARVARD
Dataverse

Harvard Dataverse > Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends >

afginc.tab

This file is part of "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends".

Version 1.0

File Citation

Koren, Ore, 2017, "afginc.tab", *Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends*, <https://doi.org/10.7910/DVN/PHHZI7/GVJOH3>, Harvard Dataverse, V1, UNF:6:nUddO4V8XEZZCnhI+9rxYg== [fileUNF]

Cite Data File - Learn about [Data Citation Standards](#).

Dataset Citation

Koren, Ore, 2017, "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends", <https://doi.org/10.7910/DVN/PHHZI7>, Harvard Dataverse, V1, UNF:6:NDQ2oRK6QGGrA+OZgUMdDA== [fileUNF]

Cite Dataset - Learn about [Data Citation Standards](#).

File Tools Metadata Versions

File Tools - Open in New Window

Tell me what you want to know

longitudes with the highest number of events

Answer please

The range of latitudes and longitudes with the highest number of events is 30.96666666666667, 61.88333333333333 to 32.0, 67.25.

Unnamed: 0	latitude	longitude	event
1	30.96666666667	61.88333333333	2
2	31	64	3
3	31.00777777778	66.40027777778	3
4	31.07444444444	53.31916666667	1

ok-ish



This chatbot only sees the tabular data but is clueless about the metadata

Q: map the range of latitudes and longitudes with the highest number of events to the names of countries

HARVARD
Dataverse

Harvard Dataverse > Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends >

afginc.tab

This file is part of "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends".

Version 1.0

File Citation
Koren, Ore, 2017, "afginc.tab", *Why Insurgents Kill Civilians in Capital Cities: A Disaggregated Analysis of Mechanisms and Trends*, <https://doi.org/10.7910/DVN/PHHZI7/GVJOH3>, Harvard Dataverse, V1, UNF:6:nUddO4V8XEZZCnhI+9rxYg== [fileUNF]

Cite Data File ▾ Learn about [Data Citation Standards](#).

Dataset Citation
Koren, Ore, 2017, "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends", <https://doi.org/10.7910/DVN/PHHZI7>, Harvard Dataverse, V1, UNF:6:NDQ2oRK6QGGrA+OZgUMdA== [fileUNF]

Cite Dataset ▾ Learn about [Data Citation Standards](#).

File Tools Metadata Versions

File Tools ▾ Open in New Window

Tell me what you want to know

map the range of latitudes and longitudes with the highest number of events to the names of countries

Answer please

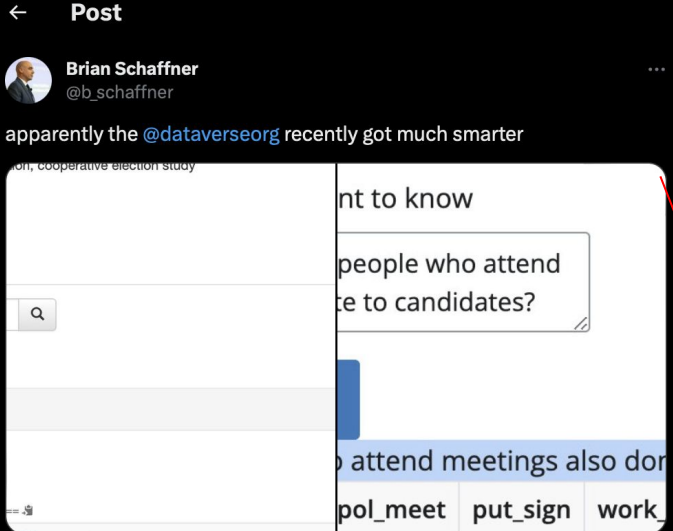
The countries with the highest number of events in the given range of latitudes and longitudes are Afghanistan, Iran, and Pakistan.

Unnamed: 0	latitude	longitude	event
1	30.9666666667	61.8833333333	2
2	31	64	3
3	31.0077777778	66.4002777778	3
4	31.0744444444	53.3191666667	1

AI kicks in



This chatbot only sees the tabular data but is clueless about the metadata



users seems to like it!

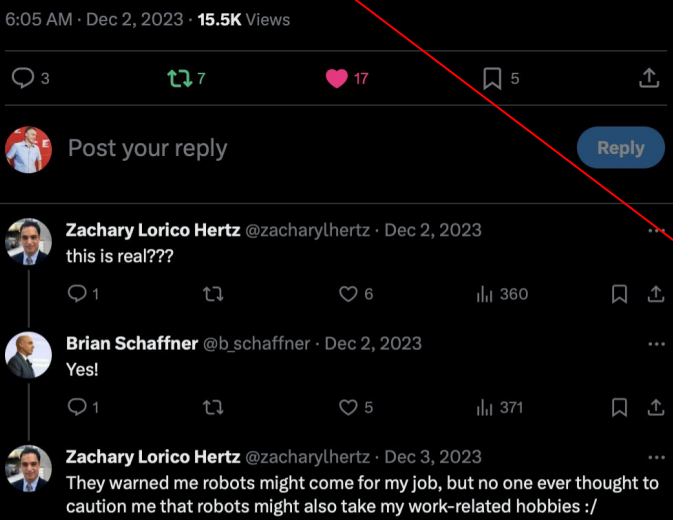
Tell me what you want to know

What proportion of people who attend meetings also donate to candidates?

Answer please

59.3% of people who attend meetings also donate to candidates.

caseid	pid3	pol_meet	put_sign	work_candidate	donate_cand
--------	------	----------	----------	----------------	-------------



'askthedata' is only 181 lines of python code (including comments) most of which is UI interaction handling via shiny app.

<https://github.com/IQSS/askdataverse/tree/main/askthedata>

Covid Project

(Iran University of Science and Technology)

Harvard Dataverse > Covid Project > Smoking Data of hospitalized Covid-19 patients >

Smoking Covid-19 Patients.tab

This file is part of "Smoking Data of hospitalized Covid-19 patients".

Version 1.0

File Citation

Bagheri Nezhad, Sina; Mozayani, Nasser; Abdi, Elham; Rostami, Setareh, 2022, "Smoking Data of hospitalized Covid-19 patients", <https://doi.org/10.7910/DVN/XPT7L0>, Harvard Dataverse, V1; Smoking Covid-19 Patients.tab [fileName], UNF:6:Btd7exaEgPzei7Qy06SPow== [fileUNF]

Cite Data File ▾ Learn about [Data Citation Standards](#).

Dataset Citation

Bagheri Nezhad, Sina; Mozayani, Nasser; Abdi, Elham; Rostami, Setareh, 2022, "Smoking Data of hospitalized Covid-19 patients", <https://doi.org/10.7910/DVN/XPT7L0>, Harvard Dataverse, V1, UNF:6:Btd7exaEgPzei7Qy06SPow== [fileUNF]

Cite Dataset ▾ Learn about [Data Citation Standards](#).

File Tools Metadata Versions

File Tools ▾ Open in New Window

Tell me what you want to know

Answer please

The data is about individuals, including their city, age, sex, and smoking status.

City	Age	Sex	Smoke
Tehran	46	M	No

Access File ▾

Contact Owner Share

File Metrics ⓘ

433 Downloads ⓘ



What happens under the hood?



We create an SQL agent and create a chain of thoughts

1. the persistent identifier (`fileId=4862482`) is sent to the external tool (a shiny dashboard)

2. the data is downloaded and loaded into a SQL dbase (duckdb in this case)

3. the agent receives the query in NL: "what is this data about?"

4. a "chain of thoughts" is started

5. NL is transformed to a SQL query

6. the final answer is sent back in NL form

```

/access/data/1c/
INFO: 127.0.0.1:58143 - "GET /?fileId=4862482&siteUrl=https://dataverse.harvard.edu HTTP/1.1" 200 OK
INFO: ('127.0.0.1', 58145) - "WebSocket/websocket/" [accepted]
INFO: connection open
Loading data
/Users/jago/anaconda3/envs/shinyapp/lib/python3.10/site-packages/duckdb_engine/___init___py:162: DuckDBEngineWarning: duckdb-engine doesn't yet support reflection on indices
  warnings.warn(

> Entering new AgentExecutor chain...
Action: sql_db_list_tables
Action Input: ""
Observation: data
Thought: There is a table named 'data'. I should check its schema to understand what kind of data it contains.
Action: sql_db_schema
Action Input: "data"
Observation:
CREATE TABLE data (
  "City" VARCHAR,
  "Age" VARCHAR,
  "Sex" VARCHAR,
  "Smoke" VARCHAR
)

/*
3 rows from data table:
City  Age  Sex  Smoke
Tehran 46  M   No
Tehran 65  M   Yes
Tehran 39  M   No
*/
Thought: The data seems to be about individuals, with information about their city, age, sex, and whether they smoke or not
Final Answer: The data is about individuals, including their city, age, sex, and smoking status.

> Finished chain.
INFO: connection closed
INFO: connection closed

```

Tell me what you want to know

Who smoke more? Men or women?

Answer please

Men smoke more than women.

City

Age

The chain of thoughts is more interesting for more interesting questions!

*/

Thought: *The "data" table has the columns "Sex" and "Smoke" which are relevant to the question. I can count the number of men and women who smoke by grouping by the "Sex" column and filtering where "Smoke" is 'Yes'. I will write a query for this and check it.*

Action: *sql_db_query_checker*

Action Input: *"SELECT Sex, COUNT(*) as Smokers FROM data WHERE Smoke = 'Yes' GROUP BY Sex"*

Observation: *The original query appears to be correct as it doesn't seem to have any of the common mistakes listed. Here is the reproduction of the original query:*

SELECT Sex, COUNT() as Smokers FROM data WHERE Smoke = 'Yes' GROUP BY Sex*

Thought: *The query is correct. Now I can execute it to get the number of smokers by gender.*

Action: *sql_db_query*

Action Input: *"SELECT Sex, COUNT(*) as Smokers FROM data WHERE Smoke = 'Yes' GROUP BY Sex"*

Observation: *[('M', 59), ('F', 1)]*

Thought: *The result shows that there are 59 male smokers and 1 female smoker. Therefore, men smoke more than women.*

Final Answer: *Men smoke more than women.*

> Finished chain.

Tell me what you want to know

Who smoke more? Men or women? By which amount? and in which cities?

Answer please

Men smoke more than women. In Tehran, 51 men smoke compared to 1 woman. In Rasht, 7 men smoke and there are no women smokers. In Bojnord, 1 man smokes and there are no women smokers.

Thought: The 'data' table contains the columns 'City', 'Sex', and 'Smoke' which are relevant to the question. I can use this information to construct a query that will return the number of men and women who smoke in each city.

Action: `sql_db_query_checker`

Action Input: `"SELECT Sex, City, COUNT(*) as Count FROM data WHERE Smoke = 'Yes' GROUP BY Sex, City ORDER BY Count DESC LIMIT 10"`

Observation: The original query seems to be correct without any common mistakes. Here is the reproduction of the original query:

```
SELECT Sex, City, COUNT(*) as Count
FROM data
WHERE Smoke = 'Yes'
GROUP BY Sex, City
ORDER BY Count DESC
LIMIT 10
```

Thought: The query is correct. Now I will execute it to get the data.

Action: `sql_db_query`

Action Input: `"SELECT Sex, City, COUNT(*) as Count FROM data WHERE Smoke = 'Yes' GROUP BY Sex, City ORDER BY Count DESC LIMIT 10"`

Observation: `[('M', 'Tehran', 51), ('M', 'Rasht', 7), ('F', 'Tehran', 1), ('M', 'Bojnord', 1)]`

Thought: The query results show that men smoke more than women. In Tehran, 51 men smoke compared to 1 woman. In Rasht, 7 men smoke and there are no women smokers. In Bojnord, 1 man smokes and there are no women smokers.

Final Answer: Men smoke more than women. In Tehran, 51 men smoke compared to 1 woman. In Rasht, 7 men smoke and there are no women smokers. In Bojnord, 1 man smokes and there are no women smokers.

> Finished chain.

Tell me what you want to know

Who is the oldest person between smokers? which age he or she has?

Answer please

The oldest male smoker is 88 years old and the oldest female smoker is 67 years old.

It may fail despite some attempt to **self correct** looking at the SQL query error results.

```
Thought:The "data" table contains the necessary information to answer the question. The "Age" column contains the age of the person, the "Smoke" column indicates whether the person is a smoker or not. I need to write a query that selects the oldest smoker.
Action: sql_db_query_checker
Action Input: "SELECT MAX(Age) as Max_Age, Sex FROM data WHERE Smoke = 'Yes'"
Observation: The original query seems to be correct as it doesn't contain any of the common mistakes listed. Here is the reproduction of the original query:

SELECT MAX(Age) as Max_Age, Sex FROM data WHERE Smoke = 'Yes'
Thought:The query is correct. Now I can execute it to get the oldest smoker and their sex.
Action: sql_db_query_checker
Action Input: "SELECT MAX(Age) as Max_Age, Sex FROM data WHERE Smoke = 'Yes'"
Observation: Error: (duckdb.BinderException) Binder Error: column "Sex" must appear in the GROUP BY clause or must be part of an aggregate function.
Either add it to the GROUP BY list, or use "ANY_VALUE(Sex)" if the exact value of "Sex" is not important.
LINE 1: SELECT MAX(Age) as Max_Age, Sex FROM data WHERE Smoke = 'Yes'
^
[SQL: SELECT MAX(Age) as Max_Age, Sex FROM data WHERE Smoke = 'Yes']
(Background on this error at: https://sqlalche.me/e/20/f405)
Thought:The error message indicates that the "Sex" column must appear in the GROUP BY clause or be part of an aggregate function. I will modify the query to include "Sex" in the GROUP BY clause.
Action: sql_db_query_checker
Action Input: "SELECT MAX(Age) as Max_Age, Sex FROM data WHERE Smoke = 'Yes' GROUP BY Sex"
Observation: The original query seems to be correct as it doesn't have any of the common mistakes mentioned. Here is the reproduction of the original query:

SELECT MAX(Age) as Max_Age, Sex FROM data WHERE Smoke = 'Yes' GROUP BY Sex
Thought:The query is correct. Now I can execute it to get the oldest smoker and their sex.
Action: sql_db_query_checker
Action Input: "SELECT MAX(Age) as Max_Age, Sex FROM data WHERE Smoke = 'Yes' GROUP BY Sex"
Observation: [(('88', 'M'), ('67', 'F'))]
Thought:I now know the final answer
Final Answer: The oldest male smoker is 88 years old and the oldest female smoker is 67 years old.

> Finished chain.
```

sometimes it fails miserably to come to a conclusion.

One more thing...

Tell me what you want to know

¿Cuántas personas fuman en este conjunto de datos?

Answer please

60 personas fuman en este conjunto de datos.

Translation comes for free

```
scutor chain...
bles

/"data" seems to be the only table in the database. I should check its schema to see if there's a column
a person smokes or not.

CREATE TABLE data (
  "City" VARCHAR,
  "Age" VARCHAR,
  "Sex" VARCHAR,
  "Smoke" VARCHAR
)
/*
3 rows from data table:
City Age Sex Smoke
Tehran 46 M No
Tehran 65 M Yes
Tehran 39 M No
*/
Thought:The "data" table has a "Smoke" column which indicates whether a person smokes or not. I can count the number of people
who smoke by querying this table.
Action: sql_db_query_checker
Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'"
Observation: The original query seems to be correct as it is simply counting the number of rows where the 'Smoke' column is 'Ye
s'. There are no joins, unions, between statements, or functions with incorrect number of arguments. The data type seems to be
correct as well, assuming 'Smoke' is a text column. The identifiers are also properly quoted.

Here is the reproduced query:

SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'
Thought:The query seems to be correct. Now I can execute it to get the number of people who smoke.
Action: sql_db_query
Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'"
Observation: [(60,)]
Thought:I now know the final answer
Final Answer: 60 personas fuman en este conjunto de datos.

> Finished chain.
```

Tell me what you want to know

¿Cuántas personas fuman en este conjunto de datos?

Answer please

60 personas fuman en este conjunto de datos.

Tell me what you want to know

combien y a-t-il de fumeurs dans cet ensemble de données ?

Answer please

Il y a 60 fumeurs dans cet ensemble de données.

Translation comes for free

ector chain...

les

"data" seems to be the only table in the database. I should check its schema to see if there's a column a person smokes or not.

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

;

Smoke
No
Yes
No

has a "Smoke" column which indicates whether a person smokes or not. I can count the number of people in this table.

Action: sql_db_query_checker

Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'"

Observation: The original query seems to be correct as it is simply counting the number of rows where the 'Smoke' column is 'Yes'. There are no joins, unions, between statements, or functions with incorrect number of arguments. The data type seems to be correct as well, assuming 'Smoke' is a text column. The identifiers are also properly quoted.

Here is the reproduced query:

SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'

Thought: The query seems to be correct. Now I can execute it to get the number of people who smoke.

Action: sql_db_query

Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'"

Observation: [(60,)]

Thought: I now know the final answer

Final Answer: 60 personas fuman en este conjunto de datos.

> Finished chain.

Tell me what you want to know

¿Cuántas personas fuman en este conjunto de datos?

Answer please

60 personas fuman en este conjunto de datos.

Tell me what you want to know

combien y a-t-il de fumeurs dans cet ensemble de données ?

Answer please

Il y a 60 fumeurs dans cet ensemble de données.

Translation comes for free

Tell me what you want to know

hvor mange røykere er det i dette datasettet?

Answer please

There are 60 smokers in this dataset.

```
calculator chain...
files

/data" seems to be the only table
a person smokes or not.

CREATE TABLE user (
  "r"+u" VARCHAR

Smoke
No
Yes
No

e has a "Smoke" column which indicates whether a person smokes or not. I can count the number of people
his table.

Action: sql_db_query_checker
Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'"
Observation: The original query seems to be correct as it is simply counting the number of rows where the 'Smoke' column is 'Yes'. There are no joins, unions, between statements, or functions with incorrect number of arguments. The data type seems to be correct as well, assuming 'Smoke' is a text column. The identifiers are also properly quoted.

Here is the reproduced query:

SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'
Thought: The query seems to be correct. Now I can execute it to get the number of people who smoke.
Action: sql_db_query
Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'"
Observation: [(60,)]
Thought: I now know the final answer
Final Answer: 60 personas fuman en este conjunto de datos.

> Finished chain.
```

Tell me what you want to know

¿Cuántas personas fuman en este conjunto de datos?

Answer please

60 personas fuman en este conjunto de datos.

Tell me what you want to know

combien y a-t-il de fumeurs dans cet ensemble de données ?

Answer please

Il y a 60 fumeurs dans cet ensemble de données.

Translation comes for free

Tell me what you want to know

hvor mange røykere er det i dette datasettet?

Answer please

There are 60 smokers in this dataset.

Tell me what you want to know

このデータセットには喫煙者が何人いますか?

Answer please

このデータセットには60人の喫煙者がいます。

for some reasons
German does not
work for this data

```

calculator chain...
files
"data" seems to be the only table
a person smokes or not.
CREATE TABLE user (
  "r"+u" VARCHAR
Smoke
No
Yes
No
e has a "Smoke" column which indic
his table.
Action: sql_db_query_checker
Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Y
Observation: The original query seems to be correct as it
s'. There are no joins, unions, between statements, or fu
correct as well, assuming 'Smoke' is a text column. The
Here is the reproduced query:
SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'
Thought:The query seems to be correct. Now I can execute it to get the number of people
Action: sql_db_query
Action Input: "SELECT COUNT(*) FROM data WHERE Smoke = 'Yes'"
Observation: [(60,)]
Thought:I now know the final answer
Final Answer: 60 personas fuman en este conjunto de datos.
> Finished chain.

```

このデータセットには60人の喫煙者がいます。

Kono datasetto ni wa 60-ri no kitsuen-sha ga imasu.

⏏ ⏮

There are 60 smokers in this dataset.

One last thing...

Dataset Citation

Paolucci, Matteo, 2020, "Stroke and COVID-19 cases in March 2020 compared to March 2019 in the Italian provinces of Forlì-Cesena and Rimini", <https://doi.org/10.7910/DVN/EVHWSB>, Harvard Dataverse, V2, UNF:6:Cs2UBASEs+8UI/398OA+GA== [fileUNF]

Cite Dataset ▾

Learn about [Data Citation Standards](#).

File Tools

Metadata

Versions

File Tools ▾

 Open in New Window

Tell me what you want to know

create a regression model that explains the number of strokes as a function of covid cases (show me the exact variable names you used). Show me the coefficients and predict the number of strokes when covid cases = 50

Added simple statistical reasoning

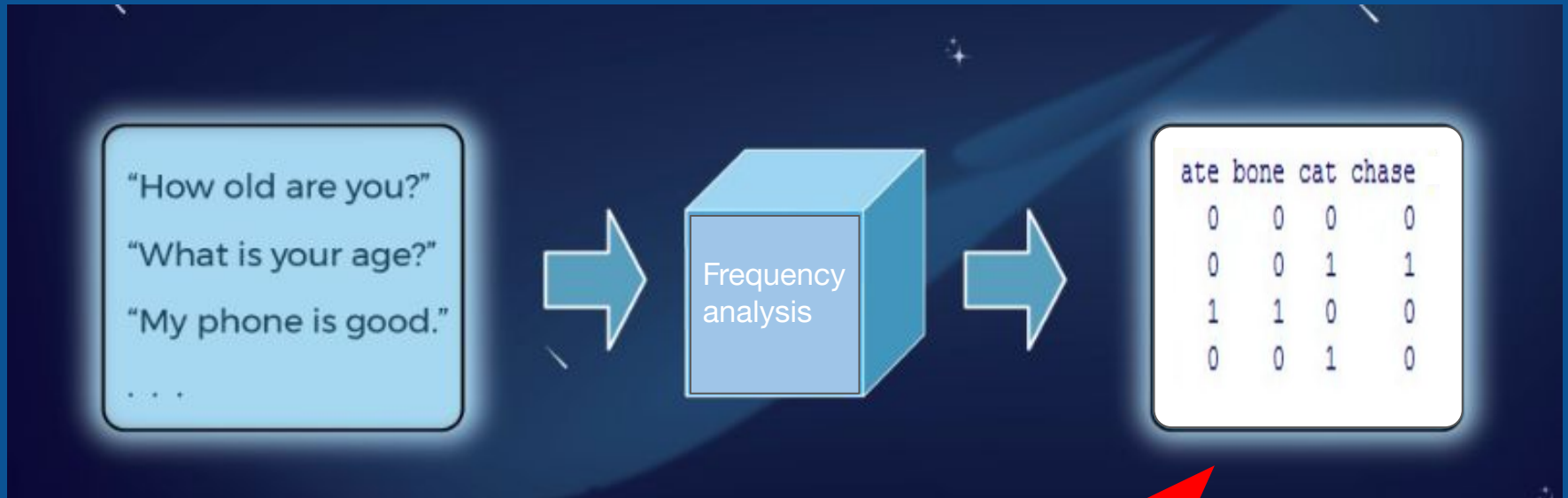
Answer please

The regression model used the variable `COVID-19FCcuminc` to explain the number of strokes (`FCstrokecuminc20`). The coefficients of the model are as follows: - Constant: 1.205324 - COVID-19FCcuminc: 5.159673 When the number of COVID cases is 50, the predicted number of strokes is approximately 259.189.

March2020	COVID-19FCcuminc	COVID-19RNcuminc	COVID-19FC+RNcuminc	E	FCstrokecuminc20	FCcumnumb20	RNstrokecuminc20	RNstrokecum
1	0.25	4.7	2.31		1.2670770307	5	0.5880191106	2
2	0.25	5.59	2.72		2.2807386552	9	1.7640573319	6
3	0.25	7.06	3.4		3.2944002798	13	3.2341051084	11
4	0.51	9.7	4.76		4.8148927166	19	4.4101433297	15
5	1.01	19.99	9.8		6.0819697473	24	5.2921719956	18
6	1.01	27.34	13.2		6.8422159657	27	6.4682102168	22
7	1.77	30.58	15.11		7.0956313718	28	7.9382579934	27
8	3.8	33.22	17.42		8.3627084025	33	8.8202866593	30
9	4.05	48.22	24.5		9.376370027	37	9.9962748806	34

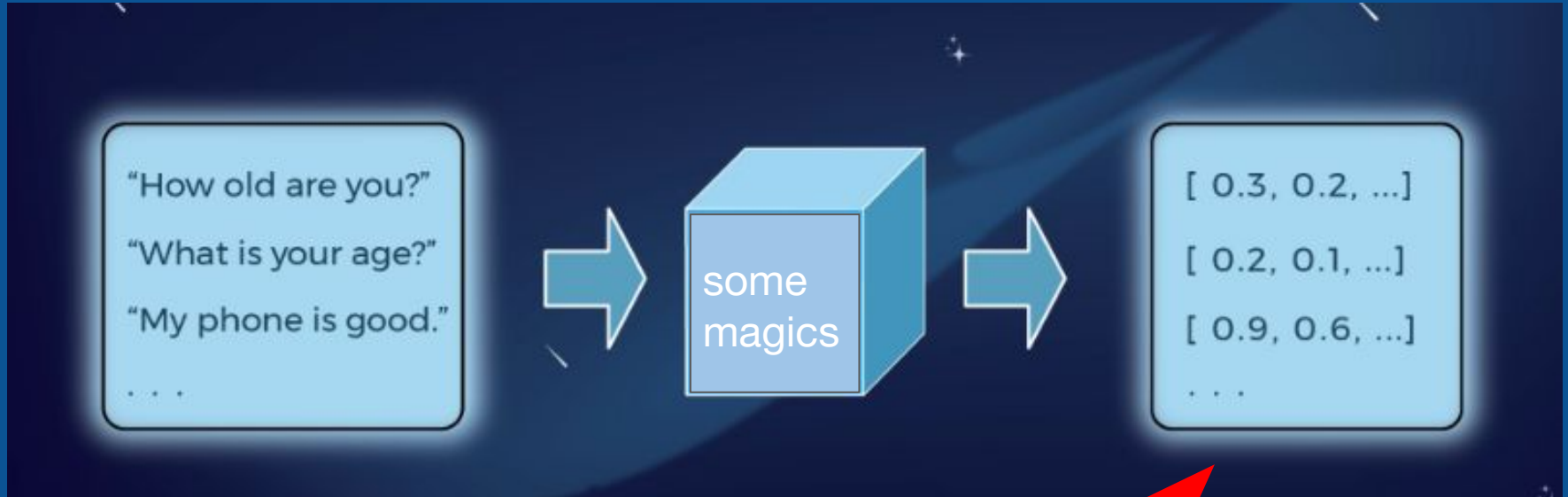
Semantic search via sentence embeddings

Algorithms **do not see words** directly but only numbers representing them (0/1 = absence/presence of a word)



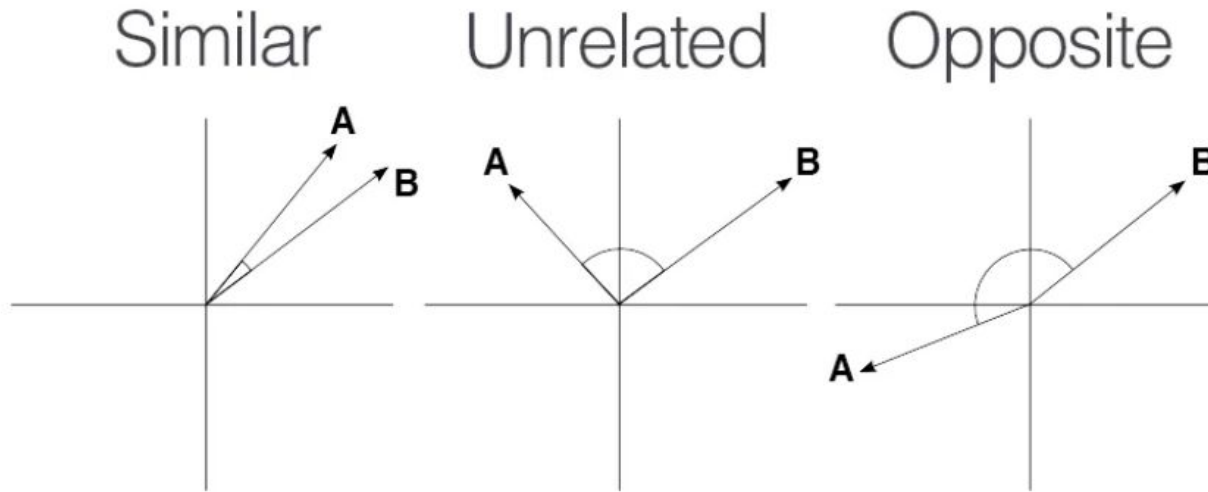
Feature vectors (rows) do not mean anything geometrically

The LLM **does not see words** either but only numerical vectors representing the words in a hypothetical geometric space



The numerical vectors are called “embedding vectors”

Vector space (and embeddings) are meant to have a semantic meaning, their distance represents “semantic closeness”.



A and **B** two different sentences

Metrics 52,999,872 Downloads

Contact Share

covid cases in Italy [Advanced Search](#)

Add Data

Dataverses (1,808)

Datasets (93,089)

Files (114,246)

Dataverse Category

- Research Project (936)
- Researcher (301)
- Organization or Institution (143)
- Research Group (135)
- Journal (81)

[More...](#)

Metadata Source

- Harvard Dataverse (140,126)
- Harvested (69,017)

Publication Year

- 2023 (32,718)
- 2022 (24,348)
- 2021 (11,975)
- 2020 (13,460)
- 2019 (6,845)

[More...](#)

Subject

- Social Sciences (38,782)
- Earth and Environmental Sciences (7,758)
- Medicine, Health and Life Sciences (6,770)
- Agricultural Sciences (4,375)
- Arts and Humanities (3,124)

[More...](#)

Author Name

- U.S. Department of Commerce, Bureau of the Census, Geography Division (3,325)
- Digital Archive of Massachusetts Anti-Slavery and Anti-Segregation Petitions, Massachusetts Archives, Boston MA (3,146)
- Government of Canada (2,668)
- Army Survey Establishment (2,666)
- Curtis A. Bradley (2,234)

[More...](#)

Author Affiliation

- Department of National Defence (2,665)
- Harvard Law School (2,246)
- Yale Law School (2,237)
- Duke Law School (2,235)
- Harvard University (1,747)

1 to 10 of 209,143 Results

Five Shades of Emerging Business Cases (Eliva Press, Moldova, EU)
Apr 18, 2022

This Dataverse has instructor manual and teaching notes for Five Shades of Emerging Business Cases: A collection of 15 case studies on emerging topics for management graduates Pages:235. Published: 2021-11-03. ISBN:978-1636483955.



National and Subnational Estimates of the Covid 19 Reproduction Number (R) for Italy Based on Test Results
Mar 23, 2022 - covid-rt

Abbott, Sam; Bennett, Christopher; Hickson, Joe; Allen, Jamie; Sherratt, Katharine; Funk, Sebastian, 2020, 'National and Subnational Estimates of the Covid 19 Reproduction Number (R) for Italy Based on Test Results', <https://doi.org/10.7910/DVN/8DU5HZ>, Harvard Dataverse, V238, UNF:6:tRgkLNaZWxEiP8D/T4rSv== [fileUNF]

... Identifying changes in the reproduction number, rate of spread, and doubling time during the course of the COVID-19 outbreak whilst accounting for potential biases due to delays in case reporting both nationally and subnationally in Italy. These results are impacted by changes in testing effort, increases ...
Geographic Coverage Country / Nation: Italy
Keyword Term: Italy



National and Subnational Estimates of the Covid 19 Reproduction Number (R) for Italy Based on Test Results
Dec 8, 2021 - covid-rt

Abbott, Sam; Bennett, Christopher; Hickson, Joe; Allen, Jamie; Sherratt, Katharine; Funk, Sebastian, 2021, 'National and Subnational Estimates of the Covid 19 Reproduction Number (R) for Italy Based on Test Results', <https://doi.org/10.7910/DVN/1H11HL>, Harvard Dataverse, V2

... Identifying changes in the reproduction number, rate of spread, and doubling time during the course of the COVID-19 outbreak whilst accounting for potential biases due to delays in case reporting both nationally and subnationally in Italy. These results are impacted by changes in testing effort, increases ...
Geographic Coverage Country / Nation: Italy
Keyword Term: Italy-full



National Reproduction Number (R) Based on Reported Cases
May 10, 2021 - covid-rt

Abbott, Sam; Bennett, Christopher; Hickson, Joe; Allen, Jamie; Sherratt, Katharine; Funk, Sebastian, 2021, 'National Reproduction Number (R) Based on Reported Cases', <https://doi.org/10.7910/DVN/1RMZQN>, Harvard Dataverse, V1

... Identifying changes in the reproduction number, rate of spread, and doubling time during the course of the COVID-19 outbreak whilst accounting for potential biases due to delays in case reporting both nationally and subnationally. These results are impacted by changes in testing effort, increases and decreases ...
Keyword Term: cases-full



National Reproduction Number (R) Based on Reported Cases
Apr 19, 2021 - covid-rt

Abbott, Sam; Bennett, Christopher; Hickson, Joe; Allen, Jamie; Sherratt, Katharine; Funk, Sebastian, 2021, 'National Reproduction Number (R) Based on Reported Cases', <https://doi.org/10.7910/DVN/ERTGJH>, Harvard Dataverse, V20, UNF:6:LICv1Yn1Miz8r5/U29Q== [fileUNF]

... Identifying changes in the reproduction number, rate of spread, and doubling time during the course of the COVID-19 outbreak whilst accounting for potential biases due to delays in case reporting both nationally and subnationally. These results are impacted by changes in testing effort, increases and decreases ...
Geographic Coverage Country / Nation: Italy
Keyword Term: cases



National Reproduction Number (R) Based on Reported Cases
Mar 24, 2022 - covid-rt

Abbott, Sam; Bennett, Christopher; Hickson, Joe; Allen, Jamie; Sherratt, Katharine; Funk, Sebastian, 2020, 'National Reproduction Number (R) Based on Reported Cases', <https://doi.org/10.7910/DVN/TLQRN>, Harvard Dataverse, V280, UNF:6:ZCentizRe+smc5jdTKQA== [fileUNF]



(Keywords) **Query:** "covid cases in Italy"

Traditional Dataverse search based on Solr (Apache Lucene)

datasets about covid cases in

Italy

Run search

Title	Description	DOI
Italy COVID-19 Case Data with Basemap (STC)	Case data from 02-24-2020 to 08-16-2020, this data repository stores COVID-19 virus case data for Italy, including daily case data, summary data, and base map. Each zip file contains weekly case data from Monday to Sunday.	doi:10.7910/DVN/4Z8ZKI
COVID-19 case data for Italy stratified by age class	This dataset contains daily data about COVID-19 cases that occurred in Italy over the period from Jan. 29, 2020 to October 15, 2021, divided into ten age classes of the population, the first class being 0-9 years, the tenth class being >90 years. The dataset contains eight columns, namely: date (day), age class, number of new cases, number of newly hospitalized patients, number of patients entering intensive care, number of deceased patients, number of recovered patients, number of active infected patients.	doi:10.7910/DVN/VSS4CO
Ruiu_Ruiu_ViolationOfLockdownNorms	This dataset contains data about daily new positive cases to COVID19 in Italy and the associated number of daily sanctions imposed by the Police for violations of lockdown norms.	doi:10.7910/DVN/WCRTS3
COVID-19 Hospitalisation, cases and tests in 18 European countries	The dataset contains information on hospitalised COVID-19 patients in acute and intensive care as well as number of COVID-19 cases and number of tests for COVID-19 in Europe. The data stem from a broad data collection initiative encompassing data on 18 European countries and two Italian regions. Countries were included in the data collection of hospitalised COVID-19 patients in acute and intensive care, which was kicked off on 20 March, 2020, once they reported at least five positive cases per 100,000 population. The database has been updated daily since. Data stems from official online sources such as websites of ministries of health, national research and public health institutes, official dashboards from national institutions and Our World in Data (OWID). More information can be found here: https://www.mig.tu-berlin.de/menu/home/akt_de/#c970629	doi:10.7910/DVN/02CFBB
Spatiotemporal Viral Cases Data Collection for COVID-19 Rapid Response	This data repository stores COVID-19 virus case in multi-scale based on ISO standard until 2020-10-21, for daily data access, please visit our GitHub site at: https://github.com/stcenter/COVID-19-Data . Datasets are organized by region area ranging from global to countries as shown below. Underneath each folder, multi-scale daily reports and summary reports are provided separately. For data for each country, start date and all the other information, you can find in the readme file.	doi:10.7910/DVN/8HGECN
Data Covid	The dataset contains information about the COVID-19 virus of different countries and cities. Also, the dataset shows record information about different countries/cities on a daily basis. Dataset also contains records for other countries and cities starting from 2020-3-11 to 2020-3-29.	doi:10.7910/DVN/URHUOV
Coronavirus Disease 2019 (COVID-19) in Italy	Coronavirus Disease (COVID-19) statistics data from Dipartimento della Protezione Civile, Presidenza del Consiglio dei Ministri, Italy, ordered by days and Italian regions. Daily update is available here:	doi:10.7910/DVN/KDFYZW

NL Query: “datasets about covid cases in Italy”



Search via embeddings

NL Query: “datasets su casi di covid in Italia” [LLM kicks in: translation out of the box]

Title	Description	DOI
Ruiu_Ruiu_ViolationOfLockdownNorms	This dataset contains data about daily new positive cases to COVID19 in Italy and the associated number of daily sanctions imposed by the Police for violations of lockdown norms.	doi:10.7910/DVN/WCRTS3
Coronavirus Disease 2019 (COVID-19) in Italy	Coronavirus Disease (COVID-19) statistics data from Dipartimento della Protezione Civile, Presidenza del Consiglio dei Ministri, Italy, ordered by days and Italian regions. /r/n /r/nDaily update is available here: https://github.com/pcm-dpc/COVID-19	doi:10.7910/DVN/KDFYZW
COVID-19 case data for Italy stratified by age class	This dataset contains daily data about COVID-19 cases that occurred in Italy over the period from Jan. 29, 2020 to October 15, 2021, divided into ten age classes of the population, the first class being 0-9 years, the tenth class being >90 years. The dataset contains eight columns, namely: date (day), age class, number of new cases, number of newly hospitalized patients, number of patients entering intensive care, number of deceased patients, number of recovered patients, number of active infected patients.	doi:10.7910/DVN/VSS4CO
Italy COVID-19 Case Data with Basemap (STC)	Case data from 02-24-2020 to 08-16-2020, this data repository stores COVID-19 virus case data for Italy, including daily case data, summary data, and base map. Each zip file contains weekly case data from Monday to Sunday.	doi:10.7910/DVN/4Z8ZKI
COVID-19 Hospitalisation, cases and tests in 18 European countries	The dataset contains information on hospitalised COVID-19 patients in acute and intensive care as well as number of COVID-19 cases and number of tests for COVID-19 in Europe. The data stem from a broad data collection initiative encompassing data on 18 European countries and two Italian regions. Countries were included in the data collection of hospitalised COVID-19 patients in acute and intensive care, which was kicked off on 20 March, 2020, once they reported at least five positive cases per 100,000 population. The database has been updated daily since. /n/nData stems from official online sources such as websites of ministries of health, national research and public health institutes, official dashboards from national institutions and Our World in Data (OWID). /n/nMore information can be found here: https://www.mig.tu-berlin.de/menue/home/akt_de/#c970629	doi:10.7910/DVN/02CFBB
Census of Italian population (aggregated by municipalities) - 2001	Census of Italian population 2001. Data are aggregated by municipalities (8101 cases).	hdl:1902.5/NS_DA0435
Longitudinal Study COVID19	Longitudinal Study COVID19 in Romania	doi:10.7910/DVN/ONZOPT
Replication Data for "Job satisfaction among healthcare workers in the aftermath of the COVID-19 pandemic"	Dataset collected with the surveys administered to Italian healthcare professionals between June 15, 2020 and August 31, 2020	doi:10.7910/DVN/KFOGPD
Spatiotemporal Viral Cases Data Collection for COVID-19 Rapid Response	This data repository stores COVID-19 virus case in multi-scale based on ISO standard until 2020-10-21, for daily data access, please visit our GitHub site at: https://github.com/stccenter/COVID-19-Data . Datasets are organized by region area ranging from global to countries as shown below. Underneath each folder, multi-scale daily reports and summary reports are provided separately. For data for each country, start date and all the other information, you can find in the readme file.	doi:10.7910/DVN/8HGEEN



Search via embeddings

(keyword based) **Query:** “Colorado Adoption Project”

The query
that drives
Sonia crazy



(keyword based)

Query: "Colorado Adoption Project"

The query that drives Sonia crazy



AI generated Sonia

Metrics 59,940,300 Downloads Contact Share

Colorado Adoption Project Advanced Search Add Data

1 to 10 of 27,352 Results Sort

Databases (622)
 Datasets (18,205)
 Files (8,525)

Dataverse Category
Research Project (362)
Researcher (75)
Research Group (58)
Organization or Institution (31)
Teaching Course (7)
More...

Metadata Source
Harvested (14,526)
Harvard Dataverse (12,826)

Publication Year
2024 (210)
2023 (1,736)
2022 (1,241)
2021 (1,278)
2020 (1,171)
More...

Subject
Social Sciences (5,015)
Earth and Environmental Sciences (4,744)
Arts and Humanities (4,352)
Agricultural Sciences (1,621)
Medicine, Health and Life Sciences (1,293)
More...

Author Name
 Esther Jansma (2,342)
Elsemieke Spoor-Hanraets (1,390)
Harvard Map Collection, Harvard College Li-
brary (1,223)

Texas Adoption Project
Apr 14, 2010 - Texas Adoption Project Dataverse
Joseph Horn; John Loehlin, 2010, "Texas Adoption Project", <https://doi.org/10.7910/DVN/XPI0OW>, Harvard Dataverse, V1, UNF:5:ov2C/MUSnhND+HKHhDinxQ== [fileUNF]
... This study presents data from two samples from the **Texas Adoption Project**. The main sample consists of 300 families who **adopted** a child from a Texas home for unwed mothers between 1963 and 1971. Included are ability and personality data from the adoptees and other children in the families, from the **adoptive** mothers and fathers, and from many of the birth mothers of the **adopted** children. ...
Related Publication Citation: Horn, J. M., & Loehlin, J. C. (2010). Heredity and Environment in 300 Families: The Texas Adoption Project. Piscataway, NJ: Transaction Publishers.
Keyword Term: Adoption

Texas Adoption Project Dataverse (The University of Texas at Austin)
Apr 15, 2009
Data from the **Texas Adoption Project**. Includes data from the main sample of 300 **adoptive** families and from an auxiliary sample of 220 families from earlier **adoptions**.

Replication data for "Firewood, forests, and fringe populations: Exploring the inequitable socioeconomic dimensions of Liquefied Petroleum Gas (LPG) adoption in India"
Mar 12, 2021 - LPG adoption in central India
Khanwilkar, Sarika, 2020, "Replication data for "Firewood, forests, and fringe populations: Exploring the inequitable socioeconomic dimensions of Liquefied Petroleum Gas (LPG) adoption in India"", <https://doi.org/10.7910/DVN/SO8P8S>, Harvard Dataverse, V2, UNF:6:AjrjdbO8PNcNe1t+z1ghw== [fileUNF]
... Replication data for "Firewood, forests, and fringe populations: Exploring the inequitable socioeconomic dimensions of Liquefied Petroleum Gas (LPG) adoption in India", Energy Research and Social Science (2020) ...

LPG adoption in central India (Columbia University)
Nov 9, 2020

Colorado Adoption Project, 1976-1989
Nov 27, 2007 - Original Murray Collection
Plomin, Robert; DeFries, John C.; Fulker, David W., 2022, "Colorado Adoption Project, 1976-1989", <https://doi.org/10.7910/DVN/BCDSEU>, Harvard Dataverse, V1, UNF:4:1:71HOxs2b3HEw/SouRIMIBHBcPMBfJaorU/MuAmXCTJM=[fileUNF]

(embeddings)

NL Query: "The Colorado Adoption Project data"

Ask me for some data

The Colorado Adoption Project data

Run search



Title	Description	DOI	Authors	Keywords	Subject	Distance
Colorado Adoption Project, 1976-1989	The Colorado Adoption Project (CAP), begun in 1976, is a longitudinal adoption study that examines genetic and environmental influence on behavioral development. Investigators employed a "full" adoption design by collecting data from the adoptive and biological parents, the adoptees and matched control parents and their children. While the entire data set includes measures from the predominantly white parents, siblings, and focal children (probands) spanning over a twenty year period, the Murray Research Archive has only acquired data on the children from the first seven years of the project and on the parents. Children were given standardized tests of mental and motor development, communication, personality, and temperament. Additional assessments included home observations, information on the physical environment, demographics, the child's birth and the Family Environment Scale. These measures were completed in the homes of the families when the children were 1, 2, 3 and 4 years old. At ages 5 and 6, the parents were surveyed by mail and phone about temperament, health, development of their child and again completed the Family Environment Scale. Murray Research Archive holdings include data from seven waves of data collection on 490 children (245 adopted and 245 controls). The Murray Archive has acquired all numeric file data on the probands and siblings from the first seven waves (ages 1-7), proband data for years 8-12, parental data, as well as videotaped data of the children interacting with their parents (in free play, semi-structured, and teaching situations) from the first three waves (ages 1-3).	doi:10.7910/DVN/BCDSEU	DeFries, John C.; Fulker, David W.; Plomin, Robert	None	Social Sciences	0.269249
Texas Adoption Project	This study presents data from two samples from the Texas Adoption Project. The main sample consists of 300 families who adopted a child from a Texas home for unwed mothers between 1963 and 1971. Included are ability and personality data from the adoptees and other children in the families, from the adoptive mothers and fathers, and from many of the birth mothers of the adopted children. The data include an initial study in the late 1970s, a ten-year follow-up of the children in the late 1980s, interviews with the parents in about half the adoptive families between 1997 and 2002, and a short mail questionnaire in 2002-2005 focused on life outcomes. The second sample consists of 220 families who adopted a child from the same institution prior to 1966. The data are comprised of item responses and scale scores from two personality inventories completed by mail in 1980 by 215 fathers, 219 mothers, 302 adopted children, and 63 biological children of the adoptive parents. The inventories were The California Psychological Inventory and the Thurstone Temperament Schedule. For depositor requirements for access to restricted data, click on 'other information' tab at the bottom of this page.	doi:10.7910/DVN/XPI0OW	None	None	None	0.303842
Denver Family Development Study, 1980-1992	This study investigated the development of marital distress and examined the effects of preventative approaches to marital discord. One hundred and fifty-six couples planning marriage were recruited for a study of relationship development. Each couple participated in two pre-assessment laboratory sessions during which they completed several self-report measures and engaged in videotaped interaction sessions. Questionnaires focused on relationship satisfaction and expectations; interaction sessions consisted of couples either discussing the most important problem in their own relationship or a hypothetical marital dispute (selected from the Inventory of Marital Conflicts). During half of these interaction sessions, couples used a "communication box" to regulate and rate their interaction. This mechanism forced the participants to take turns speaking and allowed the listener to rate the communication s/he just heard. After pre-assessment, eighty-six couples were offered the opportunity to participate in the Prevention and Relationship Enhancement Program (PREP), which is a 5-week relationship enhancement program focused around communication and problem-solving skills. Thirty-three couples completed the program, nine partially completed the program and 43 declined participation. The thirty-three couples who participated became the experimental group, while the 43 who declined and the 70 who were not offered intervention served as control groups. All couples then participated in a post-assessment laboratory session which included self-report measures and interaction sessions similar to those in the pre-assessment. Participants were followed up every year in the laboratory for the next ten years (or until divorced or separated). Follow-up sessions included self-report measures and interaction tasks similar to those in the pre- and post- assessment phases. The Murray Research Archive has 255 videotapes from 63 couples' interaction sessions. The Murray Archive also maintains numeric file data and transcripts of the couple interaction sessions from these 63 couples, and numeric file data from the Interactional Dimensions Coding System (IDCS) data.	doi:10.7910/DVN/WRXUDQ	Markman, Howard J.	None	Social Sciences	0.309214
Family Life Project: A Longitudinal Adoption Study, 1969-	Collectively, the study samples consisted of 158 African-American children ranging from birth to age two, and with approximately equal numbers of females and males. Seventy-five percent of adopting families were from the upper middle and middle class, and 25% were working class.	doi:10.7910/DVN/CRFHDU	Chicago Child Care Society; Shireman, Joan; and Vroegh, Karen	None	Social Sciences	0.310091

Generative AI in research projects

Text classification is a typical **non-scalable** task for humans.

It is also prone to **bias** and uncertainty.

In real applications, typical **inter-coder reliability** is less than 80%.

Social scientists typically apply machine learning models to analyze **large corpuses** of data.

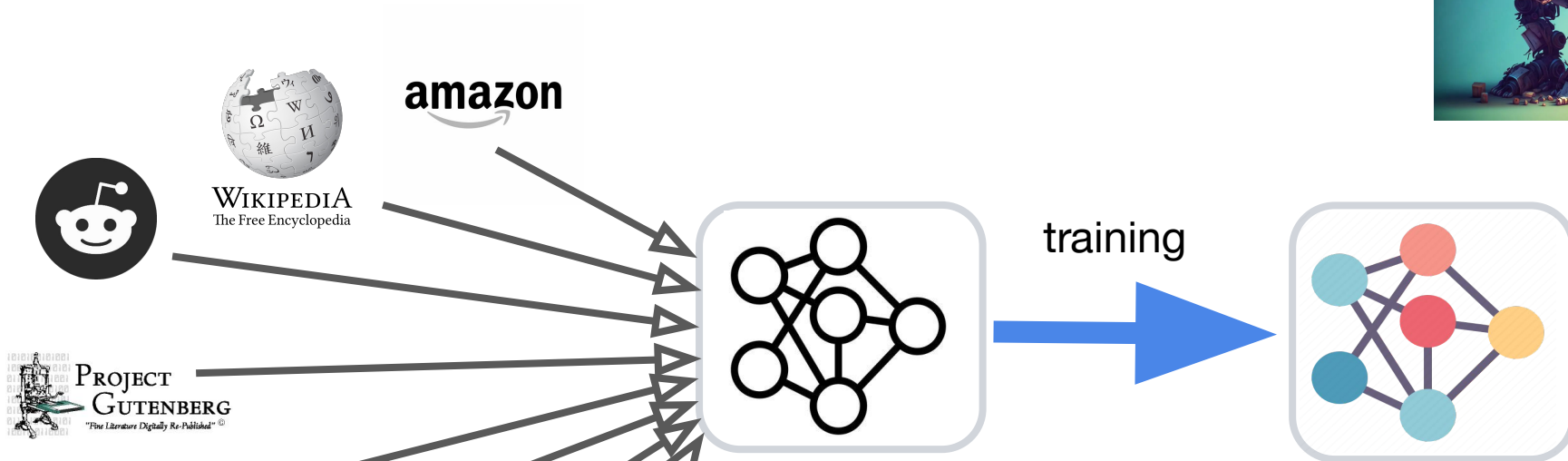
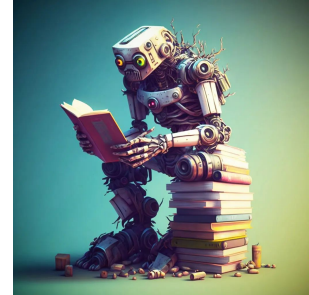
How do LLMs perform in this framework ?

We considered two different approaches:

- text embeddings => require labelled set
- direct LLM classification => does not require labelled set



Unfeasible to train LLM in-house, but fine-tuning is possible



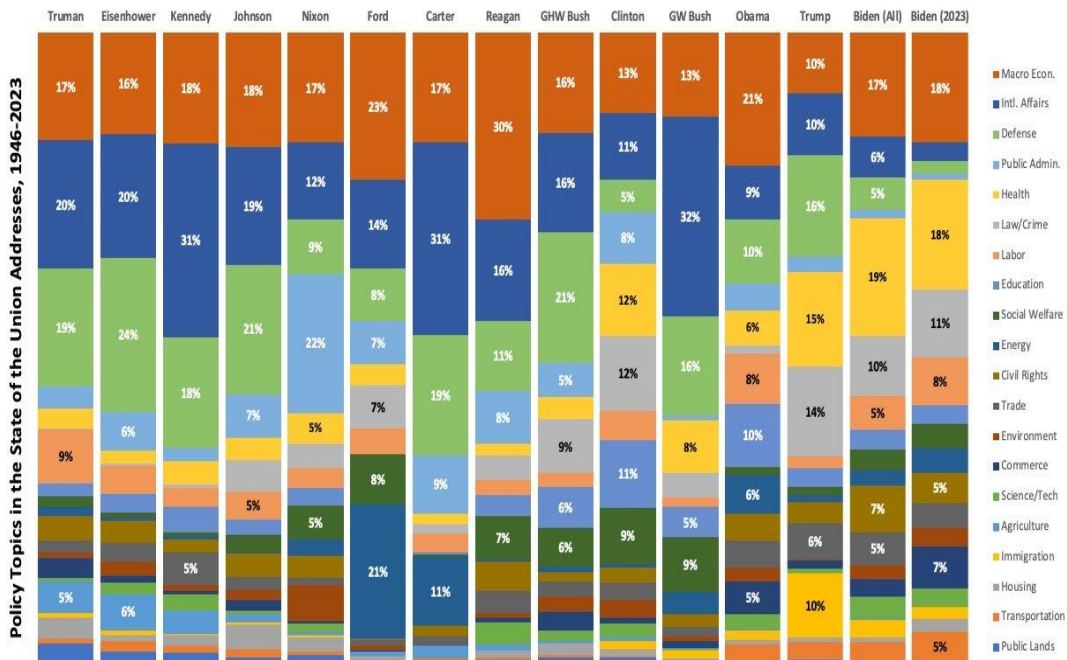
Model	Billions of Tokens (Compute-optimal)	Days to Train on MosaicML Cloud	Approx. Cost on MosaicML Cloud
GPT-1.3B	26B	0.14	\$2,000
GPT-2.7B	54B	0.48	\$6,000
GPT-6.7B	134B	2.32	\$30,000
GPT-13B	260B	7.43	\$100,000
GPT-30B *	610B	35.98	\$450,000
GPT-70B **	1400B	176.55	\$2,500,000



COMPARATIVE AGENDAS PROJECT

COMPARING POLICIES WORLDWIDE

Tracks policy attention to different topics classifying political documents using a common framework (CAP) across countries



20 macro-categories

>200 micro-categories

Best human inter-coder reliability is ~75%

Our objective is the classification of all EU parliamentary questions. As well as local national parliaments.

About 250K documents. The Institute for Quantitative Social Science

Percentages are based on the major topic policy code for presidents' SOTU as delivered over the course of their term. Only those topics with 5 percent or more are indicated with numerals. Source: Policy Agendas Project (<https://www.comparativeagendas.net/us>; accessed on 8 February 2023).



**COMPARATIVE AGENDAS
PROJECT**
COMPARING POLICIES
WORLDWIDE

Vector embeddings + RandomForest **perform as well as the best ML** learning algorithm (XLBert) known for this data.

Advantage of embeddings: multi-language. LLM do no require pre-labelled data (like XLBert).

ChatGPT4 works reasonably well but **does not scale** because of costs

Strategy of LLM classification for CAP.

Task is hard because of the nature of the data. Even human coders do not agree on the classification 25% of the times

Moreover, the **LLM is autoregressive**, i.e., it loses focus on long texts and loses efficiency when has to choose among too many categories.

Best strategy so far:

- For **each macro** category (e.g. economics) we ask the LLM to classify a text according to the **micro** categories, **or NA**.
- We keep the **macro categories** for which micro-categories apply.
- We reclassify the text according to the macro categories

This strategy performs as well as the best fine-trained ML model.

Fine-tuning

Subjective well-being, social media and AI

I. **personal well-being:**

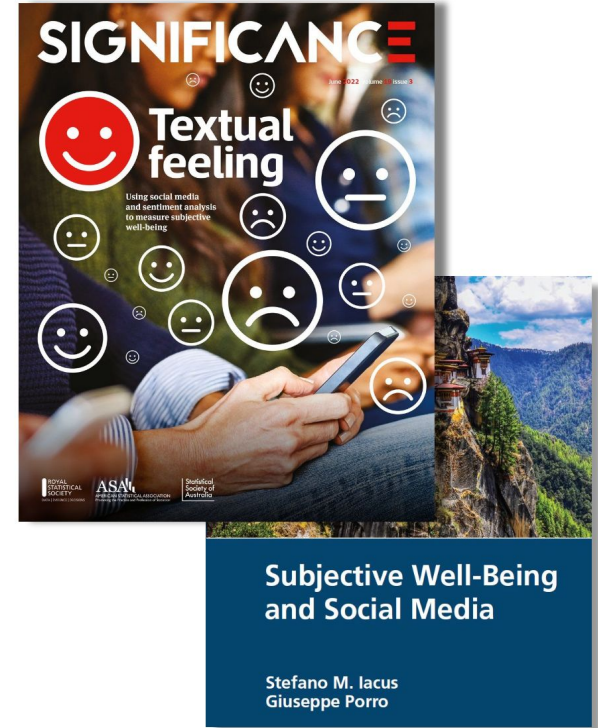
1. **emotional (emo)** well-being: *do you express positive feelings?* [the overall balance between the frequency of experiencing positive and negative emotions, with higher scores showing that positive emotions are felt more often than negative ones]
2. **satisfying life (sat)**: *are you satisfied about your life?* [having positive evaluation of one's life overall]
3. **vitality (vit)**: *do you feel healthy?* [having energy, feeling well-rested and healthy, and being physically active]
4. **resilience and self-esteem (res)**: *are you optimistic about you and your condition?* [a measure of individual psychological resources, optimism and ability to deal with life difficulties]
5. **positive functioning (fun)**: *can you make it?* [feeling free to choose and having the opportunity to do it; being able to make use of personal abilities and feeling absorbed and gratified in activities]

II. **social well-being:**

6. **trust and belonging (tru)**: *do you trust or express gratitude toward the others?* [trusting other people, feeling to be treated fairly and respectfully and feeling sentiments of belonging]
7. **relationships (rel)**: *do you feel alone?* [extent and quality of interactions in close relationships with family, friends and others who provide support]

III. **well-being at work:**

8. **quality of job (wor)**: *are you satisfied in your daily work?* [feeling job satisfaction, satisfaction with work-life balance, evaluating the emotional experiences of work and work conditions]



8 dimensions of subjective well-being. Now extending the results the 43 dimensions of the Human Flourishing program



The Human Flourishing Program
at Harvard's Institute for Quantitative Social Science

Aim: classify 10B tweets of the CGA Geotweet Archive using LLMs.

```
N = 5976, id = 1627621459290890242
tweet = Still in celebratory mode, slide through to find out what we have been able to achieve within two y
ears of our establishment >>>> https://t.co/zGTjCjd8gw
```

```
{
  "Happiness": "high", 0.8,
  "Resilience": "high", 0.7,
  "Self-esteem": "high", 0.6,
  "Life satisfaction": "high", 0.9,
  "Fear of future": "low", 0.3,
  "Vitality": "high", 0.8,
  "Having energy": "high", 0.7,
  "Positive functioning": "high", 0.8,
  "Expressing job satisfaction": "medium", 0.5,
  "Expressing optimism": "high", 0.7,
  "Peace with thoughts and feelings": "high", 0.6,
  "Purpose in life": "high", 0.7,
  "Depression": "low", 0.2,
  "Anxiety": "low", 0.3,
  "Suffering": "low", 0.4,
  "Feeling pain": "low", 0.5
}
```

```
llama_print_timings:      load time = 3048.05 ms
llama_print_timings:      sample time = 201.78 ms / 236 runs ( 0.86 ms
per second)
llama_print_timings: prompt eval time = 1966.44 ms / 193 tokens ( 10.19 ms
per second)
llama_print_timings:      eval time = 18594.90 ms / 235 runs ( 79.13 ms per token, 12.64 tokens
per second)
llama_print_timings:      total time = 21031.11 ms
Llama.generate: prefix-match hit
```

The screenshot shows the Harvard Dataverse interface for the 'Harvard CGA Geotweet Archive v2.0' dataset. The page includes a breadcrumb trail: 'Harvard Dataverse > Center for Geographic Analysis (CGA) Dataverse > Harvard CGA Geotweet Archive >'. The dataset title is 'Harvard CGA Geotweet Archive v2.0' with a 'Version 2.10' label. A citation is provided: 'Lewis, Benjamin; Kakkar, Devika, 2016, "Harvard CGA Geotweet Archive v2.0", <https://doi.org/10.7910/DVNI/3NCMB6>, Harvard Dataverse, V2'. There are links for 'Cite Dataset' and 'Learn about Data Citation Standards'. A 'Description' section explains that the archive maintains a global record of tweets from 2010 to the present, stored on Harvard's High Performance Computing (HPC) cluster. It also includes a 'Subject' section with 'Arts and Humanities; Earth and Environmental Sciences; Social Sciences' and a 'License/Data Use Agreement' section with 'Custom Dataset Terms'.

1) The human reviews the the classification proposed by the LLM.

2) correct classification is used to fine-tune the LLMs

3) fine-tuned LLM run at scale on data

better usage of human resources

CheckIt! Post-verification of text coding | © 2023 SMI

Select your role
none

Total file coded: 0

Text id: 1630812673922277376

RT #URGENT #AB+ blood group required at Gangaram hospital #Lahore. Please contact on 03048621898. #SpreadTheWord #donateblood #savelife #BeAHero #BDP Courtesy @uali337

Show me the tweet!

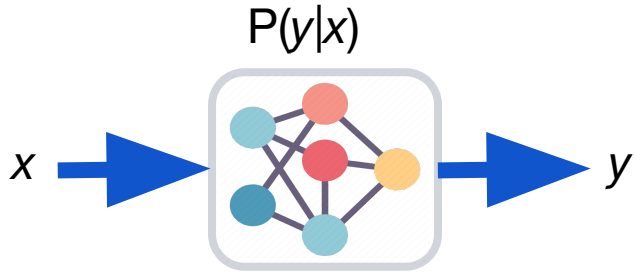
Coding: select only the rows that apply! Skip if uncertain.

dimension	value
expressing altruism	high
volunteering	high
expressing empathy	high

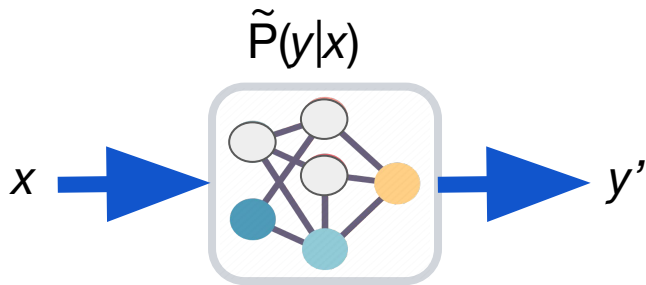
Do not forget to download your codings and the file to this mail.

Download

LLM task: predicting the next word



Fine-tuning does not add knowledge but helps the LLM to focus on certain regions of the NN to produce the answer according to context or examples



○ = nodes of the NN becoming less relevant for the task after fine-tuning (well, the colored ones become more important)

LLAMA2-70B open source version

=====

high medium low not present

high	0.18	0.00	0.00	0.10
medium	0.00	0.04	0.00	0.02
low	0.00	0.00	0.08	0.01
not present	0.17	0.08	0.32	0.00

N = 24243

Accuracy = 30%

systematic
error

ChatGPT-4 (closed model)

=====

high medium low not present

high	0.31	0.00	0.00	0.31
medium	0.00	0.05	0.00	0.07
low	0.00	0.00	0.03	0.14
not present	0.05	0.02	0.02	0.00

N = 10767

Accuracy = 38%

systematic
errors

~

LLAMA2-7B fine tuned

=====

high medium low not present

high	0.30	0.01	0.00	0.18
medium	0.01	0.03	0.00	0.05
low	0.00	0.00	0.07	0.07
not present	0.15	0.03	0.09	0.00

N = 12153

Accuracy = 40%

LLAMA2-70B fine tuned

=====

high medium low not present

high	0.42	0.00	0.00	0.13
medium	0.02	0.06	0.00	0.04
low	0.00	0.00	0.11	0.05
not present	0.09	0.02	0.06	0.00

N = 10671

Accuracy = 59%

LLAMA2-13B fine tuned

=====

high medium low not present

high	0.40	0.01	0.00	0.15
medium	0.01	0.06	0.00	0.04
low	0.00	0.00	0.09	0.07
not present	0.09	0.03	0.04	0.00

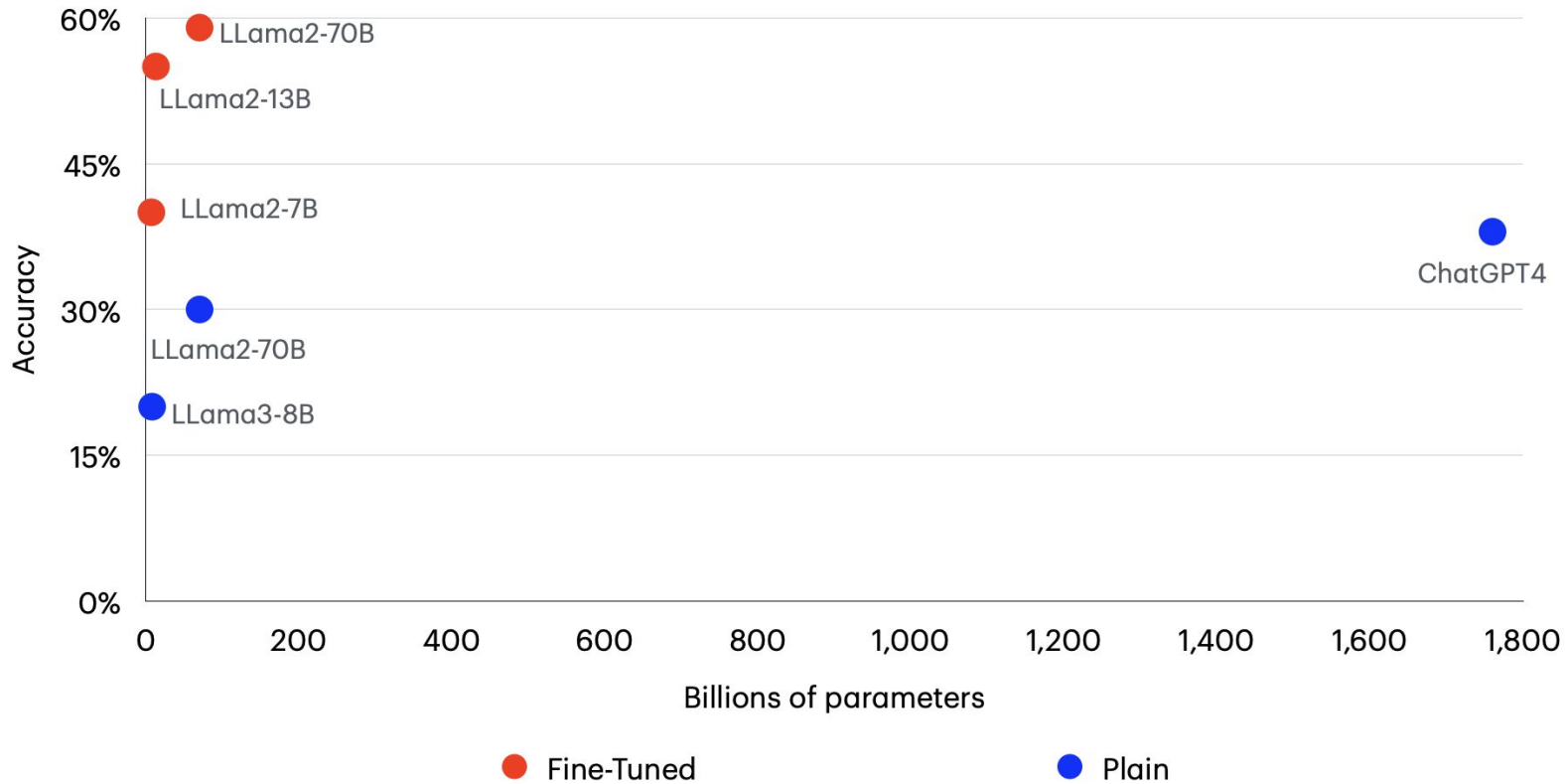
N = 10505

Accuracy = 55%

Commercial models not necessarily good at all tasks, cannot be fine-tuned, expensive

Open models can be fine-tuned up to high accuracy, inexpensive

We are testing LLama3 these days.



Thanks!



Contact information: siacus@iq.harvard.edu



HARVARD
Faculty of Arts and Sciences