CMSC 398z
Effective use of AI Coding Assistants and Agents

Bill Pugh and Derek Willis

Oct 10th, 2025

Readings for next week

A video of a chat Derek and Bill had with Simon Willison

A post by Simon on Vibe engineering

Additional link worth reading:

Political Email Extraction Leaderboard

https://thescoop.org/LLM-Extraction-Challenge/

Today

Mostly time for you to all work on data extraction

Bill has spend hours playing with this, so I'm sure you can keep busy

Feel free to continue the dataset you had been working on, or switch to a new one

Some notes:

- I was surprised that it would take 30-60 seconds to do an FEMA extraction
- Might be worth trying other models, need to add llm plugins to use other models
- With claude, I had to add a delay so I didn't hit the server more than once a minute
- You might need to separately create an API key and provide billing for it
- Command line tool you may not be familiar with: sdiff - side by side diff

sdiff of extraction from sanctions - color added

```
{                                                                  {
  "items": [                                                         "items": [
    {                                                                  {
      "name": "DAVIS, Mary Elizabeth",                                   "name": "DAVIS, Mary Elizabeth",
      "sanction": "Suspension for thirty days (stayed)",        |        "sanction": "Suspension for thirty days (stayed under c
      "requested": "2024-07-09",                                         "requested": "2024-07-09",
      "description": "Reciprocal action from D.C.; representi |        "description": "Reciprocal action from the District of
    },                                                                 },
    {                                                                  {
      "name": "ELAN, Evan Stuart",                                       "name": "ELAN, Evan Stuart",
      "sanction": "Disbarment by Consent",                               "sanction": "Disbarment by Consent",
      "requested": "2024-11-25",                                         "requested": "2024-11-25",
      "description": "Reciprocal action from Virginia and D.C |        "description": "Reciprocal action from Virginia and the
    },                                                                 },
    {                                                                  {
      "name": "FRANKLIN, Jamel R.",                                      "name": "FRANKLIN, Jamel R.",
      "sanction": "Disbarment by Consent",                               "sanction": "Disbarment by Consent",
      "requested": "2024-12-23",                                         "requested": "2024-12-23",
      "description": "Criminal convictions: pled guilty to on |        "description": "Committed criminal acts reflecting adve
    },                                                                 },
    {                                                                  {
      "name": "GALLAGHER, Michele Yvonne",                               "name": "GALLAGHER, Michele Yvonne",
      "sanction": "Indefinite Suspension by Consent",                    "sanction": "Indefinite Suspension by Consent",
      "requested": "2024-09-23",                                         "requested": "2024-09-23",
      "description": "Committed a criminal act reflecting adv |        "description": "Indefinite suspension with reinstatemen
    },                                                                 },
```

Questions for data extraction

- Initial extraction
- Understanding initial variations
- What is the use you might make of this extraction?
- Improving prompt & schema
- Comparing different models, prompts and schemas
- Ways to better prepare source materials
- Examining consistency
  - Figuring out significant vs insignificant differences
  - Will majority voting give you high confidence
- Confidence that the results are "correct"

Unparsable Addresses

Seems like it should be the easiest extraction problem

Might want to use this at huge scale

How to deal with addresses that would be ambiguous to a human

8808-10 Woodyard Service RD

Working with other test data

- Addresses on which regular expression approach produced an answer
  - you might find problems with the regular expression approach
- Synthetic data

FEMA emergency declarations

- Simon just used this as an example of the use of extraction
- You might want to turn several years worth of PDFs into a database
- Different models have different error rates and kinds
  - None of the ones I tried seemed to have problems with character recognition
  - Of the models I tried, the OpenAI models were the worse, including GPT-5
- Some errors were rare (5%), others high (40+%), for exactly the same kind of data
- Some differences in text were insignificant
  - but often consistent across an entire run

preprocessing FEMA declarations

We might want to do this a lot, over several years of data and/or 20 times per date

Would preprocessing the PDF file to extract just the pages with the table help?

- Might make it cheaper/faster
- Would it change the results of the data extraction?

Sanctions database

A challenging example. The sanction field might be underspecified

```
egrep sanction 00.json | sort | uniq -c | sort -r -n > 0.s

egrep sanction 01.json | sort | uniq -c | sort -r -n > 1.s

sdiff -w 150 0.s 1.s
```

```
   3       "sanction": "Disbarred",                                         3       "sanction": "Disbarred",
   3       "sanction": "Disbarment by Consent",                             3       "sanction": "Disbarment by Consent",
   1       "sanction": "Temporary Suspension",                              1       "sanction": "Temporary Suspension",
   1       "sanction": "Suspension for thirty days (stayed)",             |  1       "sanction": "Suspension for thirty days (stayed under conditio
   1       "sanction": "Suspension by Consent for sixty days (all but thi |  1       "sanction": "Suspension by Consent for sixty days (with stay a
   1       "sanction": "Suspension by Consent for six months",              1       "sanction": "Suspension by Consent for six months",
   1       "sanction": "Suspension by Consent for 120 days",                1       "sanction": "Suspension by Consent for 120 days",
   1       "sanction": "Reprimand by Consent",                              1       "sanction": "Reprimand by Consent",
   1       "sanction": "Indefinite Suspension",                             1       "sanction": "Indefinite Suspension",
   1       "sanction": "Indefinite Suspension by Consent",                  1       "sanction": "Indefinite Suspension by Consent",
   1       "sanction": "Commission Reprimand",                              1       "sanction": "Commission Reprimand",
```

Description field is way underspecified

For both of these, might be good to manually extract the desired data from a dozen examples.

Might lead to a refining what the description field should contain, or breaking it down into multiple separate fields

Providing examples to the model can help with getting desired/consistent extraction

Ways of determining if two descriptions are "similar"

- We will get to that next week

Looking at the file

DAVIS, Mary Elizabeth – Suspension for thirty days on July 9, 2024, effective nunc pro tunce to July 29, 2023, stayed under the conditions imposed by the Supreme Court of Maryland, in a reciprocal action from the District of Columbia for representing a client involving a conflict of interest and engaging in conduct that is prejudicial to the administration of justice. The Respondent failed to obtain her client's informed consent regarding a conflict of interest.

ELAN, Evan Stuart – Disbarment by Consent on November 25, 2024, effective immediately, in a reciprocal action from Virginia and the District of Columbia, for failing to represent his clients competently and diligently, failing to adequately communicate with his clients, failing to take steps to protect his clients' interests upon termination of the representation, and engaging in conduct involving dishonesty, fraud, deceit, or misrepresentation. The Respondent abandoned representation of multiple clients.

FRANKLIN, Jamel R. – Disbarment by Consent on December 23, 2024, effective immediately, for committing a criminal act that reflects adversely on his honesty, trustworthiness, or fitness as an attorney; and engaging in conduct involving dishonesty, fraud, deceit, or misrepresentation. The Respondent pled guilty to one count of perjury in violation of Maryland Code, Criminal Law § 9-101; and one count of theft over $100,000.00 in violation of Maryland Code, Criminal Law § 7-104.

GALLAGHER, Michele Yvonne – Indefinite Suspension by Consent on September 23, 2024, effective immediately, with reinstatement conditioned on the satisfactory report of a healthcare professional, for committing a criminal act that reflects adversely on her honesty, trustworthiness, or fitness as an attorney; engaging in conduct involving dishonesty, fraud, deceit, or misrepresentation; and engaging in conduct that is prejudicial to the administration of justice. The Respondent entered an Alford Plea to one count of conspiracy to commit second degree assault.

Census data

CITY	NEGRO POPULATION 1930	1920	1910	Increase or decrease (—) 1920 to 1930 Number	Percent	1910 to 1920 Number	Percent	Percent of total population 1930	1920	1910	PERCENT INCREASE OR DECREASE (—) IN WHITE POPULATION 1920–1930 [1]	1910–1920
New York, N. Y.	327,706	152,467	91,709	175,239	114.9	60,758	66.3	4.7	2.7	1.9	20.7	16.9
Chicago, Ill.	233,903	109,458	44,103	124,445	113.7	65,355	148.2	6.9	4.1	2.0	20.5	21.0
Philadelphia, Pa.	219,599	134,220	84,459	85,370	63.6	49,770	58.9	11.3	7.4	5.5	2.4	15.4
Baltimore, Md.	142,106	108,322	84,749	33,784	31.2	23,573	27.8	17.7	14.8	15.2	5.9	32.1
Washington, D. C.	132,068	109,966	94,446	22,102	20.1	15,520	16.4	27.1	25.1	28.5	8.3	38.4
New Orleans, La.	129,632	100,930	89,262	28,702	28.4	11,668	13.1	28.3	26.1	26.3	14.9	14.6
Detroit, Mich.	120,066	40,838	5,741	79,228	194.0	35,097	611.3	7.7	4.1	1.2	51.4	107.0
Birmingham, Ala.	99,077	70,230	52,305	28,847	41.1	17,925	34.3	38.2	39.3	39.4	47.9	35.1
Memphis, Tenn.	96,550	61,181	52,441	35,369	57.8	8,740	16.7	38.1	37.7	40.0	54.8	28.7
St. Louis, Mo.	93,580	69,854	43,960	23,726	34.0	25,894	58.9	11.4	9.0	6.4	3.5	9.4
Atlanta, Ga.	90,075	62,796	51,902	27,279	43.4	10,894	21.0	33.3	31.3	33.5	30.8	34.0
Cleveland, Ohio	71,899	34,451	8,448	37,448	108.7	26,003	307.8	8.0	4.3	1.5	8.6	38.1
Houston, Tex.	63,337	33,960	23,929	29,377	86.5	10,031	41.9	21.7	24.6	30.4	116.3	90.2
Pittsburgh, Pa.	54,983	37,725	25,623	17,258	45.7	12,102	47.2	8.2	6.4	4.8	11.6	8.3
Richmond, Va.	52,988	54,041	46,733	−1,053	−1.9	7,308	15.6	29.0	31.5	36.6	10.5	45.4
Jacksonville, Fla.	48,196	41,520	29,293	6,676	16.1	12,227	41.7	37.2	45.3	50.8	62.7	76.4
Cincinnati, Ohio	47,818	30,079	19,639	17,739	59.0	10,440	53.2	10.6	7.5	5.4	8.6	7.9
Louisville, Ky.	47,354	40,087	40,522	7,267	18.1	−435	−1.1	15.4	17.1	18.1	33.7	6.2
Indianapolis, Ind.	43,967	34,678	21,816	9,289	26.8	12,862	59.0	12.1	11.0	9.3	14.5	31.9
Norfolk, Va.	43,942	43,392	25,039	550	1.3	18,353	73.3	33.9	37.5	37.1	18.4	70.5

Standard ways to measure confidence on extraction

Can run OCR multiple times

look for major or minor differences

compare different models

Preprocess image

Random sample

forensic accounting techniques

evaluate confidence

Making use of redundant information

- # in 1920
- # change from 1920 to 1930
- # in 1930

If they aren't consistent, we know at least one is wrong

minor differences may be due to difference in less significant digit

Making more use of redundant information

- # in 1920
- # change from 1920 to 1930
- % change from 1920 to 1930
  - approximate
- # in 1930

If 3 are consistent, they are likely correct and can predict the 4th