

The logo for Carnegie Mellon University, featuring the text "Carnegie Mellon University" in a white serif font. The text is set against a dark blue background with a grid of colorful lines (red, green, yellow, blue) that create a sense of depth and movement.

Carnegie
Mellon
University

Large Language Models and Empirical Legal Studies

Artificial Intelligence in Research and Applications Seminar (AIRA)
March 5, 2026

Jaromir Savelka

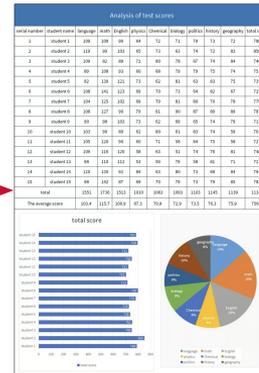
A decorative graphic in the bottom right corner of the slide, consisting of a series of overlapping, curved lines in various colors (red, green, blue, yellow) that create a sense of motion and depth.

Overview

- Developments in natural language processing
- In-context learning
- Examples of empirical legal studies
 - Predicting Legal Factors from Court Opinions
 - Inductive Analysis of Criminal Court Opinions on Thefts

Empirical Legal Studies (ELS)

- Empirical legal studies (ELS) is an approach to the study of law through empirical methods typical of economics, psychology, and sociology.
- Law is a heavily text-based discipline → ELS frequently focuses on text analytic methods
 - **deductive coding** - applying a fixed set of codes to a dataset
 - **inductive coding** - discovery of the codes from the data and their application



Typical ELS Project (text analysis focused)

Deductive Coding (*hired annotators case*)

- Specify task (annotation guidelines)
- Pilot study (update guidelines)
- Hire and train human annotators
- Annotate the data set
 - Monitor performance
 - Re-train annotators
 - Quality assurance (QA)
- Analyze results

Inductive Coding (*single researcher case*)

- Familiarization with data
- Generating initial codes
- Searching for themes
- Reviewing themes
- Defining and naming themes
- Projecting themes onto the data
- Analyze results



Developments in Natural Language Processing

Artificial Intelligence: Definition

AI is a family of technologies and scientific field that enables/studies: (i) **automation**, (ii) **acceleration**, and (iii) **extreme scalability** of human (i) **perception**, (ii) **reasoning**, and (iii) **decision making**.

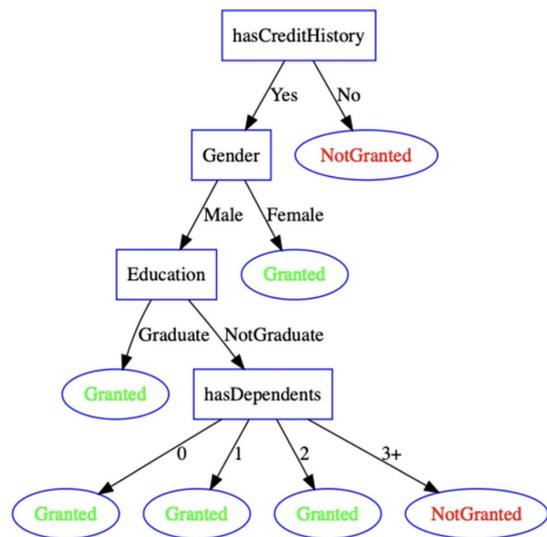
- **Narrow AI** - focused, specialised intelligent behavior approximating human performance.
- **Extended AI** - outperforming human in specialised intelligent behavior in scale, speed, precision or reliability and robustness.
- **General AI** - holistic intelligent behavior in general sense (creativity, intensionality, transfer learning capability)
- **Super AI** - outperforming humans in the general AI in scale, speed, etc.

Artificial Intelligence: Evolving Paradigms

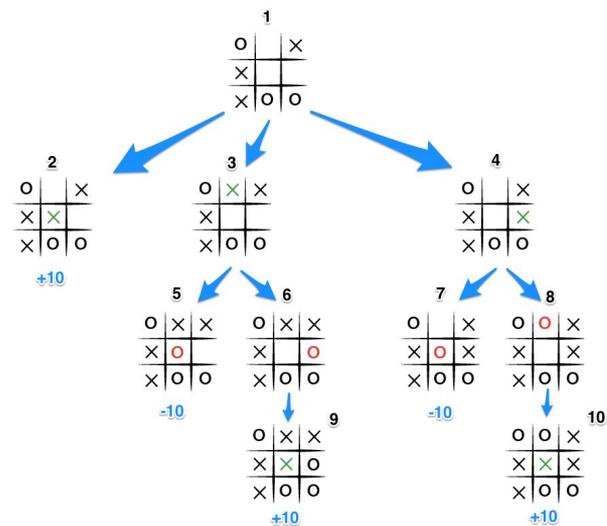
- **Symbolic AI** (GOFAI) - AI solvers, automated reasoning, planning & optimisation
- **Statistical AI** - machine learning (e.g., neural networks, deep learning)
 - Probabilistic data-driven approaches (since late eighties)
 - Fully **supervised** machine learning (since nineties)
 - Traditional ML - heavy reliance on feature engineering.
 - Deep neural networks - architecture engineering and need for (extremely) large datasets.
 - **Pre-train** -> **Fine-tune** paradigm (since 2017)
 - A model with a fixed architecture is pre-trained to predict the probability of observed data.
 - A model is then adapted to different downstream tasks.
 - Pre-train -> (Fine-tune)? -> **Adapt** paradigm (since 2021)
 - Prompt engineering, in-context learning
 - Instruction following, reinforcement learning (e.g., RLHF)
- **Distributed AI** - **agentic** systems, decentralised control, game theory (recently)

Artificial Intelligence: Evolving Paradigms

- **Symbolic AI** (GOFAI) - AI solvers, automated reasoning, planning & optimisation



$P \rightarrow P$
 $Q \rightarrow (P \rightarrow Q)$
 $P \rightarrow ((P \rightarrow Q) \rightarrow Q)$
 $(P \rightarrow Q) \rightarrow ((Q \rightarrow R) \rightarrow [P \rightarrow R])$
 $(Q \rightarrow R) \rightarrow ((P \rightarrow Q) \rightarrow [P \rightarrow R])$
 $(P \rightarrow [Q \rightarrow R]) \rightarrow ((P \rightarrow Q) \rightarrow [P \rightarrow R])$
 $((P \rightarrow Q) \rightarrow [P \rightarrow R]) \rightarrow (P \rightarrow [Q \rightarrow R])$
 $(P \rightarrow [Q \rightarrow R]) \rightarrow (Q \rightarrow [P \rightarrow R])$
 $(P \rightarrow [P \rightarrow Q]) \rightarrow (P \rightarrow Q)$
 $((P \rightarrow Q) \rightarrow Q) \rightarrow ((Q \rightarrow P) \rightarrow P)$
 $\sim P \rightarrow P$
 $P \rightarrow \sim P$
 $(P \rightarrow Q) \rightarrow (\sim Q \rightarrow \sim P)$
 $(P \rightarrow \sim Q) \rightarrow (Q \rightarrow \sim P)$
 $(\sim P \rightarrow Q) \rightarrow (\sim Q \rightarrow P)$
 $(\sim P \rightarrow \sim Q) \rightarrow (Q \rightarrow P)$



Typical ELS Project (text analysis focused)

Deductive Coding (*hired annotators case*)

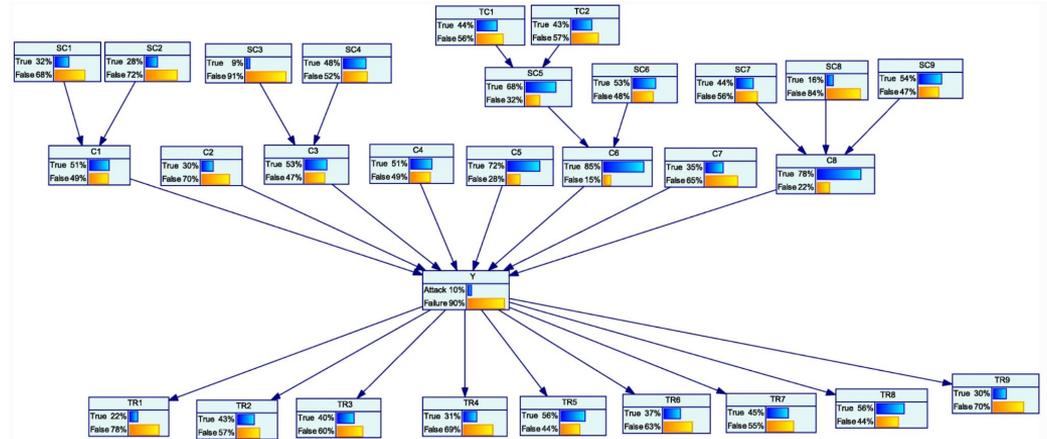
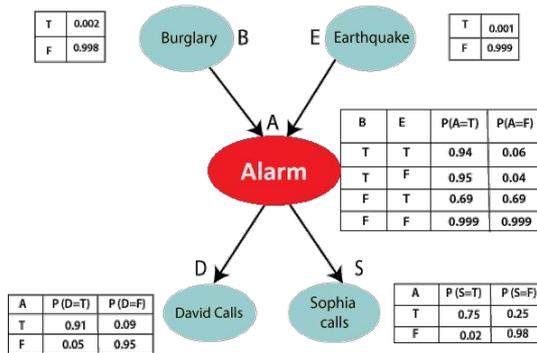
- Specify task (annotation guidelines)
- Pilot study (update guidelines)
- **Annotate the data set**
 - Formulate explicit rules
 - Apply rules for annotation
 - Quality assurance (QA)
- Analyze results

Inductive Coding (*single researcher case*)

- Familiarization with data
- Generating initial codes
- Searching for themes
- Reviewing themes
- Defining and naming themes
- **Projecting themes onto the data**
- Analyze results

Artificial Intelligence: Evolving Paradigms

- **Symbolic AI** (GOFAI) - AI solvers, automated reasoning, planning & optimisation
- **Statistical AI** - machine learning (e.g., neural networks, deep learning)
 - Probabilistic data-driven approaches (since late eighties)



Typical ELS Project (text analysis focused)

Deductive Coding (*hired annotators case*)

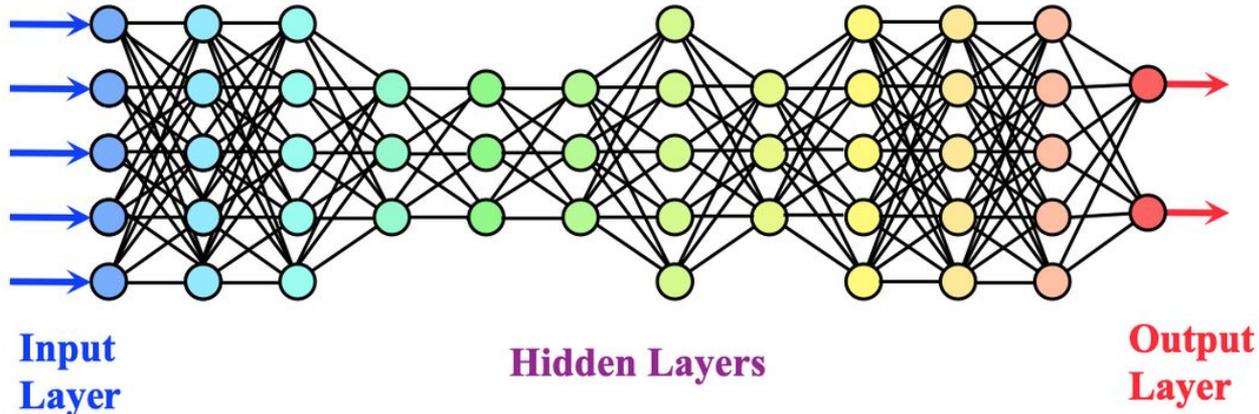
- Specify task (annotation guidelines)
- Pilot study (update guidelines)
- Hire and train human annotators
- Annotate the data set
 - Monitor performance
 - Re-train annotators
 - Quality assurance (QA)
- **Use machine learning to annotate additional data (train, evaluate)**
- Analyze results

Inductive Coding (*single researcher case*)

- Familiarization with data
- Generating initial codes
- **Searching for themes → Clustering?**
- Reviewing themes
- Defining and naming themes
- Projecting themes onto the data
- **Use machine learning to project themes onto additional data**
- Analyze results

Artificial Intelligence: Evolving Paradigms

- **Symbolic AI** (GOFAI) - AI solvers, automated reasoning, planning & optimisation
- **Statistical AI** - machine learning (e.g., neural networks, deep learning)
 - Probabilistic data-driven approaches (since late eighties)
 - Fully **supervised** machine learning (since nineties)
 - Traditional ML - heavy reliance on feature engineering.
 - Deep neural networks - architecture engineering and need for (extremely) large datasets.



Artificial Intelligence: Evolving Paradigms

- **Symbolic AI** (GOFAI) - AI solvers, automated reasoning, planning & optimisation
- **Statistical AI** - machine learning (e.g., neural networks, deep learning)
 - Probabilistic data-driven approaches (since late eighties)
 - Fully **supervised** machine learning (since nineties)
 - **Traditional ML** - heavy reliance on feature engineering.
 - **Deep neural networks** - architecture engineering and need for (extremely) large datasets.
 - **Pre-train** -> **Fine-tune** paradigm (since 2017)
 - A model with a fixed architecture is pre-trained to predict the probability of observed data.
 - A model is then adapted to different downstream tasks.

Artificial Intelligence: Evolving Paradigms

- **Symbolic AI** (GOFAI) - AI solvers, automated reasoning, planning & optimisation
- **Statistical AI** - machine learning (e.g., neural networks, deep learning)
 - Probabilistic data-driven approaches (since late eighties)
 - Fully **supervised** machine learning (since nineties)
 - Traditional ML - heavy reliance on feature engineering.
 - Deep neural networks - architecture engineering and need for (extremely) large datasets.
 - **Pre-train** -> **Fine-tune** paradigm (since 2017)
 - A model with a fixed architecture is pre-trained to predict the probability of observed data.
 - A model is then adapted to different downstream tasks.
 - Pre-train -> (Fine-tune)? -> **Adapt** paradigm (since 2021)
 - Prompt (context) engineering, in-context learning
 - Instruction following, reinforcement learning (e.g., RLHF)

Typical ELS Project (text analysis focused)

Deductive Coding (*hired annotators case*)

- Specify task (annotation guidelines)
- Pilot study (update guidelines)
- **Use machine learning to annotate data**
- Hire and train human annotators
- **Annotate data set for evaluation**
- Analyze results

Inductive Coding (*single researcher case*)

- Familiarization with data
- **Generating initial codes**
- **Use machine learning to define themes**
- **Use machine learning to project themes onto additional data**
- Evaluate?
- Analyze results

Typical ELS Project (text analysis focused)

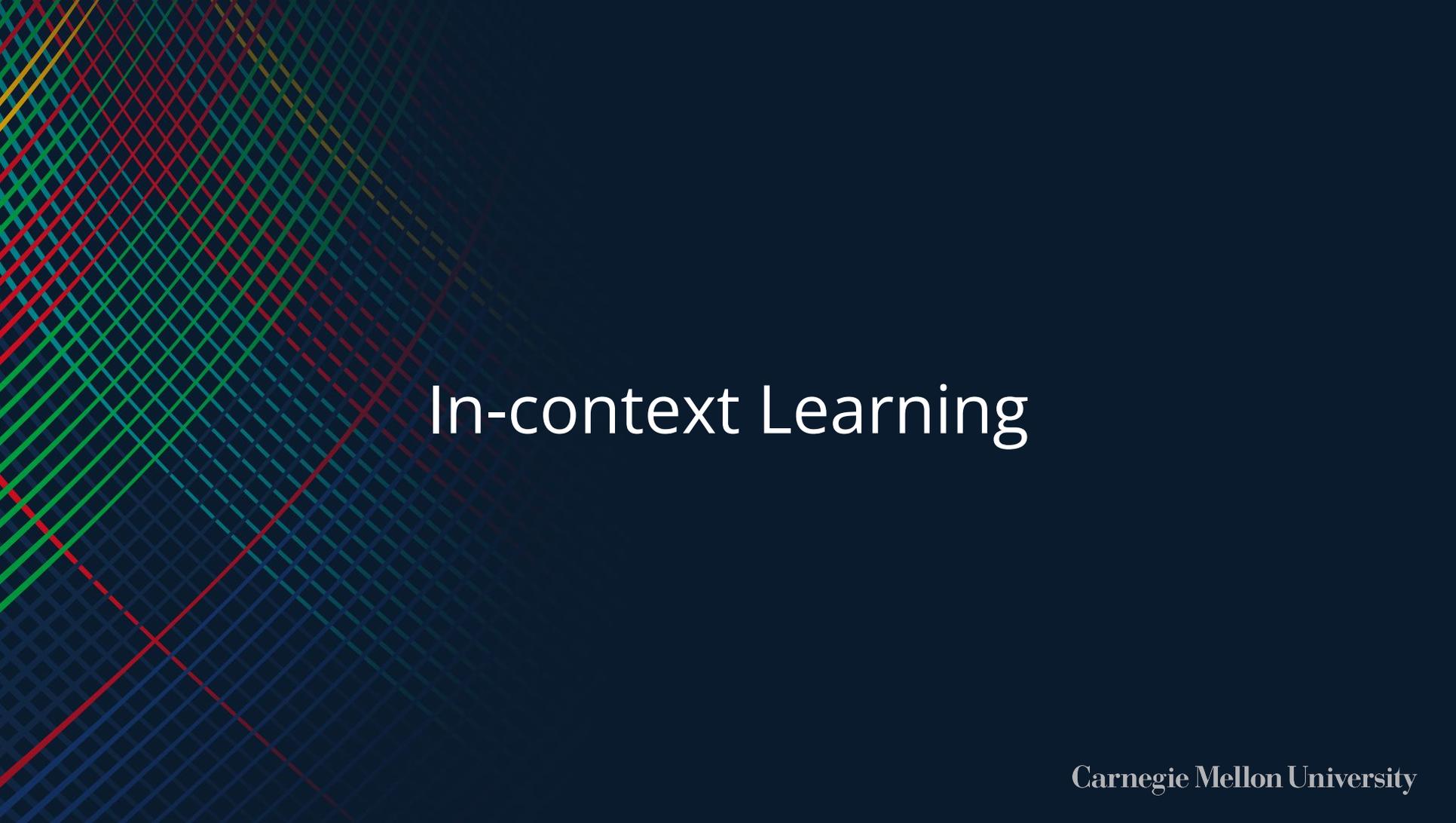
Deductive Coding (*hired annotators case*)

- AI performs the whole task end-to-end

Inductive Coding (*single researcher case*)

- AI performs the whole task end-to-end





In-context Learning

In-context Learning

All the information necessary to get the LM to do the task is included as part of the textual prompt input to the LM.

LLM zero-shot learning: A prompt that contains instructions for the task, but no actual examples of the task being performed.

LLM few-shot learning: A prompt that contains several examples of the task being performed and optionally the instructions.

Example Prompt: Text Summarization

JS

Antibiotics are a type of medication used to treat bacterial infections. They work by either killing the bacteria or preventing them from reproducing, allowing the body's immune system to fight off the infection. Antibiotics are usually taken orally in the form of pills, capsules, or liquid solutions, or sometimes administered intravenously. They are not effective against viral infections, and using them inappropriately can lead to antibiotic resistance.

Explain the above in one sentence:

Antibiotics are medications that treat bacterial infections by killing bacteria or stopping their reproduction, but they don't work against viruses and improper use can cause antibiotic resistance.



   [Retry](#) 

Claude can make mistakes. Please double-check responses.

Example Prompt: Information Extraction

JS Author-contribution statements and acknowledgements in research papers should state clearly and specifically whether, and to what extent, the authors used AI technologies such as ChatGPT in the preparation of their manuscript and analysis. They should also indicate which LLMs were used. This will alert editors and reviewers to scrutinize manuscripts more carefully for potential biases, inaccuracies and improper source crediting. Likewise, scientific journals should be transparent about their use of LLMs, for example when selecting submitted manuscripts. Mention the large language model based product mentioned in the paragraph above:

The large language model-based product mentioned in the paragraph above is **ChatGPT**.



   Retry ▾

Claude can make mistakes. Please double-check responses.

Example Prompt: Question Answering

JS Answer the question based on the context below. Keep the answer short and concise. Respond "Unsure about answer" if not sure about the answer. Context: Teplizumab traces its roots to a New Jersey drug company called Ortho Pharmaceutical. There, scientists generated an early version of the antibody, dubbed OKT3. Originally sourced from mice, the molecule was able to bind to the surface of T cells and limit their cell-killing potential. In 1986, it was approved to help prevent organ rejection after kidney transplants, making it the first therapeutic antibody allowed for human use. Question: What was OKT3 originally sourced from?

OKT3 was originally sourced from mice.



Retry ▾

Claude can make mistakes. Please double-check responses.

Example Prompt: Text Classification

JS Classify the text into neutral, negative or positive. Text: I think the food was okay.
Sentiment:

Sentiment: **Neutral**

The text expresses a mild, non-committal opinion about the food. The word "okay" suggests the food was acceptable but not particularly impressive or disappointing—it's a middle-ground assessment without strong positive or negative feelings.



Retry ▾

Claude can make mistakes. Please double-check responses.

Example Prompt: Conversation

JS Hello, who are you?

Hello! I'm Claude, an AI assistant made by Anthropic. I'm here to help you with a wide range of tasks - whether that's answering questions, having conversations, helping with analysis or writing, coding, problem-solving, or creative projects.

Is there something specific I can help you with today?

JS Can you tell me about the creation of blackholes?

Black holes form through several different processes, all involving extreme gravitational collapse:

Stellar Black Holes

The most common formation process involves massive stars. When a star at least 20-25 times the mass of our Sun exhausts its nuclear fuel, it can no longer support itself against gravity. The core collapses catastrophically in a supernova explosion. If the remaining core mass is above about 3 solar masses (the Tolman-Oppenheimer-Volkoff limit), not even neutron degeneracy pressure can stop the collapse, and it continues shrinking until it forms a black hole.

Supermassive Black Holes

These giants, found at the centers of most galaxies (including our own Milky Way), contain

Example Prompt: Code Generation

JS Ask the user for their name and say "Hello" in Python.

I'd be happy to help you create a Python program that asks for a name and says hello!

Here's a simple way to do it:

```
python
name = input("What is your name? ")
print(f"Hello, {name}!")
```

This code:

1. Uses `input()` to ask the user for their name and stores it in the `name` variable
2. Uses `print()` with an f-string to display a personalized greeting

When you run this, it will prompt the user to enter their name, wait for them to type it and press Enter, then display "Hello, [their name]!"

Would you like me to create a file with this code, or would you like to see any variations (like a more elaborate greeting)?



   Retry ▾

Claude can make mistakes. Please double-check responses.

Why does ICL work?

- Instances of the task **exist in the pre-training data**.
 - Example: Translation data on the internet
- This means, the LLM already contains the knowledge it needs to do the task.
 - Additional instructions allow appropriate knowledge to be accessed.
 - Few-shot examples simply “teach” the LLM the template of the task.

“I’m not the cleverest man in the world, but like they say in French: **Je ne suis pas un imbecile [I’m not a fool]**.”

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: “**Mentez mentez, il en restera toujours quelque chose,**” which translates as, “**Lie lie and something will always remain.**”

“I hate the word ‘**perfume,**’” Burr says. ‘It’s somewhat better in French: ‘**parfum.**’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “**-Comment on fait pour aller de l’autre coté? -Quel autre coté?**”, which means “**- How do you get to the other side? - What side?**”.

If this sounds like a bit of a stretch, consider this question in French: **As-tu aller au cinéma?**, or **Did you go to the movies?**, which literally translates as Have-you to go to movies/theater?

“**Brevet Sans Garantie Du Gouvernement**”, translated to English: “**Patented without government warranty**”.

Table 1. Examples of naturally occurring demonstrations of English to French and French to English translation found throughout the WebText training set.

Prompts Are Not Created Equal

The cause for different behavior of various prompts remains unclear so far.

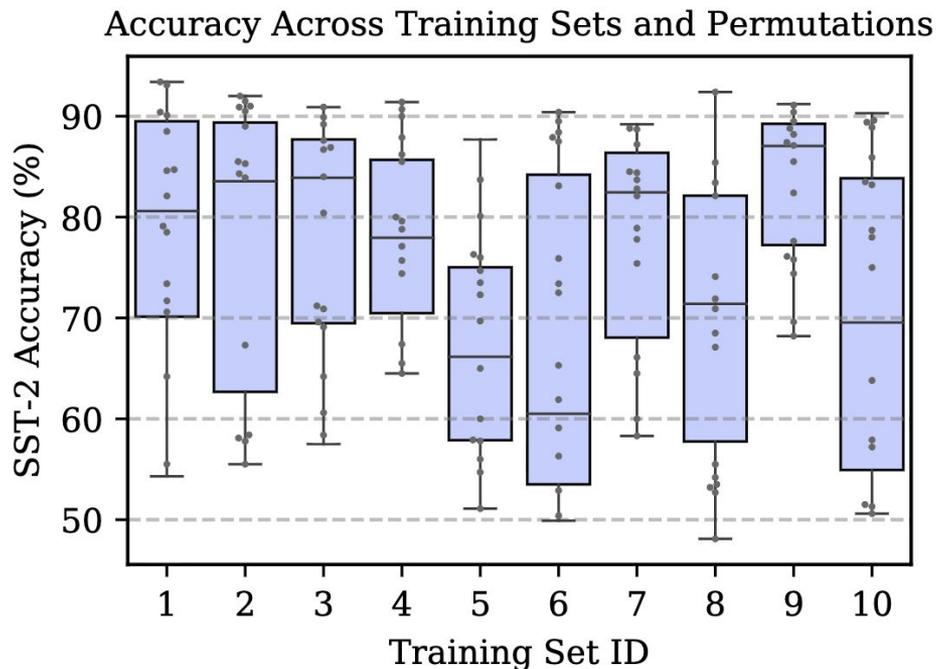
Proposal: *The more frequently a prompt appears in some variation in the data, the better it works for the task.*

| Prompt | Accuracy |
|--|-----------------|
| What is this piece of news regarding? | 40.9 |
| What is this article about? | 52.4 |
| What is the best way to describe this article? | 68.2 |
| What is the most accurate label for this news article? | 71.2 |

OPT 175B accuracies on AG News classification task

Exemplar Selection and Ordering

- Sentiment analysis task
- 4-shot with GPT-3 2.7B
- Randomly sampled exemplars
- Permutations



Correctness of the Labels

ICL performance drops only marginally when labels in the exemplars are random.

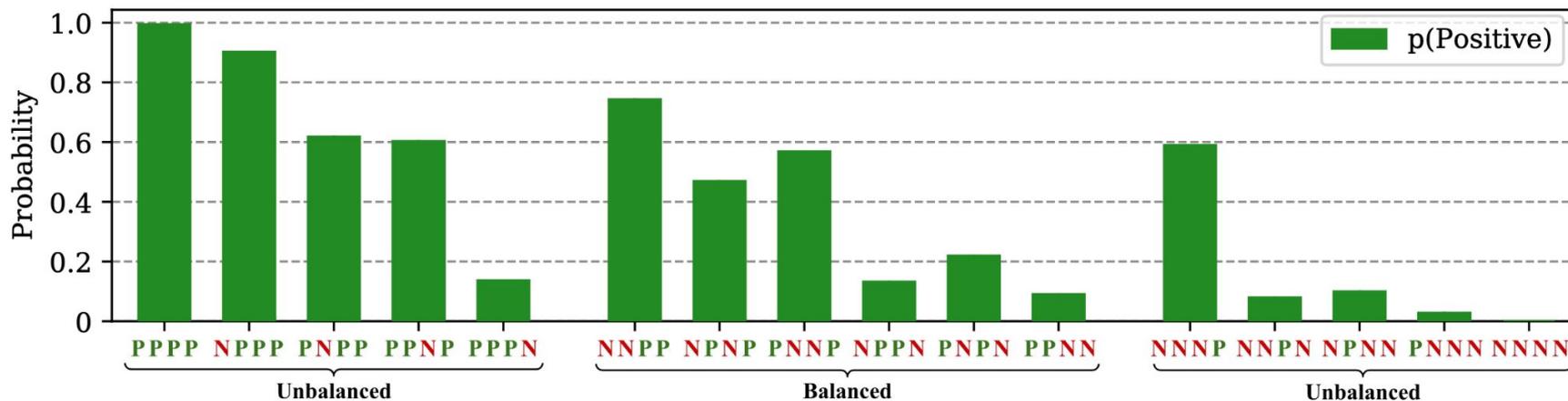
- The input-label mapping
- The distribution of the input text
- The label space
- The format



Other Considerations

Majority label: The exemplar space is dominated by a single (or a subset) of labels.

Recency bias: Labels seen later in the sequence.



In-Context Learning (Summary)

In-context learning is NOT robust to:

- The language of the prompt
- The language of the few-shot exemplars
- The ordering of few-shot exemplars
- The labels of the few-shot exemplars



Predicting Legal Factors from Court Opinions

Drug Interdiction Vehicle Stops: Legal Standards

- Courts assess whether police had:
 - Reasonable Suspicion → Brief detention for canine unit
 - Probable Cause → Immediate vehicle search
- Common "**Factors of Suspicion**":
 - Rental vehicles
 - Unusual travel plans
 - Strong air fresheners
 - Other observable behaviors
- Judicial Review Process:
 - Officers make initial determination
 - Judge reviews whether detention/search was legally justified
- Thousands of cases litigated at state, federal, and Supreme Court levels



Factors of Suspicion

| 1 Occupant Appearance or Behavior | 2 Occupant Status |
|--|--|
| 1A Furtive Movement | 2E Motorist License |
| 1B Physical Appearance of Nervousness | 2F Driver Status |
| 1C Nervous Behavior | 2G Refused Consent |
| 1D Suspicious or Inconsistent Answers | 2H Legal Indications of Drug Use |
| | 2I Motorist's Appearance Related to Drug Use |
| 3 Travel Plans | 4 Vehicle |
| 3J Possible Drug Route | 4L Expensive Vehicle |
| 3K Unusual Travel Plans | 4M Vehicle License Plate or Registration |
| | 4N Unusual Vehicle Ownership |
| 5 Vehicle Status | 6 Other Annotation Labels |
| 5O Indicia of Hard Travel | 6T Other |
| 5P Masking Agent | 6U Possibly Off Point |
| 5Q Vehicle Contents Suggest Drugs | 6V Suspicion Found? - No |
| 5R Suspicious Communication Device | 6W Suspicion Found? - Yes |
| 5S Suspicious Storage | |

Typical ELS Project (text analysis focused)

Deductive Coding

- Specify task (annotation guidelines)
- Pilot study (update guidelines)
- **Use machine learning to annotate data**
- Hire and train human annotators
- Annotate data set for evaluation
- Analyze results

Automatic Factor Annotation

| 1 Occupant Appearance or Behavior | 2 Occupant Status |
|---------------------------------------|--|
| 1A Furtive Movement | 2E Motorist License |
| 1B Physical Appearance of Nervousness | 2F Driver Status |
| 1C Nervous Behavior | 2G Refused Consent |
| 1D Suspicious or Inconsistent Answers | 2H Legal Indications of Drug Use |
| | 2I Motorist's Appearance Related to Drug Use |
| 3 Travel Plans | 4 Vehicle |
| 3J Possible Drug Route | 4L Expensive Vehicle |
| 3K Unusual Travel Plans | 4M Vehicle License Plate or Registration |
| | 4N Unusual Vehicle Ownership |
| 5 Vehicle Status | 6 Other Annotation Labels |
| 5O Indicia of Hard Travel | 6T Other |
| 5P Masking Agent | 6U Possibly Off Point |
| 5Q Vehicle Contents Suggest Drugs | 6V Suspicion Found? - No |
| 5R Suspicious Communication Device | 6W Suspicion Found? - Yes |
| 5S Suspicious Storage | |

| | |
|---------------------------------------|------|
| minimum number of sentences in case: | 33 |
| maximum number of sentences in case: | 1090 |
| average number of sentences in case: | 224 |
| median number of sentences in case: | 185 |
| number of sentences in 25th quartile: | 123 |
| number of sentences in 75th quartile: | 288 |

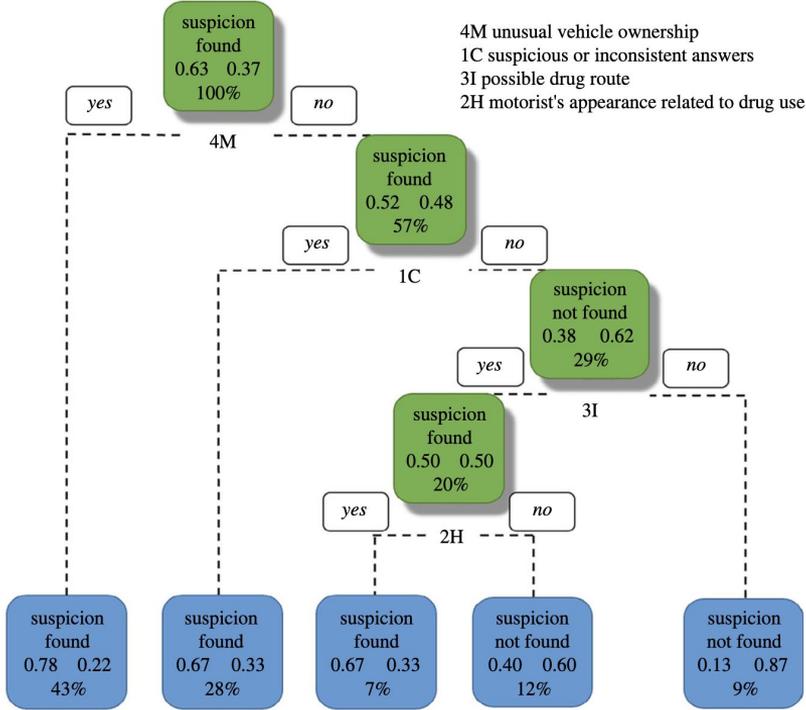
211 cases overall

| | P | R | F1 | n |
|---|------|------|------|------|
| no_type | 0.99 | 0.99 | 0.99 | 8764 |
| 3J Possible Drug Route | 0.90 | 0.94 | 0.92 | 47 |
| 5S Suspicious Storage | 0.94 | 0.91 | 0.92 | 32 |
| 1B Physical Appearance of Nervousness | 0.88 | 0.92 | 0.90 | 126 |
| 1D Suspicious or Inconsistent Answers | 0.86 | 0.87 | 0.87 | 127 |
| 4N Unusual Vehicle Ownership | 0.87 | 0.88 | 0.87 | 75 |
| 5P Masking Agent | 0.86 | 0.86 | 0.86 | 28 |
| 2E Motorist License or Identification | 0.94 | 0.79 | 0.86 | 19 |
| 6T Other | 0.86 | 0.86 | 0.86 | 14 |
| 5O Indicia of Hard Travel | 0.90 | 0.82 | 0.86 | 11 |
| 5R Suspicious Communication Device | 0.88 | 0.82 | 0.85 | 17 |
| 6W Suspicion Found? - Yes | 0.82 | 0.85 | 0.83 | 53 |
| 2G Refused Consent | 0.78 | 0.88 | 0.83 | 49 |
| 6V Suspicion Found? - No | 0.83 | 0.83 | 0.83 | 30 |
| 3K Unusual Travel Plans | 0.88 | 0.76 | 0.82 | 38 |
| 2H Legal Indications of Drug Use | 0.83 | 0.79 | 0.81 | 57 |
| 1C Nervous Behavior | 0.78 | 0.82 | 0.80 | 96 |
| 5Q Vehicle Contents Suggest Drugs | 0.77 | 0.73 | 0.75 | 41 |
| 4M Vehicle License Plate or Registration | 0.64 | 0.90 | 0.75 | 10 |
| 2I Appearance Related to Drug Use or Sale | 0.75 | 0.71 | 0.73 | 17 |
| 1A Furtive Movement | 0.79 | 0.62 | 0.70 | 24 |
| 4L Expensive Vehicle | 0.33 | 1.00 | 0.50 | 1 |
| 2F Driver Status | 1.00 | 1.00 | 1.00 | 1 |

Automatic Factor Annotation (cont.)

| GPT model | all cases | | | short cases | | | long cases | | |
|---------------|-----------|------|---------|-------------|------|---------|------------|------|---------|
| | 3.5 | 4 | 4-Turbo | 3.5 | 4 | 4-Turbo | 3.5 | 4 | 4-Turbo |
| accuracy | 0.79 | 0.91 | 0.89 | 0.77 | 0.91 | 0.88 | 0.86 | 0.89 | 0.93 |
| intersection | 0.81 | 0.97 | 0.87 | 0.80 | 0.97 | 0.87 | 0.86 | 0.99 | 0.86 |
| false factors | 2.9 | 2.6 | 1.5 | 2.7 | 2.4 | 1.3 | 4.0 | 4.6 | 2.6 |

Decision Tree from Training Data



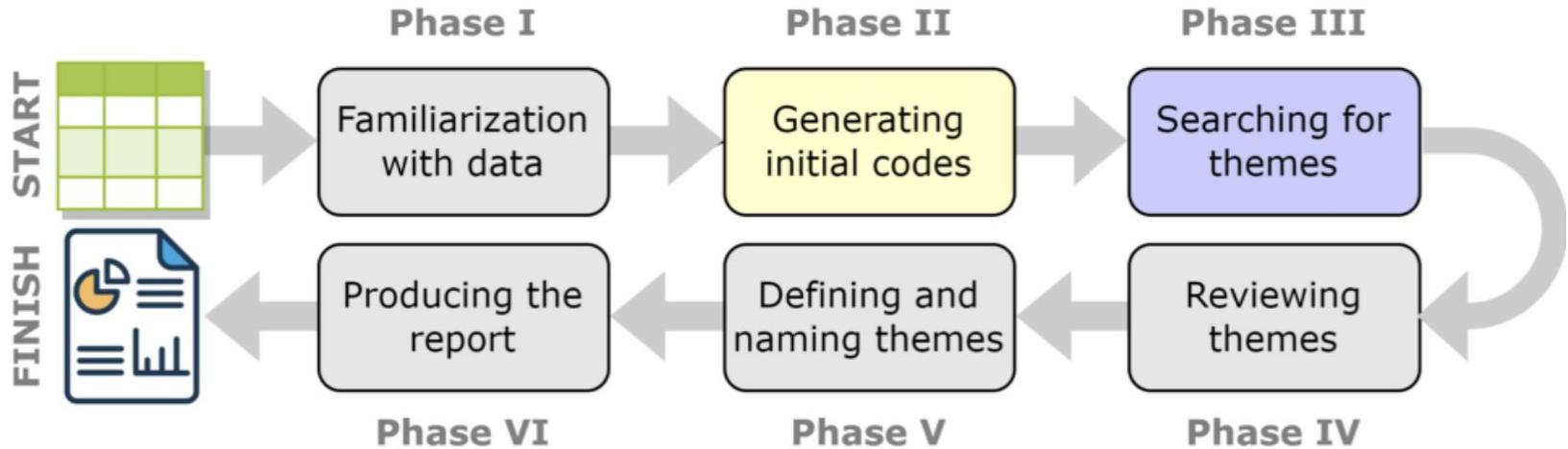


Inductive Analysis of Criminal Court Opinions on Thefts

Motivation

- Thematic analysis of criminal courts' opinions (focused on the criminal offense of theft in Czechia)
- Criminal offense categories (e.g., theft, murder) are usually defined in statutory law
- The individual criminal acts are described by courts when they apply the law to factual circumstances of the cases
- To get insight into what behavior is criminalized and how effectively, it is necessary to identify shared features of criminal acts, and generalize them into "**typical crimes**"
→ This is akin to performing thematic analysis
- Such analysis is an expensive and time-consuming endeavour

Six Phases of Thematic Analysis



Research Questions

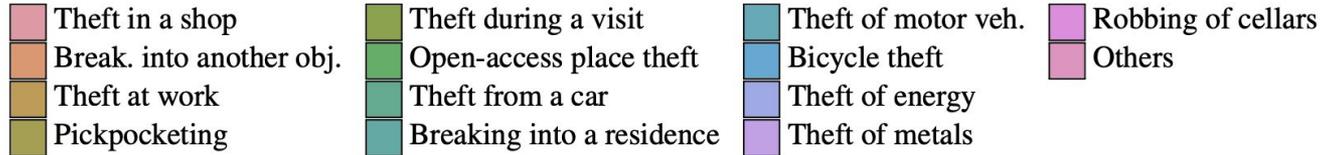
(RQ1) How successfully can the LLM perform initial coding of the data?

(RQ2) To what degree can a subject matter expert improve the quality of the initial codes via natural language feedback?

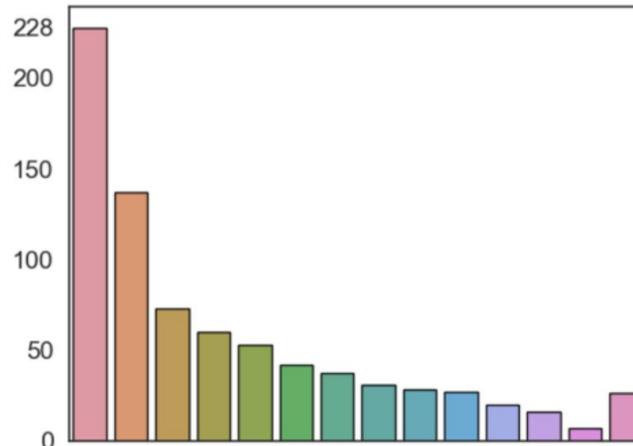
(RQ3) How successfully can the LLM predict themes for the analyzed data points?

(RQ4) How successfully can the LLM autonomously discover themes and associate them with the analyzed data?

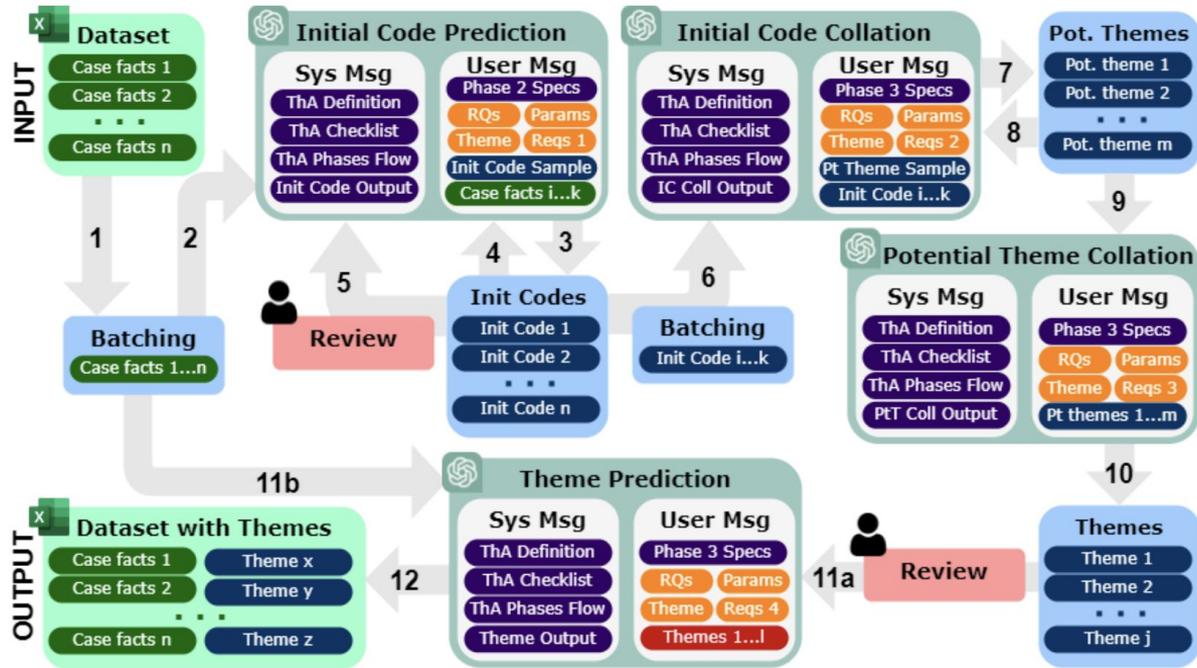
Dataset (human annotated - inductive)



At an undetermined time between 18:00 on May 12, 2017 and 06:00 on May 13, 2017, at the parked delivery vehicle branded Peugeot Boxer, an unknown individual used an unidentified object to pry open the locks of the driver's door, the passenger door, and then the cargo space. The individual **entered the vehicle and stole from it** a car radio, a demolition hammer, an electric saw, a drill, and other work tools, all valued at 8,700 CZK [...] By damaging the door lock, he caused damage worth 3,500 CZK. The stolen items were sold to unknown persons.



Proposed Framework



Initial Codes (RQ1 + RQ2)

Initial Codes Evaluation Scheme

1. \neg How: If the code does not address (even implicitly) how the theft happened; \rightarrow
2. \neg What: If the code does not address (even implicitly) what was stolen; \rightarrow
3. Ok: The code addresses how the theft happened and what was stolen. \square

| | \neg How | \neg What | Ok | \neg Ok |
|-------------------------------|-------------|-------------|-------------|-------------|
| Before expert feedback | 104 (13.2%) | 111 (14.1%) | 570 (72.6%) | 215 (27.4%) |
| After expert feedback | 16 (2.0%) | 72 (9.2%) | 697 (88.8%) | 88 (11.2%) |

| Before expert feedback | After expert feedback |
|--|--|
| Private theft of cash from a residential space | Theft of large quantity of cash from relative's home |
| Forced entry and theft involving an automobile | Burglary and theft of work tools from vehicle |
| Shoplifting - personal care items | Shoplifting of shaving equipment from drugstore |

Typical ELS Project (text analysis focused)

Inductive Coding *(single researcher case)*

- Familiarization with data
- **Generating initial codes**
- Searching for themes
- Reviewing themes
- Defining and naming themes
- Projecting themes onto the data
- Analyze results

Theme Prediction (RQ3)

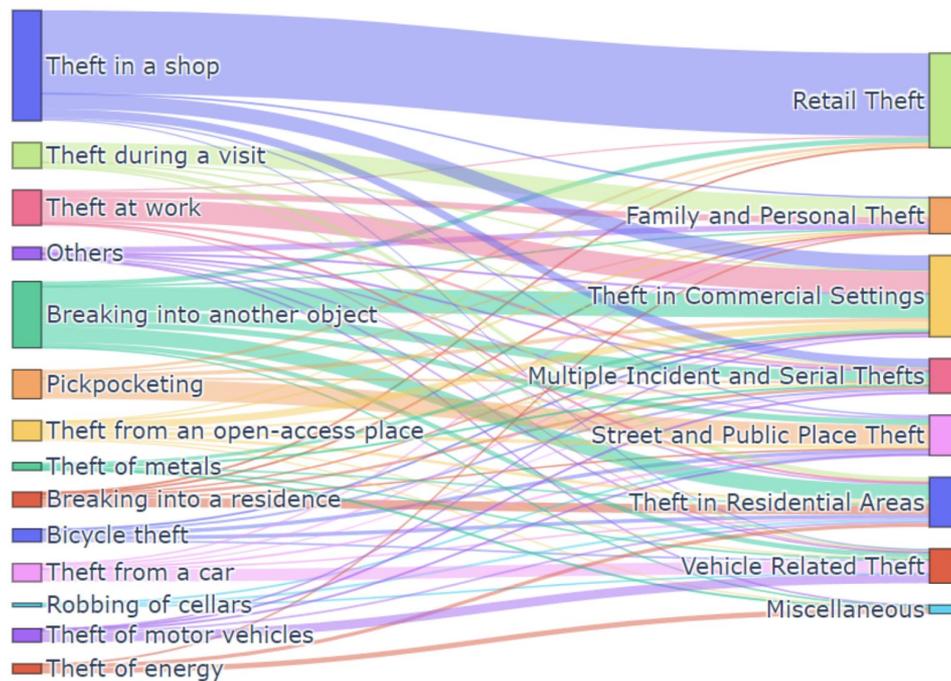
| Manual Theme | R@1 | R@3 | Manual Theme | R@1 | R@3 |
|---------------------------------|------------|------------|---------------------------|-----|-----|
| Theft in a shop | .95 | .96 | Breaking into a residence | .52 | .71 |
| Theft during a visit | .75 | .91 | Bicycle theft | .74 | .96 |
| Theft at work | .71 | .86 | Theft from a car | .70 | .84 |
| Breaking into another object | .35 | .67 | Robbing of cellars | .14 | .57 |
| Pickpocketing | .68 | .87 | Theft of motor vehicles | .50 | .75 |
| Theft from an open-access place | .21 | .29 | Theft of energy | 1.0 | 1.0 |
| Theft of metals | .69 | .88 | Others | .23 | .73 |
| Overall | .66 | .82 | | | |

Typical ELS Project (text analysis focused)

Inductive Coding *(single researcher case)*

- Familiarization with data
- Generating initial codes
- Searching for themes
- Reviewing themes
- Defining and naming themes
- **Projecting themes onto the data**
- Analyze results

Theme Discovery and Prediction (RQ4)



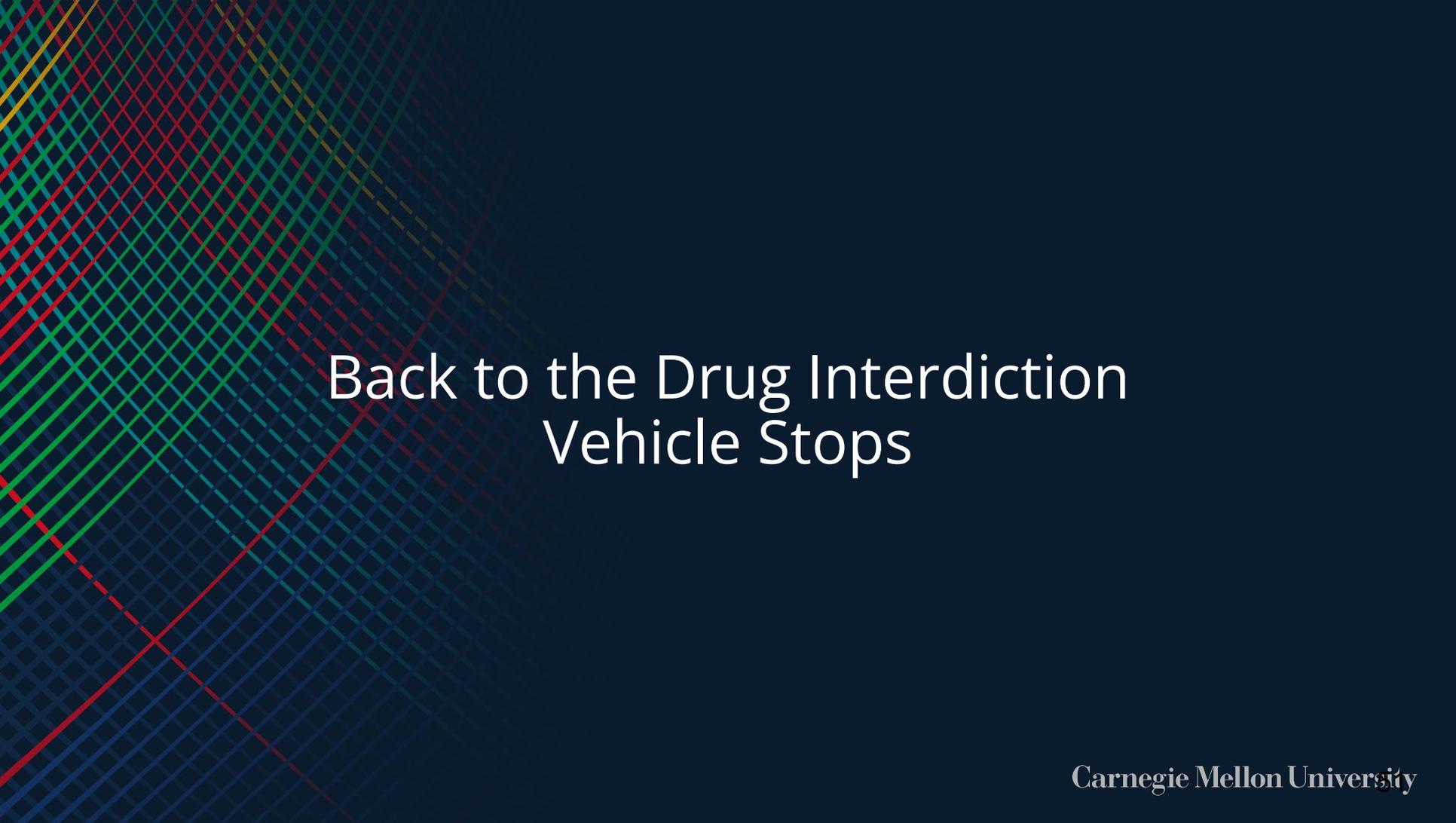
Typical ELS Project (text analysis focused)

Inductive Coding *(single researcher case)*

- Familiarization with data
- Generating initial codes
- Searching for themes
- Reviewing themes
- Defining and naming themes
- Projecting themes onto the data
- Analyze results

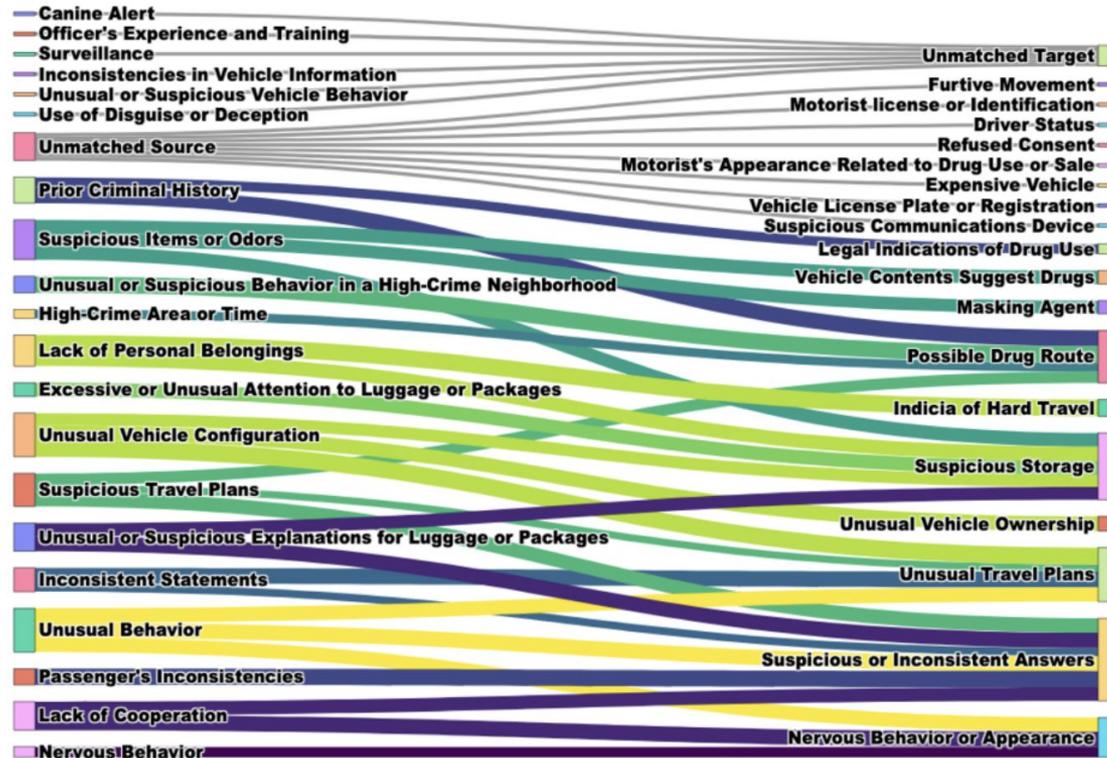
Overall

- The initial coding of data was performed with reasonable quality (RQ1).
- The coding further improved when expert feedback was provided (RQ2).
- The performance on zero-shot classification of facts descriptions in terms of categories of theft was promising (RQ3) but could likely benefit from expert feedback (future work).
- The evaluation of the end-to-end performance of the pipeline on discovering and predicting themes suggested viability of the proposed framework (RQ4)
→ subject matter expert supervision important



Back to the Drug Interdiction Vehicle Stops

Automatically Induced Themes for Drug Interdiction Vehicle Stops





Burning Questions

Typical ELS Project (text analysis focused)

Deductive Coding (*hired annotators case*)

- AI performs the whole task end-to-end

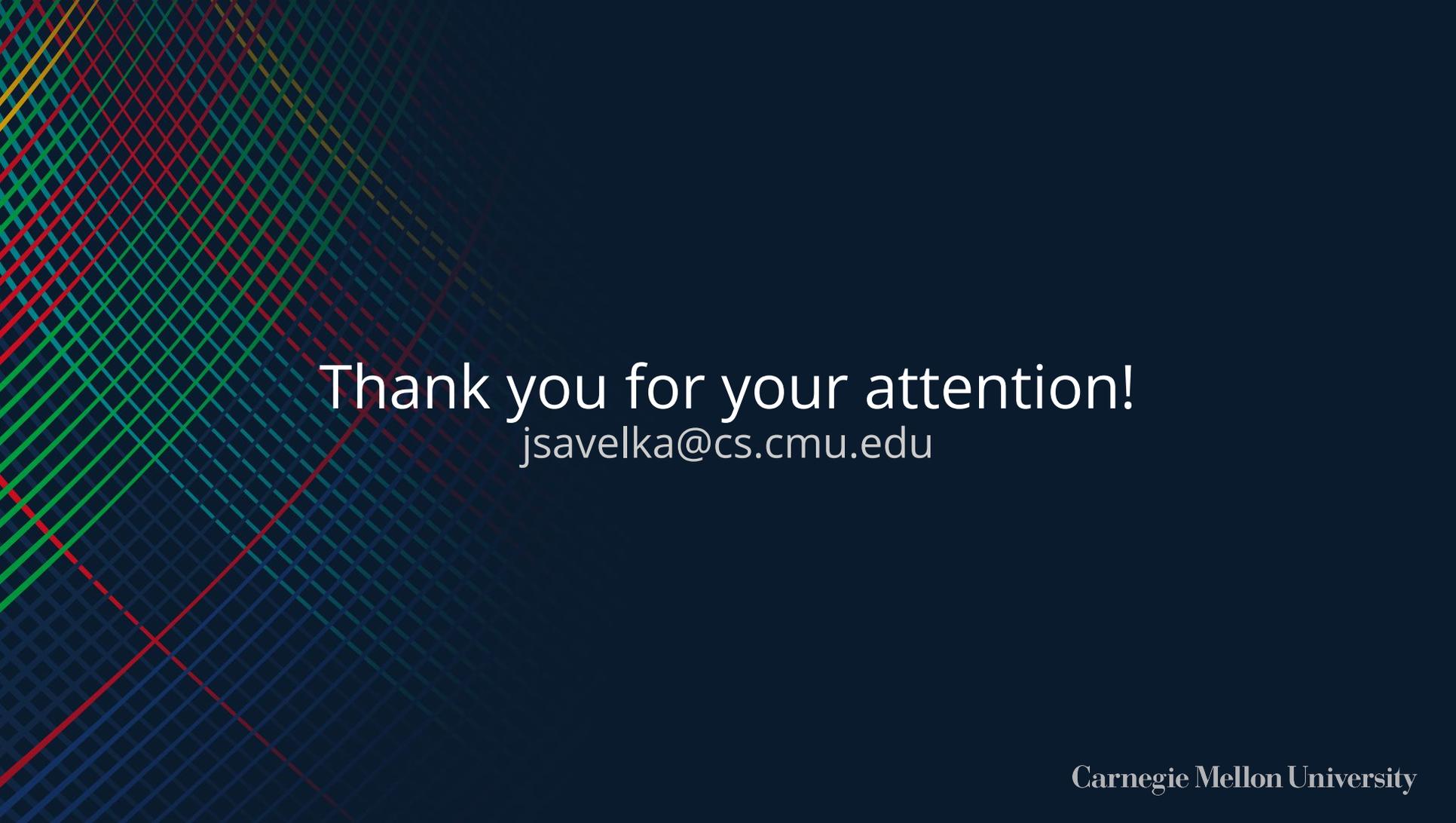
Inductive Coding (*single researcher case*)

- AI performs the whole task end-to-end



Burning Questions

1. How can we validate the results produced by the automatic pipeline?
2. How do we design interactions between AI frameworks and humans to ensure continued growth of a human as an expert?



Thank you for your attention!
jsavelka@cs.cmu.edu