Using LLMs to Support Survey Coding

April 17, 2024

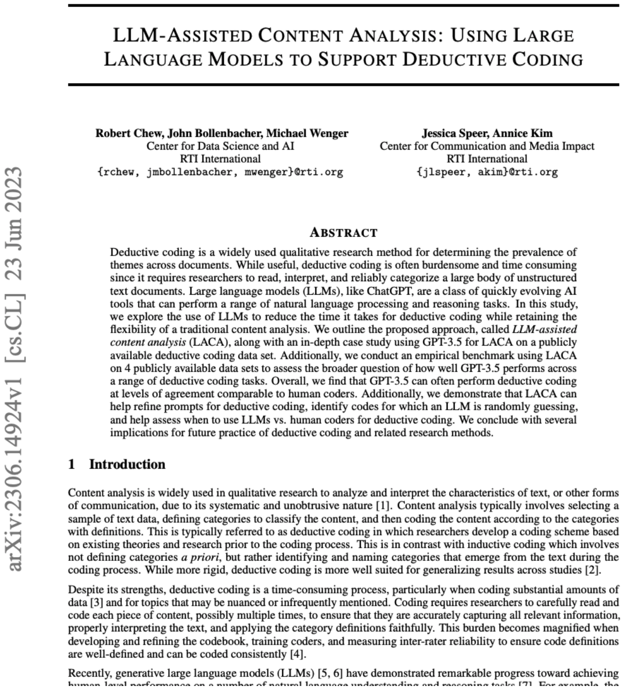
Robert Chew
John Bollenbacher
Michael Wenger
Jess Speer
Annice Kim

RTI INTERNATIONAL

Want to Learn More?

Check out the pre-print!

https://arxiv.org/abs/2306.14924



LLM-ASSISTED CONTENT ANALYSIS: USING LARGE LANGUAGE MODELS TO SUPPORT DEDUCTIVE CODING

Robert Chew, John Bollenbacher, Michael Wenger
Center for Data Science and AI
RTI International
{rchew, jmbollenbacher, mwenger}@rti.org

Jessica Speer, Annice Kim
Center for Communication and Media Impact
RTI International
{jlspeer, akim}@rti.org

ABSTRACT

Deductive coding is a widely used qualitative research method for determining the prevalence of themes across documents. While useful, deductive coding is often burdensome and time consuming since it requires researchers to read, interpret, and reliably categorize a large body of unstructured text documents. Large language models (LLMs), like ChatGPT, are a class of quickly evolving AI tools that can perform a range of natural language processing and reasoning tasks. In this study, we explore the use of LLMs to reduce the time it takes for deductive coding while retaining the flexibility of a traditional content analysis. We outline the proposed approach, called *LLM-assisted content analysis* (LACA), along with an in-depth case study using GPT-3.5 for LACA on a publicly available deductive coding data set. Additionally, we conduct an empirical benchmark using LACA on 4 publicly available data sets to assess the broader question of how well GPT-3.5 performs across a range of deductive coding tasks. Overall, we find that GPT-3.5 can often perform deductive coding at levels of agreement comparable to human coders. Additionally, we demonstrate that LACA can help refine prompts for deductive coding, identify codes for which an LLM is randomly guessing, and help assess when to use LLMs vs. human coders for deductive coding. We conclude with several implications for future practice of deductive coding and related research methods.

1 Introduction

Content analysis is widely used in qualitative research to analyze and interpret the characteristics of text, or other forms of communication, due to its systematic and unobtrusive nature [1]. Content analysis typically involves selecting a sample of text data, defining categories to classify the content, and then coding the content according to the categories with definitions. This is typically referred to as deductive coding in which researchers develop a coding scheme based on existing theories and research prior to the coding process. This is in contrast with inductive coding which involves not defining categories *a priori*, but rather identifying and naming categories that emerge from the text during the coding process. While more rigid, deductive coding is more well suited for generalizing results across studies [2].

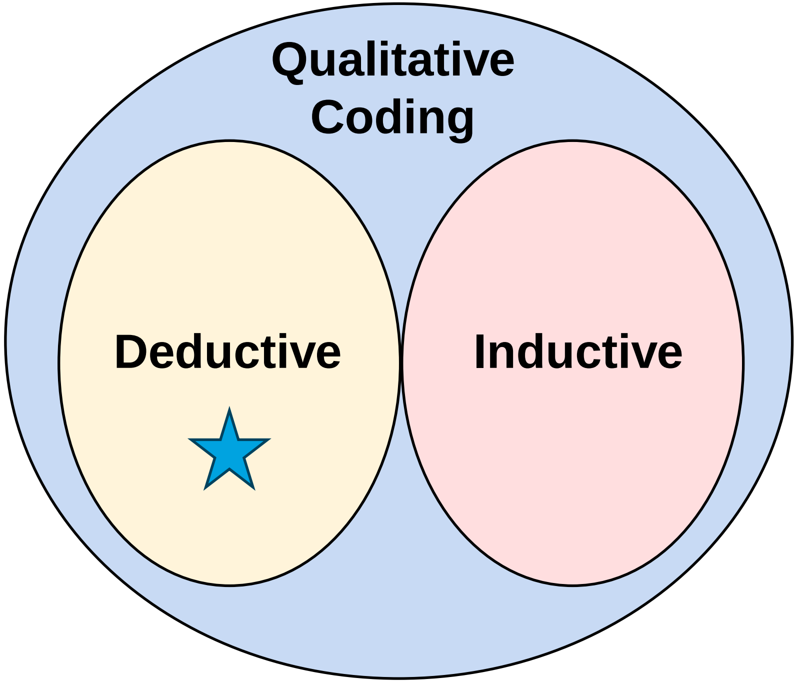
Despite its strengths, deductive coding is a time-consuming process, particularly when coding substantial amounts of data [3] and for topics that may be nuanced or infrequently mentioned. Coding requires researchers to carefully read and code each piece of content, possibly multiple times, to ensure that they are accurately capturing all relevant information, properly interpreting the text, and applying the category definitions faithfully. This burden becomes magnified when developing and refining the codebook, training coders, and measuring inter-rater reliability to ensure code definitions are well-defined and can be coded consistently [4].

Recently, generative large language models (LLMs) [5, 6] have demonstrated remarkable progress toward achieving

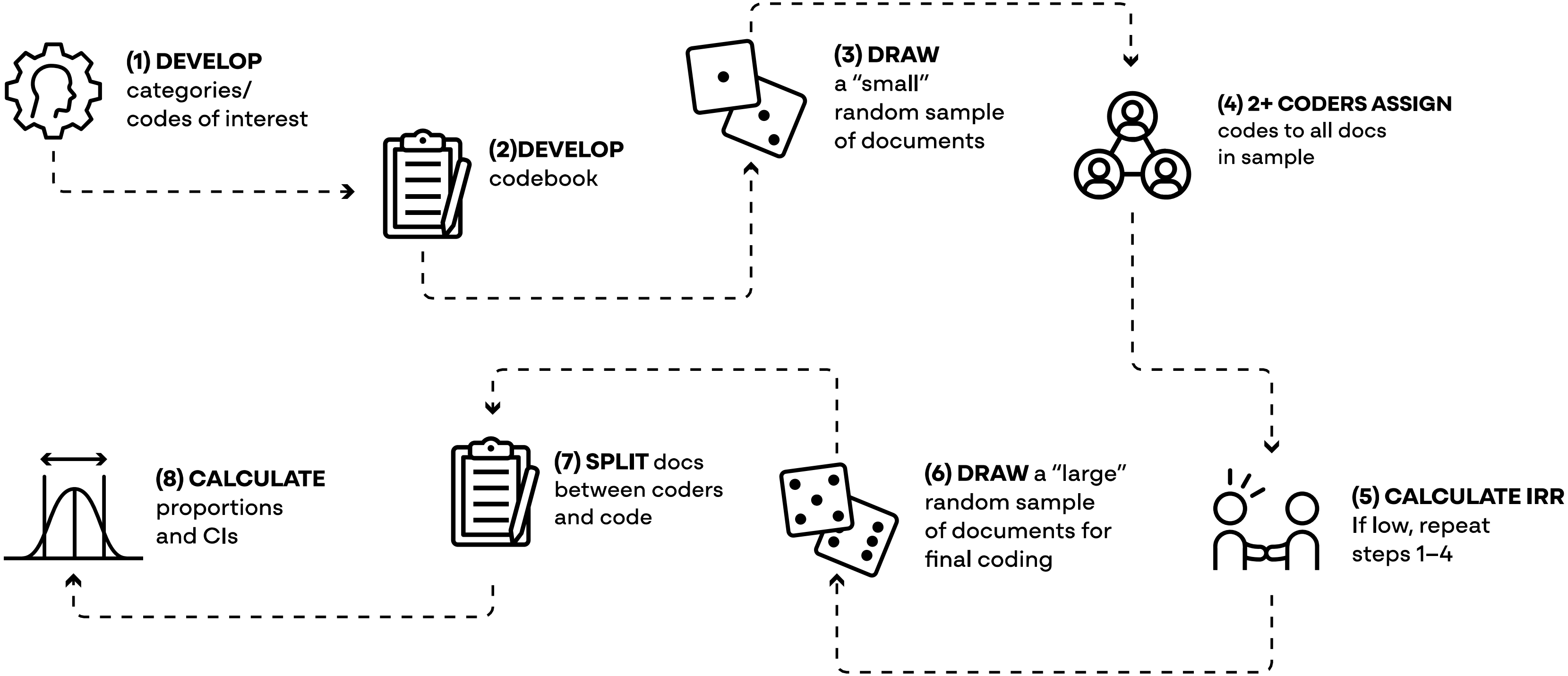
Background

Survey coding is an essential operation for analyzing text data.

However, coding can be **slow, expensive, and error prone**.



Deductive Coding



Neuendorf (2017). *The content analysis guidebook*.

Research Question

How well does ChatGPT perform deductive coding compared to humans?

1. Inter-rater Reliability (IRR)
2. Coding Time

Publicly Available Datasets

Data Set	Doc Type	Mutually Exclusive	# Codes	# Docs	Notes
Trump Tweets	Tweets	No	13	2,083	Codebook written informally with short descriptions
Contrarian Claims	Blog Posts	Yes	28	2,904	Mutually exclusive, hierarchical code set. Codes nuanced and may have definitions with conceptual overlap
BBC News	News Articles	Yes	5	2,225	No formal codebook, only class names (e.g., business)
Ukraine Water Problems	Water Quality Reports	No	5	100	Brief codebook, but technically complex classes

Current case studies discussed in this webinar are exploratory only and should not be used for any other purpose.

Example Prompt

```
You are a qualitative coder who is annotating news stories. To code this text, do the following:

- First, read the codebook and the text.
- Next, decide which code is most applicable and explain your reasoning for the coding decision.
- Finally, print the most applicable code and your reason for the coding decision.

Use the following format:

Codebook:
---
{codebook here}
---

Text:
---
{text here}
---

Code:
---
business, entertainment, politics, sport, or tech
---

Code:
```

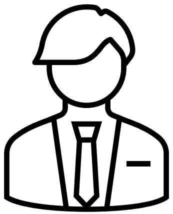
Coding instructions

Text document

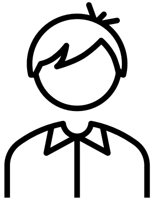
Coding decision and reason for decision

Human-Human vs. Human Model Agreement

Human-Human Agreement



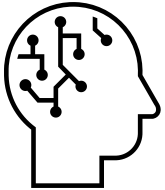
Published
Data



Our Coded
Data

Human-Model Agreement

Published
Data



ChatGPT
Predictions

Agreement Metric

Gwet’s AC1

Results: Reliability

GPT-3.5 often coded at levels of agreement **comparable to humans**

Table 7: Summary Benchmark Results Across Data Sets

Dataset	Code	Gwet's AC1 Human-Human	Gwet's AC1 Human-Model	Tests of Randomness (p-value)
Trump Tweets	HSTG	0.96	0.18	0.19
Trump Tweets	ATSN	1.00	0.58	0.92
Trump Tweets	CRIT	0.73	0.76	0.00
Trump Tweets	MEDI	1.00	0.96	0.00
Trump Tweets	FAMY	0.97	0.96	0.00
Trump Tweets	PLCE	1.00	0.98	0.00
Trump Tweets	MAGA	0.99	0.98	0.00
Trump Tweets	CAPT	0.93	0.36	0.76
Trump Tweets	INDV	0.79	0.50	0.19
Trump Tweets	MARG	0.97	0.94	0.00
Trump Tweets	INTN	0.86	0.81	0.00
Trump Tweets	PRTY	0.81	0.76	0.00
Trump Tweets	IMMG	0.99	0.97	0.00
Ukraine Water	env_problems	0.23	0.64	0.62
Ukraine Water	pollution	0.59	0.55	0.62
Ukraine Water	treatment	0.84	0.88	0.00
Ukraine Water	climate	0.97	0.87	0.00
Ukraine Water	biomonitoring	0.51	0.86	0.00
BBC News	All	0.76	0.85	0.00
Contrarian Claims	All	0.65	0.59	0.00

Results: Reliability

Our method was **able to predict when GPT-3.5 fails** at coding (p-values)

Table 7: Summary Benchmark Results Across Data Sets

Dataset	Code	Gwet's AC1		Tests of Randomness (p-value)
		Human-Human	Human-Model	
Trump Tweets	HSTG	0.96	0.18	0.19
Trump Tweets	ATSN	1.00	0.58	0.92
Trump Tweets	CRIT	0.73	0.76	0.00
Trump Tweets	MEDI	1.00	0.96	0.00
Trump Tweets	FAMY	0.97	0.96	0.00
Trump Tweets	PLCE	1.00	0.98	0.00
Trump Tweets	MAGA	0.99	0.98	0.00
Trump Tweets	CAPT	0.93	0.36	0.76
Trump Tweets	INDV	0.79	0.50	0.19
Trump Tweets	MARG	0.97	0.94	0.00
Trump Tweets	INTN	0.86	0.81	0.00
Trump Tweets	PRTY	0.81	0.76	0.00
Trump Tweets	IMMG	0.99	0.97	0.00
Ukraine Water	env_problems	0.23	0.64	0.62
Ukraine Water	pollution	0.59	0.55	0.62
Ukraine Water	treatment	0.84	0.88	0.00
Ukraine Water	climate	0.97	0.87	0.00
Ukraine Water	biomonitoring	0.51	0.86	0.00
BBC News	All	0.76	0.85	0.00
Contrarian Claims	All	0.65	0.59	0.00

New Results! GPT4

Using a **better model (GPT-4)** with same prompts **improved IRR** for many categories which GPT-3.5 struggled.

Dataset	Code	Gwet's AC1		
		Orginal-Replicated	Original-GPT3.5	Original-GPT4
Trump Tweets	HSTG	0.96	0.18	**0.97**
Trump Tweets	ATSN	**1**	0.58	**1**
Trump Tweets	CRIT	0.73	0.76	**0.89**
Trump Tweets	MEDI	**1**	0.96	**1**
Trump Tweets	FAMY	0.97	0.96	**0.99**
Trump Tweets	PLCE	**1**	0.98	0.99
Trump Tweets	MAGA	**0.99**	0.98	**0.99**
Trump Tweets	CAPT	**0.93**	0.36	0.72
Trump Tweets	INDV	0.79	0.5	**0.87**
Trump Tweets	MARG	**0.97**	0.94	0.95
Trump Tweets	INTN	0.86	0.81	**0.95**
Trump Tweets	PRTY	0.81	0.76	**0.83**
Trump Tweets	IMMG	**0.99**	0.97	**0.99**
Ukraine Water	env_problems	0.23	0.64	**0.7**
Ukraine Water	pollution	0.59	0.55	**0.62**
Ukraine Water	treatment	0.84	**0.88**	0.83
Ukraine Water	climate	**0.97**	0.87	0.86
Ukraine Water	biomonitoring	0.51	0.86	**0.92**
BBC	All	0.76	0.85	**0.99**
Contrarian Claims	All	**0.65**	0.59	0.52

Results: Coding Time

GPT-3.5 substantially faster than humans, especially for long docs with many categories

Table 8: Coding Time per Document

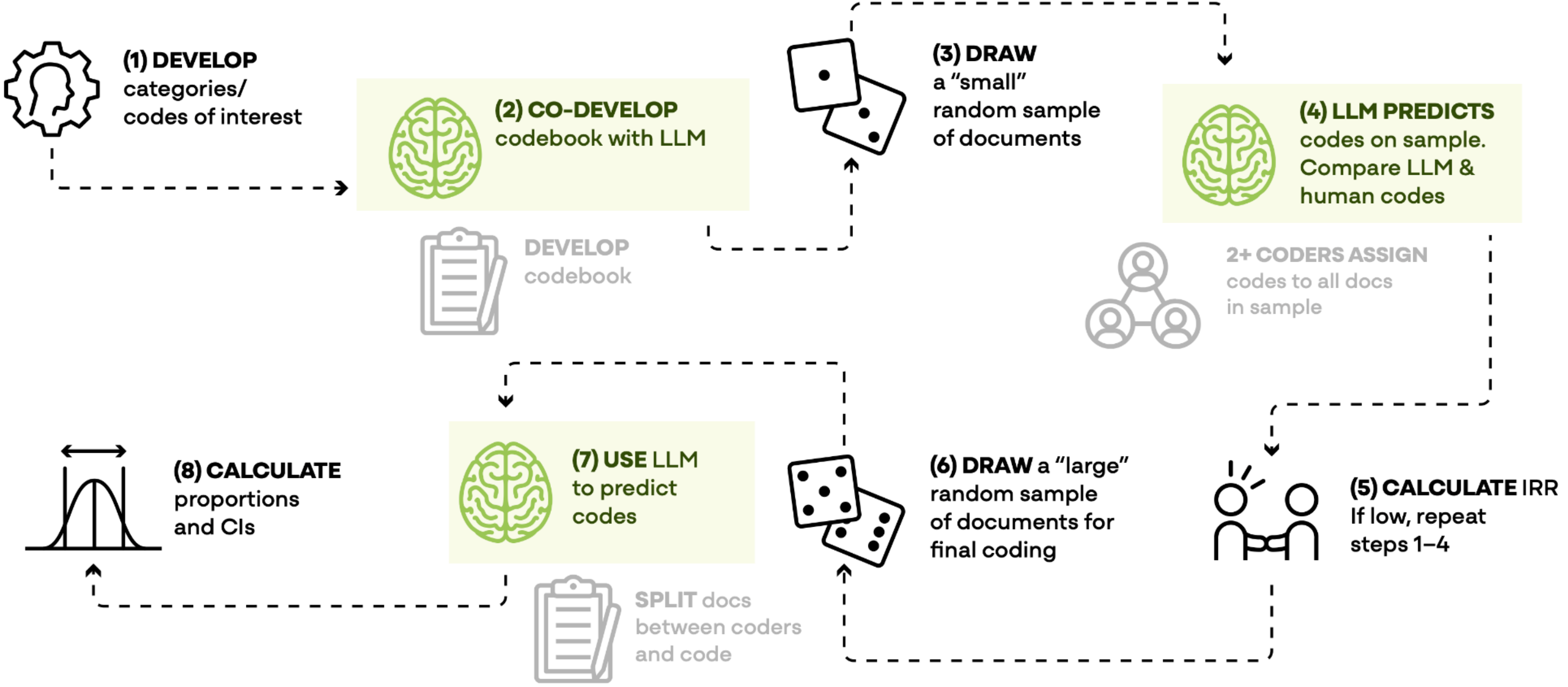
Dataset	**Coding Time (seconds / document)**	
	Human Coder	**LLM Coder**
Trump Tweets	72	52
Ukraine Water	108	16
BBC News	72	4
Contrarian Claims	144	4

36x faster!

Discussion

- Based on coding time and reliability, LLMs appears promising for deductive coding.
- Use of LLMs for deductive coding will likely require different types of reporting and documentation for reproducibility and critique.
- **We do not consider LLMs as a replacement for qualitative coders,** but rather, a tool to help accelerate the latter stages of deductive coding that tend to be more manually taxing and repetitive.

LLM-Assisted Content Analysis (LACA)



Limitations

- To match the original data sets, we forced ChatGPT to choose Yes / No or a single code (no “I don’t know” option).
- We only assessed ChatGPT and not a wider variety of Large Language Models (LLMs).
- Implementing LACA would mean researchers read less documents, which may limit new theory development and discovering themes not proposed by the research team *a priori*.

Questions?

Rob Chew | rchew@rti.org