

From Codebooks to Promptbooks

Extracting Information from Text
with Generative LLMs

Oscar Stuhler
Cat Dang Ton
Étienne Ollion

Special Issue on GenAI in *Sociological Methods and
Research*, 2025



Sven SachsAlbert, **Needle in a Haystack**

Introduction

Generative AI as a Tool for Social Scientific Research

- Agent-based Models (Gao, 2023)
- Simulate Human Subjects (Argyle et al., 2023, [Boelaert et al., 2024](#))
- Help in questionnaire design (Götz, 2024)
- Missing data imputation (Kim and Lee, 2023)
- Content classification: text ([Do et al., 2022](#); [Gilardi et al., 2023](#)), photo ([Cornet et al., 2024](#)), video ([Hwang & Naik, 2023](#); Davidson, 2024)
- ... (see Bail, 2024)

Introduction

What about *Information Extraction*?

- A classic NLP task (Grishman 2022)
- From unstructured to structured data

Introduction

What about *Information Extraction*?

Doing IE with Generative AI

- Does it work? How complex is it?
⇒ An experiment
- How is it affecting the way we do content analysis?

Introduction

Argument

- We can now do a lot of IE, at scale, with gen LLMs
- This is an opportunity to increase transparency about data production

1. What is IE
2. Case: Obits
3. Methods
4. Results

What is IE, and why does it Matter?

What is IE, and why does it Matter?

Information extraction

“The process by which the unstructured information embedded in texts is identified and turned into structured data”

What is IE, and why does it Matter?

Unstructured text

News

John Doe, the highly regarded and long-serving CEO of Company A, which has been at the forefront of innovation for over two decades, announced today during a press conference in Baltimore that the company will acquire Company B, a leading competitor in the ABC industry. The deal is valued at a staggering \$6.2 billion and will be financed through a complex combination of cash and stock. [...]

Text with identified entities and events

[PERJohn Doe], the highly regarded and long-serving [POSITIONCEO] of [ORGCompany A], which has been at the forefront of innovation for [TIMEover two decades], announced today during a press conference in [LOCBaltimore] that the company will [EVENTacquire] [ORGCompany B], a leading competitor in the [INDUSTABC industry]. The deal is valued at a staggering [MONEY\$6.2 billion] and will be financed through a complex combination of cash and stock."

Filled information templates

Event type: Acquisition

Acquiring	"Company A"
Target	"Company B"
Deal value	"\$6.3 billion"

Relation type: Employment

Person	"John Doe"
Role	"CEO"
Organization	"Company A"

What is IE, and why does it Matter?

A Rare but not Absent Endeavor in Sociology

- Franzosi (1990) Subject-Action-Object [later Tilly 1997]
- Practices & Identities (Mohr, 1994; Carley, 1994)
- Representations & action (Bearman & Stovel, 2000)
- Events (Braun & Koopmans, 2014)

What is IE, and why does it Matter?

Although

- Dependency parsers & NER (Mohr et al. 2013; Knight (2022)
- SVO Triplets (Goldenstein & Poschmann, 2019)
- Dependency Parsers (Stuhler, 2021; 2024)

Reason?

- Many tasks, many tools, many options

What is IE, and why does it Matter?

Towards more

- Accessibility
- Flexibility
- Reproducibility

A Rich Case Study: Obits

EUROPE Silvio Berlusconi, a Showman Who Upended Italian Politics and Culture, Dies at 86

June 12, 2023

[Leer en español](#)

ROME — Silvio Berlusconi, the brash media mogul who revolutionized Italian television using privately owned channels to become the country's most polarizing and prosecuted prime minister over multiple stints in office and an often scandalous quarter-century of political and cultural influence, died on Monday at San Raffaele Hospital in Milan. He was 86.

His death was confirmed in a statement by Prime Minister Giorgia Meloni, with whom he was a coalition partner in the current Italian government. No cause of death was given, but he was hospitalized last week as part of his treatment for chronic leukemia and other ailments.

To Italians, Mr. Berlusconi was constant entertainment — both comic and tragic, with more than a touch of off-color material — until they booed him off the stage. But he kept coming back. To economists, he was the man who helped drive the Italian economy into the ground. To political scientists, he represented a bold new experiment in television's impact on voters. And to tabloid reporters, he was a delicious fount of scandal, gaffes, ribald insults and sexual escapades.

A gifted orator and showman who sang on cruise ships as a young

June 12, 2023

Leer en español

ROME — Silvio Berlusconi, the brash media mogul who revolutionized Italian television using privately owned channels to become the country's most polarizing and prosecuted prime minister over multiple stints in office and an often scandalous quarter-century of political and cultural influence, died on Monday at San Raffaele Hospital in Milan. He was 86.

His death was confirmed in a statement by Prime Minister Giorgia Meloni, with whom he was a coalition partner in the current Italian government. No cause of death was given, but he was hospitalized last week as part of his treatment for chronic leukemia and other ailments.

To Italians, Mr. Berlusconi was constant entertainment — both comic and tragic, with more than a touch of off-color material — until they booed him off the stage. But he kept coming back. To economists, he was the man who helped drive the Italian economy into the ground. To political scientists, he represented a bold new experiment in television's impact on voters. And to tabloid reporters, he was a delicious fount of scandal, gaffes, ribald insults and sexual escapades.

A gifted orator and showman who sang on cruise ships as a young man, Mr. Berlusconi was first elected prime minister in 1994, after

Name	Silvio Berlusconi
Date of birth	Sept. 29, 1936
Age at Death	86
Born in	Milan
Cause of death	Chronic leukemia
Children	5 (3 daughters, 2 sons)
...	

A Rich Case Study: Obits

Why Obits?

- The death (beat) is not a dead end: Sources for Cultural Sociology
- Formulaic, but writing is diverse
 - "Using as few words as possible to say as much as possible" (a NYT writer in the documentary *Obituaries*)
- Several tasks

⇒ Quite representative of the diversity of available tasks

A Rich Case Study: Obits

Data

- *New York Times* Obits, 1980-2024
- 80,000 unique articles
- Average token length = 706 words, important variations

Methods

Variables and Tasks

Variable	Task type	Format	Notes
Age (in years)	String extraction	Integer	Omnipresent, place, wording varies
Cause of death	String extraction	Text string	Omnipresent, but not standardized (wording, place)
Institutions of higher education attended	String extraction, inference	Text string	Inference “played Baseball when at Michigan” (positive), “received a honorary degree from Texas A&M” (negative)
Religious affiliation	String extraction, inference	Text string	Rare. Only a person’s voluntary association or lifestyle (e.g., donated to a specific church or active roles at places of worship)
Origin	String extraction, inf, exogenous knowledge	Pre-formatted text string	Numerous places often listed (inference)
Place lived last	String extraction, inf, exogenous knowledge	Pre-formatted text string	Inference because we wanted to distinguish place of death from this

Variable	Task type	Format	Notes
Gender	Inference	Categorical (3)	From textual cues
Military Service	Inference	Binary	Sometimes not explicit
Number of Children	Inference	Integer	Scattered throughout the text, wording sometimes complex
Highest level of education	Inference	Categorical (5)	Zero-shot, 5 valid options
Survivors	Summarization	Pre-formatted text string	Standardizing information contained in various places
Occupation	Summarization, inference	Text string	Many occupations

Methods

Prompting

- In natural language

```
Below I will provide an obituary of a deceased
person
Based on the text, infer the gender of the deceased
person. Provide a one-word response
from only one of the following options: [male],
[female], [other]
Your answer should be formatted like this:
[gender]
Here is the obituary date, title, and text: {print
obituary}
```

Methods

Prompting

- In natural language
- Built iteratively (dev set = 300 obits)
- Long, very specific (up to 400 words)

Methods

Model

- Llama 3 70-B, quantized (Q5), temperature 0
 - 50 GB
 - Ran on 2 NVIDIA V100 PCIe GPUs with 32 GB
- Test set: 200 obituaries (random)

Methods

What about contamination?

retrieve the obituary for etienne ollion

After conducting a search, I found the obituary for Étienne Ollion, a French sociologist. Here is the relevant information:

Name: Étienne Ollion

Born: 1982

Died: May 25, 2019

Age: 37

Cause of Death: Cancer

Occupation: Sociologist

Obituary Excerpt:

"Étienne Ollion, a brilliant and promising young sociologist, died on May 25, 2019, at the age of 37, after a long battle with cancer. He was a researcher at the CNRS and a lecturer at the École des Hautes Études en Sciences Sociales (EHESS). His work focused on the sociology of elites, social networks, and the study of power relations."

Source: Le Monde, June 2019



Everything except
CNRS is wrong on this
page...

Methods

What about contamination?

ROME — [John Doe](#), the brash media mogul who revolutionized Italian television using privately owned channels to become the country's most polarizing and prosecuted prime minister over multiple stints in office and an often scandalous quarter-century of political and cultural influence, died on Monday at San Raffaele Hospital in Milan. He was 86.

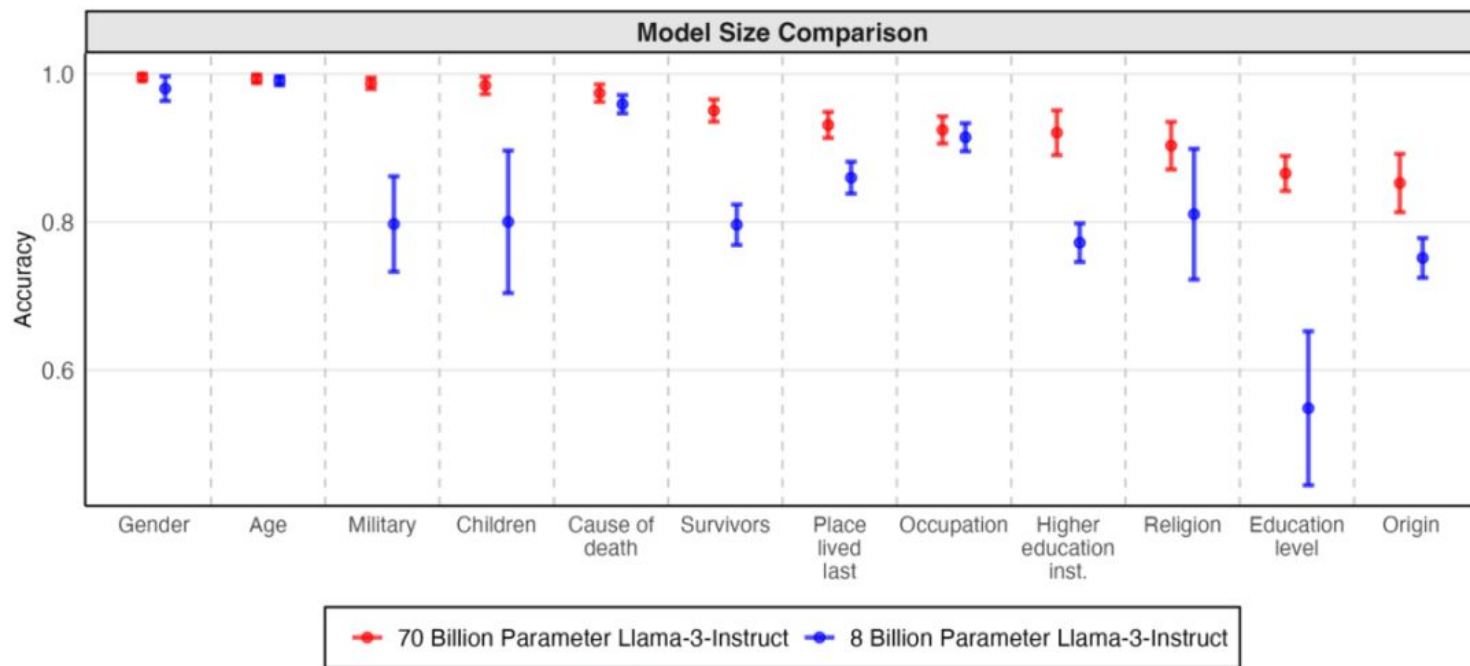
His death was confirmed in a statement by Prime Minister Giorgia Meloni, with whom he was a coalition partner in the current Italian government. No cause of death was given, but he was hospitalized last week as part of his treatment for chronic leukemia and other ailments.

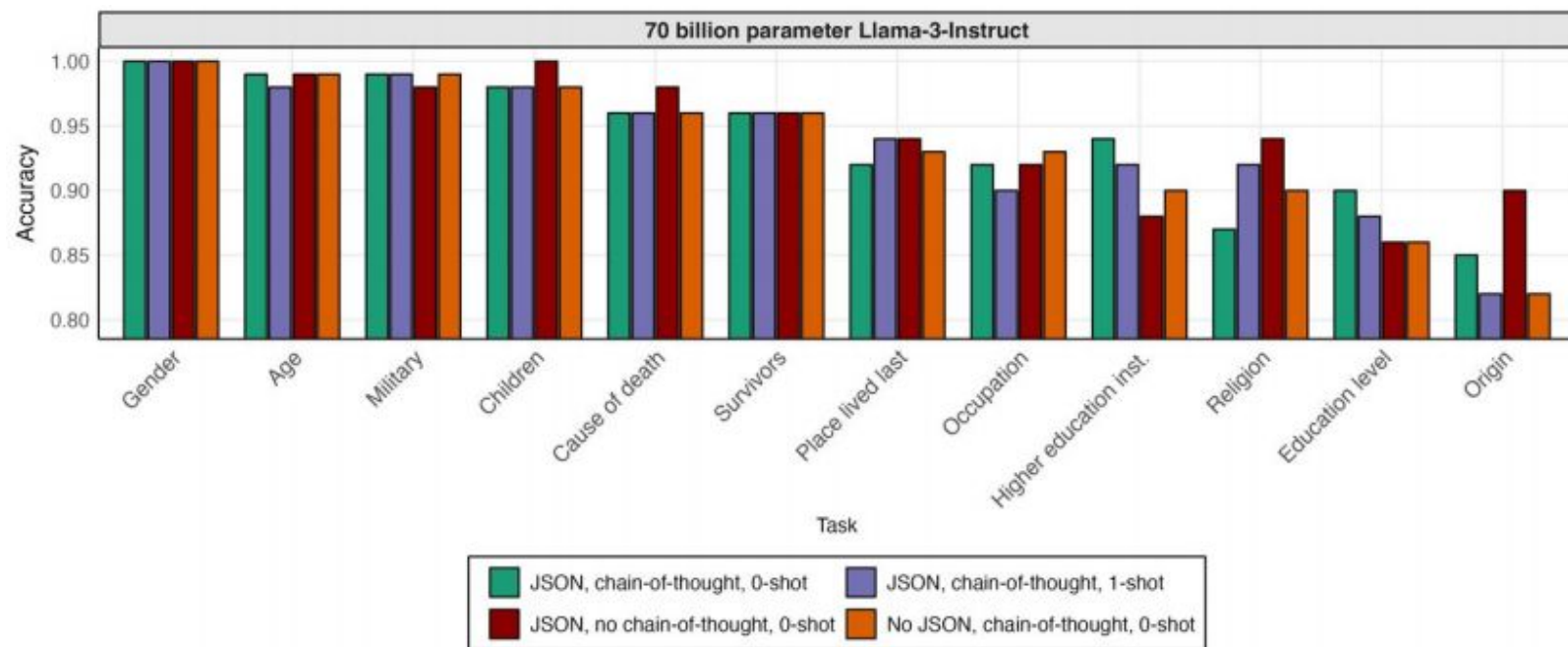
To Italians, Mr. [Doe](#) was constant entertainment — both comic and tragic, with more than a touch of off-color material — until they booed him off the stage. But

...

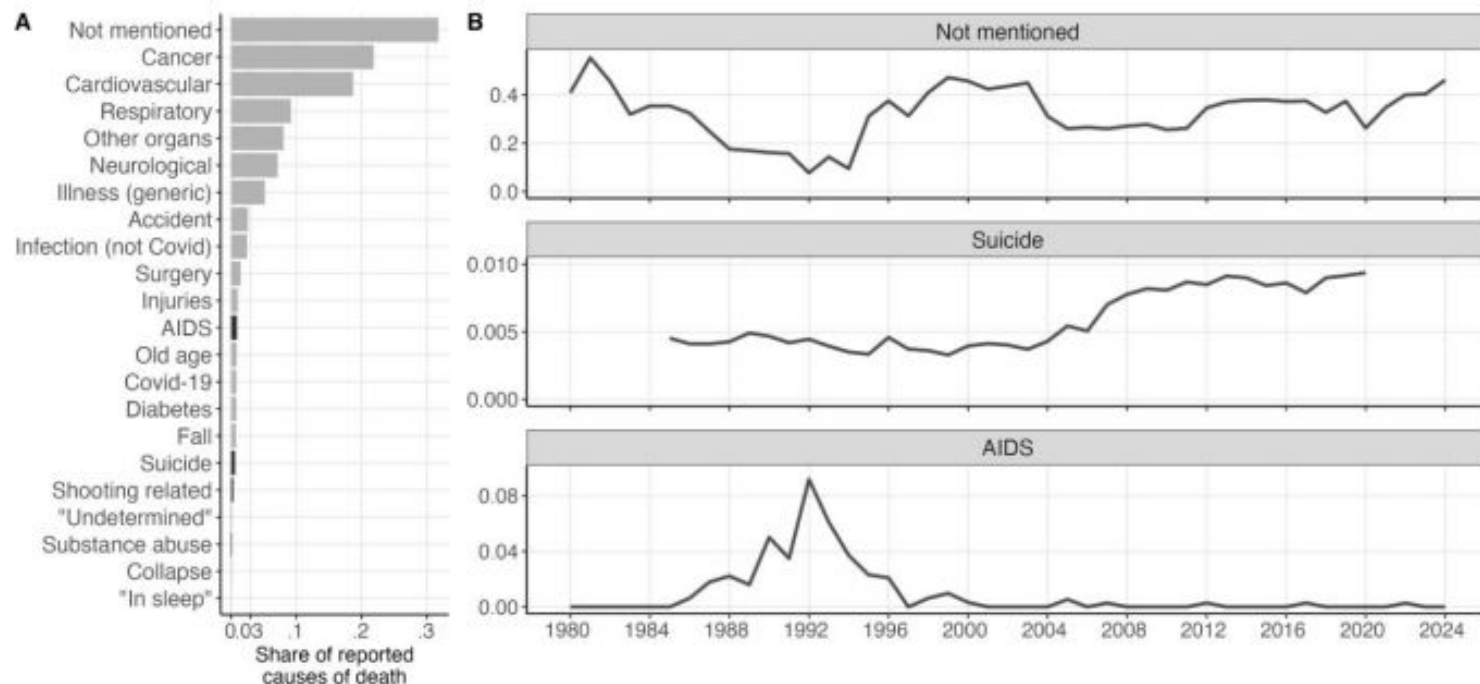
Results

Results





Results



Discussion

- Accessibility/Flexibility
- Reproducibility
- Costs

Conclusion

From codebook to promptbook

Does the obituary mention that the person served in the army? If so, put "yes." Otherwise, put "not mentioned".

If a person served in a foreign (that is non-US) army, also put "yes." Advisory roles do not count as having served in the army.

Below I will provide an obituary of a deceased person.

Does the obituary mention that the person served in the military? If so, respond with [yes]. Otherwise, respond with [not mentioned].

If a person served in a foreign (that is non-US) military, also respond with [yes.] Advisory roles do not count as having served in the military.

Your answer should be formatted like this: [yes/not mentioned]

Here is the obituary date, title, and text: {insert obituary text}

Conclusion

Promptbooks for social sciences

SC. I]	KING LEAR	99	Standby Sound Q3
	And from some knowledge and assurance offer This office to you.		
Gent.	I will talk further with you.		
Kent.	No, do not.		
	For confirmation that I am much more Than my out-wall, open this purse, and take What it contains. If you shall see Cordelia,—	45	
	As fear not but you shall—show her this ring, And she will tell you who that fellow is That yet you do not know. Fie on this storm! I will go seek the King.	50	Sound Q3 GO [Thunder]
Gent.	Give me your hand. Have you no more to say?		
Kent.	Few words, but, to effect, more than all yet; That, when we have found the King, in which your pain		Standby LX Qs 10-14 and Sound Qs 4-6
	That way, I'll this, he that first lights on him Holla the other. <i>[Exeunt severally.]</i>	55	LX Q 10 GO [Dim scene change]
SCENE II.— <i>[Another part of the Heath.] Storm still.</i>			
<i>Enter LEAR and Fool.</i>			<i>When the stage is clear</i>
Lear.	Blow, winds, and crack your cheeks! rage! blow! You cataracts and hurricanoes, spout		LX Q 11 GO [Heath state – dim]
	Till you have drench'd our steeples, drown'd the cocks! You sulph'rous and thought-executing fires, Vaunt-couriers of oak-cleaving thunderbolts,	5	LXQ 12 GO [Lightning flash]
	Singe my white head! And thou, all-shaking thunder, Strike flat the thick rotundity o'th'world! Crack Nature's moulds, all germens spill at once That makes ingrateful man!		Follow on Sound Q4 GO [Thunder rumble]
Fool.	O Nuncle, court holy-water in a dry house is better than this rain-water out o'door. Good Nuncle, in, ask thy daughters blessing; here's a night pities neither wise men nor Fools.	10	LXQ13 GO [Lightning flash]
Lear.	Rumble thy bellyful! Spit, fire! spout, rain! Nor rain, wind, thunder, fire, are my daughters: I tax you not, you elements, with unkindness; I never gave you kingdom, call'd you children, You owe me no subscription: then let fall Your horrible pleasure; here I stand, your slave,	15	Follow on Sound Q5 GO [Thunder crack]
			LXQ14 GO [Lightning flash]
			Follow on Sound Q6 GO [Thunder rumble]