Learning Theories and Learning Engineering

ISEA Session 7

Min Sun and Alex Liu
University of Washington
3.8.2024

IES Institute of Education Sciences
AmplifyLearn.AI
UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS
UNIVERSITY OF OREGON

Warm Up

> **Is programming a skill still needed in the AI era, particularly with the unprecedented speed of advancement of Large Language Modeling (LLM)?**

Today's Agenda

1. Natural Language Processing (NLP)
2. Theme identification, text summarization
    1. Traditional NLP—Topic Modeling
    2. LLM Prompting Engineering
    3. Application in one paper: analyzing interview data
3. The benefits of human and computer interactive learning in the AI era
4. Code and program
5. Classification Models (next week)
    1. Lexicon-based sentiment analysis versus LLM
    2. Multi-tasks classification based on neural network models
    3. RADAR for personalization
    4. Code

Natural Language Processing (NLP)

- “Natural language processing (NLP) combines computational linguistics, machine learning, and deep learning models to process human language.”
  - Amazon Web Services (AWS)

Natural Language Processing (NLP)

- **Computational linguistics**: the science of *understanding and constructing human language models* with computers and software tools. Tools like: Language translators, Text-to-speech synthesizers; Speech recognition software are based on computational linguistics.
- **Machine Learning:** A technology trains a computer with sample data to improve its efficiency. Machine learning (ML) methods teach NLP applications to *recognize and accurately understand human language*.
- **Deep Learning:** A subset of machine learning *that teaches computers to learn, think, act like humans*. It involves neural network models to enable computers to recognize, classify, and co-relate complex patterns in language data.

Traditional Text as Data Methods

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

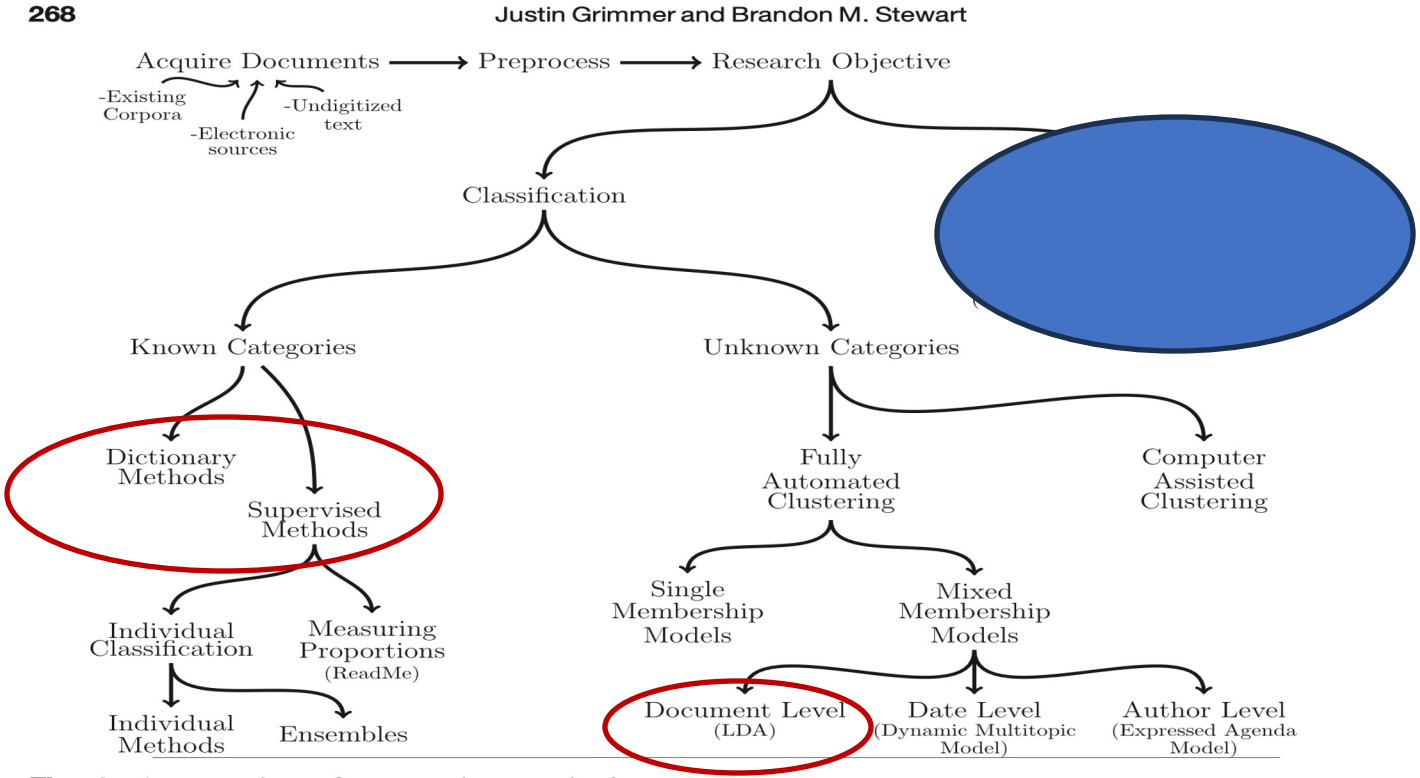


Fig. 1 An overview of text as data methods.

Topic Modeling: latent Dirichlet allocation (LDA, see Blei, 2012; Blei, Ng, & Jordan, 2003)

LDA is a generative statistical model that identifies the latent topics and corresponding proportions that compose a document.

LDA assumes that each document is a mixture of topics. For each document, π_{ik}, represents the proportion of task *i* dedicated to topic *k*. Each task collects the proportions across topics, as $\pi_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik})$. LDA represents a given task as a set of topic weights, rather than assigning it to one of those topics.
(see an example next slide)

Text Analysis: latent Dirichlet allocation (LDA, see Blei, 2012; Blei, Ng, & Jordan, 2003)

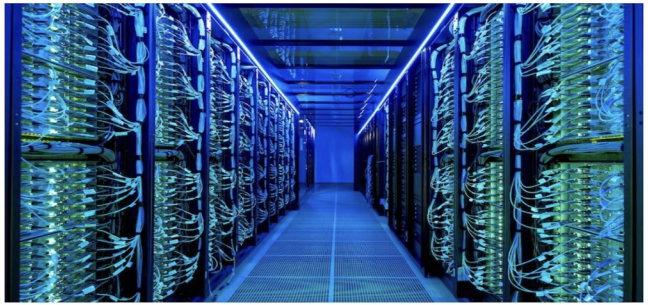
	Topic 1: Leadership	Topic 2: Teacher Capacity Building	Topic 3: Engaging Parents
1	0.6	0.1	0.3
W-2	0.4	0.5	0.1
+3	0.1	0.1	0.8

Large Language Models (LLM)

Training them is more involved.

Think of it like compressing the internet.

Chunk of the internet,
~10TB of text



6,000 GPUs for 12 days, ~$2M
~1e24 FLOPS

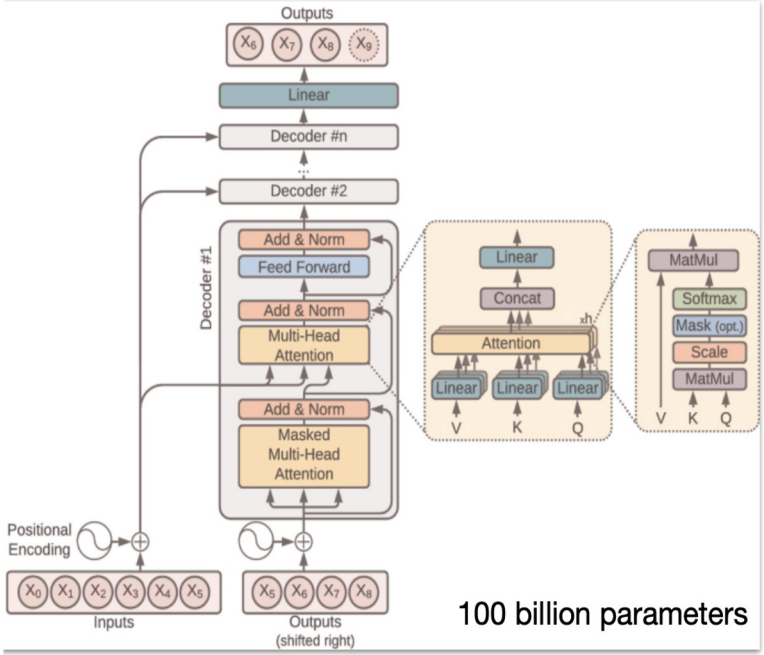
parameters.zip

~140GB file

*numbers for Llama 2 70B

Andrej Karpathy (2023): Intro to LLM for busy people.
https://www.youtube.com/watch?v=zjkBMFhNj_g&ab_channel=AndrejKarpathy

LLM: How does it work?



"Little is known in full detail...

- **Billions of parameters are dispersed through the network.**
- **We know how to iteratively adjust them to make it better at prediction. We can measure that this works, but we don't really know how the billions of parameters collaborate to do it.**
- **They build and maintain some kind of knowledge database, but it is a bit strange and imperfect."**

- Andrej Karpathy (2023): Intro to LLM for busy people. https://www.youtube.com/watch?v=zjkBMFhNj_g&ab_channel=AndrejKarpathy
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). **Attention is all you need**. *Advances in neural information processing systems*, *30*. https://arxiv.org/pdf/1706.03762.pdf

From Voices to Validity: Leveraging Large Language Models (LLMs) for Textual Analysis of Policy Stakeholder Interviews

Liu, A., & Sun, M. (2023). From Voices to Validity: Leveraging Large Language Models (LLMs) for Textual Analysis of Policy Stakeholder Interviews. *arXiv preprint arXiv:2312.01202*.

This study is supported by the Baller Group and William T. Grant Foundation (Grant No. 190735) and the National Science Foundation (Grant No. 2055062). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Liu & Sun assume equal authorship.

IES Institute of Education Sciences

UNIVERSITY of WASHINGTON

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Introduction

1. One important source of data to support policy discourse and decision-making involves **stakeholders'** lived experiences about the implementation of current policy and their opinions about how to improve.
2. Stakeholders' voices may be collected from **interviews**, open-ended survey responses, or texts obtained from social media posts.
3. The **cost of manually** analyzing even a moderately sized text may hinder the actual use of stakeholders' voices.
4. Data science methods—like topic modeling (LDA), sentiment analysis, and large language models (LLMs, notably ChatGPT)—may offer the efficiency, but can be constrained by **the lack of domain and contextual knowledge.**
5. The central aim of this study is to examine **the validity of LLMs** to analyze interview data about a specific domain—education policies and programs—in a specific context—Washington state's K-12 school system.

Research Questions

A large study of identifying policies and programs that either advance or hinder racial and economic equity in Washington (WA) State's K-12 public school system in 2022.

Substance Research Questions:

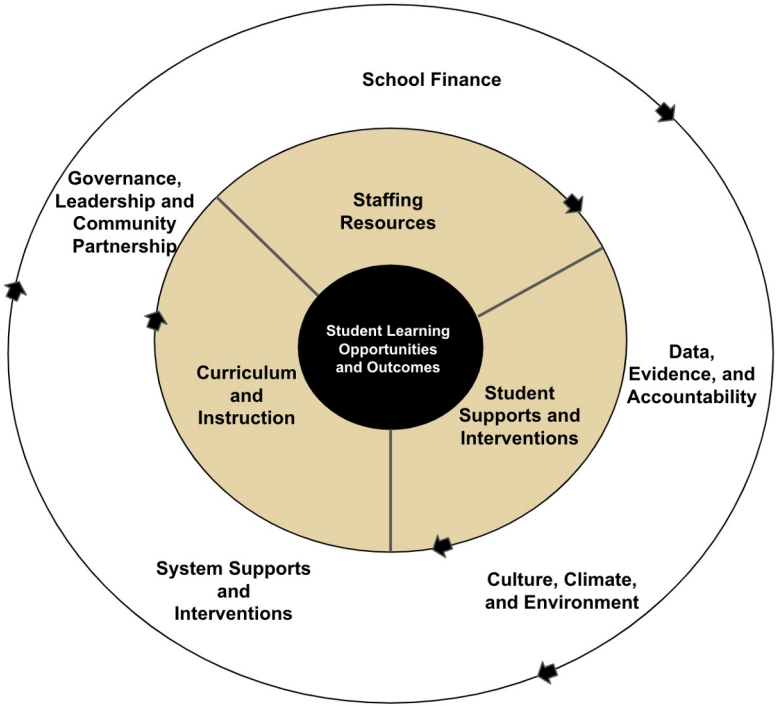
1. What are the **key themes that WA stakeholders voiced** about K-12 public school system?
2. Which themes did stakeholders **recognize as advancing educational equity (positive)**? Conversely, which areas were mentioned **as needing improvement or hinder (negative)** educational equity?

Methodological Research Questions:

1. How accurate and valid are GPT-4 labels of key themes when comparing to human experts' labels and traditional topic modeling results?
2. How accurate and valid are GPT-4 sentiment classifications when comparing to human experts' and lexicon-based sentiment analysis?

Conceptual Framework

Resources Equity Framework Embedded in a Data-Informed Iterative Improvement Cycle



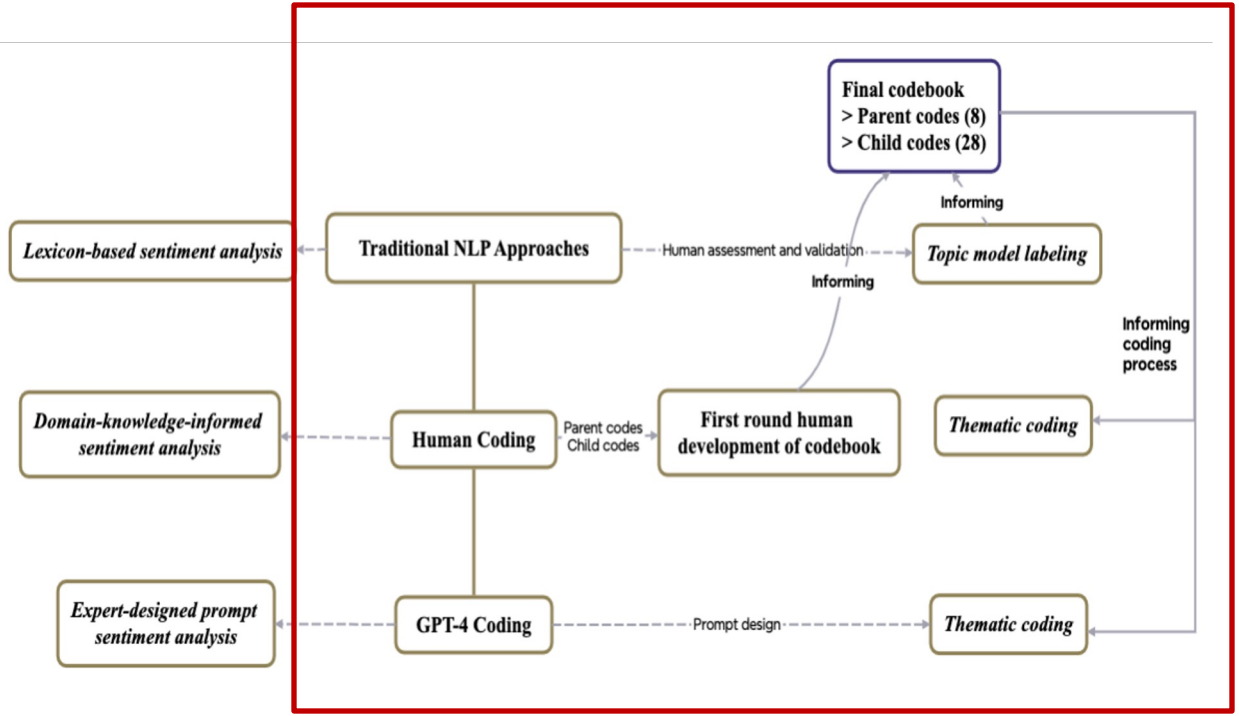
Data Collection and Preprocessing

1. **24 interviews (45-60 minutes)** with stakeholders:
   - **Administrators:** state legislators, other state-level policymakers, school district administrators;
   - **Non-Profit and Advocates:** teacher union representatives, policy advocates, and community leaders;
   - **Educators:** teachers, teacher coaches or mentors
2. Tidytext-format data contains about 1,700 entries (i.e. documents).
   - One complete thought (one long or several short sentences)
   - Filtered out stop words and words like “um,” “so,” and “you know”
   - Stemming
3. Contains interviewees’ research ID, demographics, job roles, and job location.

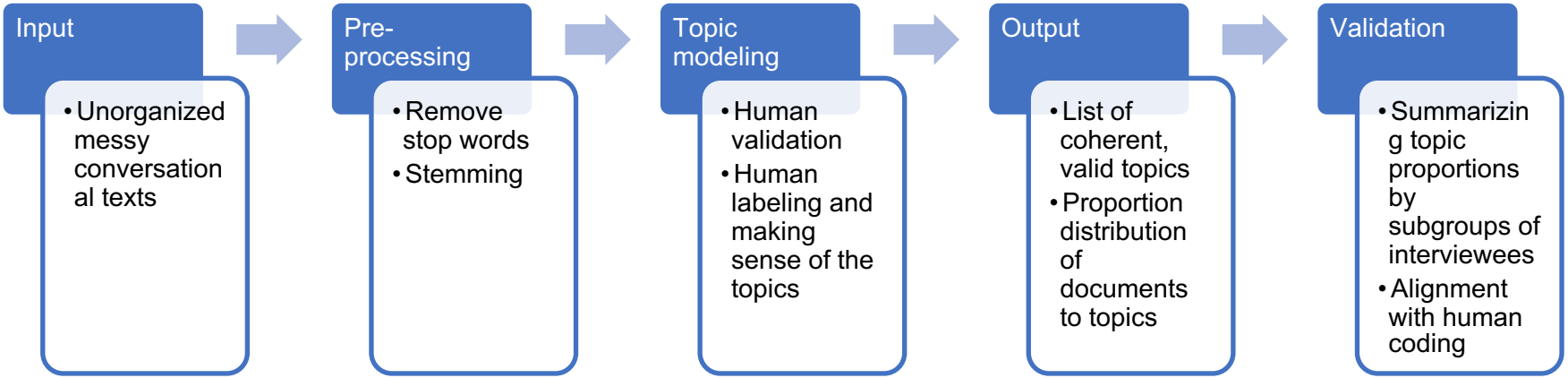
Methods: Human-Computer Interactive Approach

"This human-computer partnership approach (socio-technical) perspective on learning highlights how technology exerts agency in a relationship with educators and learners, as generative AI learns more with the quantity and quality of what humans provide to the technology, and vice versa. The reciprocal nature of AI learning and teaching interactions means that less bias and more flexibility in the tool are possible with increased human use."

-- Katie Headrick Taylor, Alex Liu, Min Sun

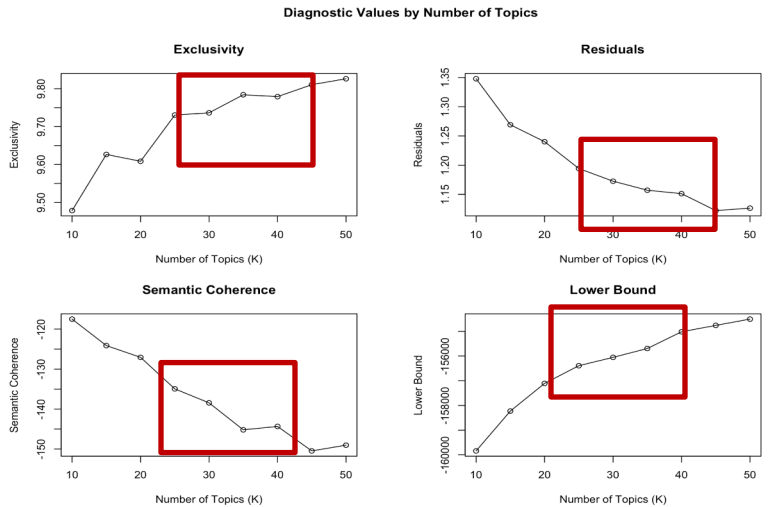


LDA

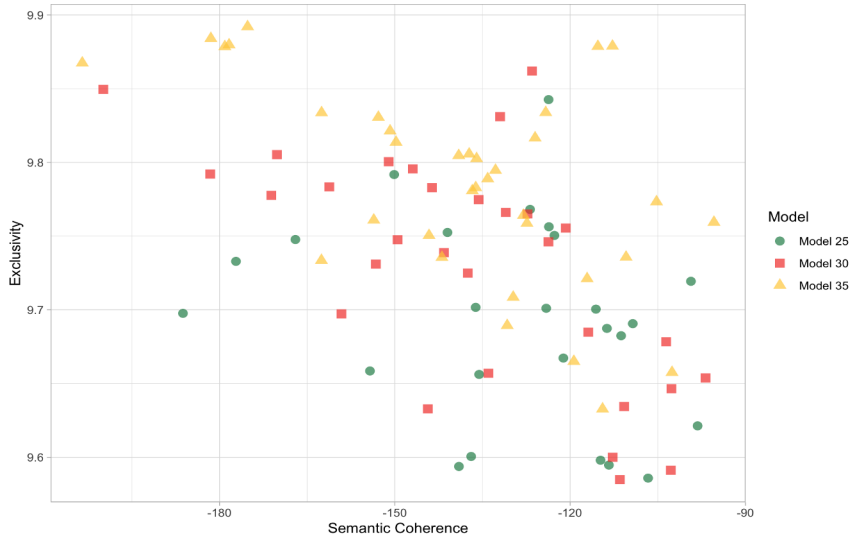

Input
• Unorganized messy conversational texts
Pre-processing
• Remove stop words
• Stemming
Topic modeling
• Human validation
• Human labeling and making sense of the topics
Output
• List of coherent, valid topics
• Proportion distribution of documents to topics
Validation
• Summarizing topic proportions by subgroups of interviewees
• Alignment with human coding

LDA

Diagnostic Values for a Number of Topics



Exclusivity vs. Semantic Coherence



IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Prompt Engineering

- **Zero-shot**
  - Use the codebook to label three child and/or parent codes for each document; if no child code works, only label the parent codes.
- **Few shot:**
  - Zero shot + here are a few examples.
- **Chain of thoughts**
  - Step 1: label three parent codes with reasoning; Step 2: label child code within each parent code. If none of the child code applies, keep the parent code.

Resources to learn more about prompt engineering:

1. https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/
2. https://www.youtube.com/watch?v=dOxUroR57xs&ab_channel=ElvisSaravia

GPT-4 Thematic Analysis: Chain-of-Thought (CoT) Prompt

***prompt_base = f*"""**

Task: As a policy researcher, you've been provided with a paragraph extracted from an interview with an education policy stakeholder. Utilize the attached "Voice to validity codebook " (in excel format) to code the paragraph. The Codebook comprises four columns: 'Parent', 'Child', 'Child_description', and 'Key words'.

Role, context, and overall task

Steps:

1. Identify Salient Themes:
   - Understand the paragraph's content within the context of the Washington State K-12 public school system.
   - Refer to the 'Parent' column in the Codebook for broader thematic categories.
   - Pinpoint up to three salient themes from these 'Parent' categories.
   - These themes should highlight the most significant ideas in the paragraph.
   - Label the paragraph with the chosen 'Parent' themes.
2. Dive into Child Themes:
   - The 'Child' column in the Codebook lists detailed thematic subcategories, which fall under the broader 'Parent' categories.
   - The 'Child_description' elaborates on the 'Child' categories, and the 'Key words' column lists pertinent terms for each 'Child' category.
3. Associate with Child Categories:
   - Revisit the paragraph, keeping the Washington State K-12 public school system context in mind.
   - For each previously identified 'Parent' theme, pinpoint the appropriate 'Child' subcategories from the Codebook. The 'Child_description' and 'Key words' columns can aid your decision.
   - Ensure the 'Child' categories align with the paragraph's content. If there's no fit or you're uncertain, label it as 'None'.
   - From your identified 'Parent' and 'Child' pairs, pick the top three pairs that encapsulate the paragraph's central ideas.
   - Label the paragraph with these three 'Parent' and corresponding 'Child' pairs.

Codebook: see attachment

Paragraph for Analysis:

[[[TEXTGOHERE]]]

Response Format:

Frame your answer as a JSON object containing the keys: 'Parent 1', 'Child 1', 'Parent 2', 'Child 2', 'Parent 3', 'Child 3', and 'Reasoning'.

"""

Codebook

> Snippet of the codebook

Appendix A.3. Final Version of Codebook with Agreement Ratings

Topic # from LDA	Parent	Child	Child description	Agreed rating	Key words
23	Culture, climate and environment	Trauma at home	Struggling home and family experiences of children of high poverty and of color negatively influence their school learning and graduation pathways post pandemic	4	bad, kid, children, home, school, famili, grade, covid-19, hispan, pandem, third, rate, happen, stop, number, class, get, year, somebody, want
14	Culture, climate and environment	Anti-racism	Talking about whiteness, success, and anti-racism	3	pull, white, child, success, stay, built, job, indic, keep, whole, four, barrier, stori, middle, racial, ago, pretti, feel, understand, day
None	Curriculum and instruction	Instructional programs	AP/IB courses, college classes/credits in high school, special education programs, bilingual programs, ethnic studies		Equity, Students, School, District, Policy, Data, Programs, Honors, Course, AP, IB, State, Access, Advanced, Highly Capable, College, Racial, Learning, Services, Practices.
13	Curriculum and instruction	Curriculum development and instructional delivery	School curriculum development and instructional delivery, specially including culturally responsive teaching and equitable pedagogical practices that influence students' experience in the classroom	3	onlin, teach, taught, teacher, high, learn, class, middle, science, elementari, experi, student, day, next, sit, first, noth, life, school, potenti
11	Data, evidence, and accountability	Data access, analysis, reporting, use, quality and transparency	Data collection, access, analysis, and use to help practitioners improve their practices/strategies and to help policymakers (at all levels - e.g., school building, district, state legislators) make new policies or refine current policies. This also pertains to data sharing and reporting, as well as data transparency and quality issues.	4	dashboard, data, inform, assumess, collect, disaggreg, report, websit, access, use, ospi, yes, effect, analysi, tool, together, good, stori, wsif, point

IES Institute of Education Sciences

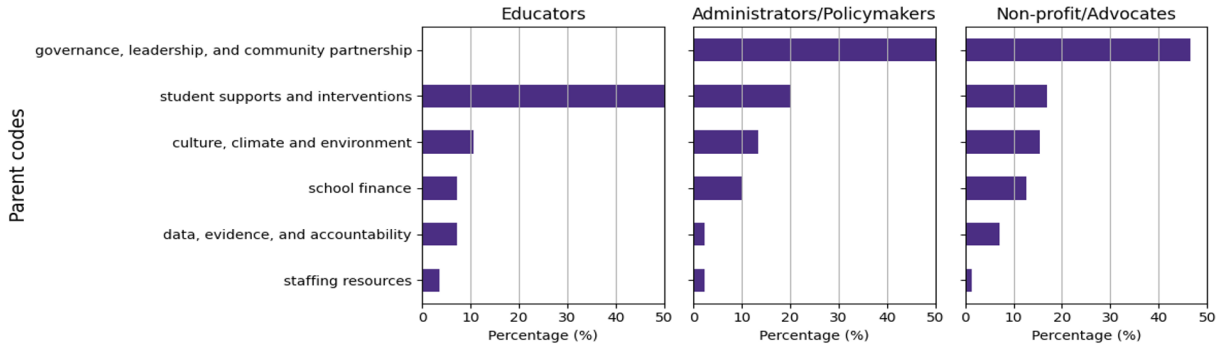
Amplify

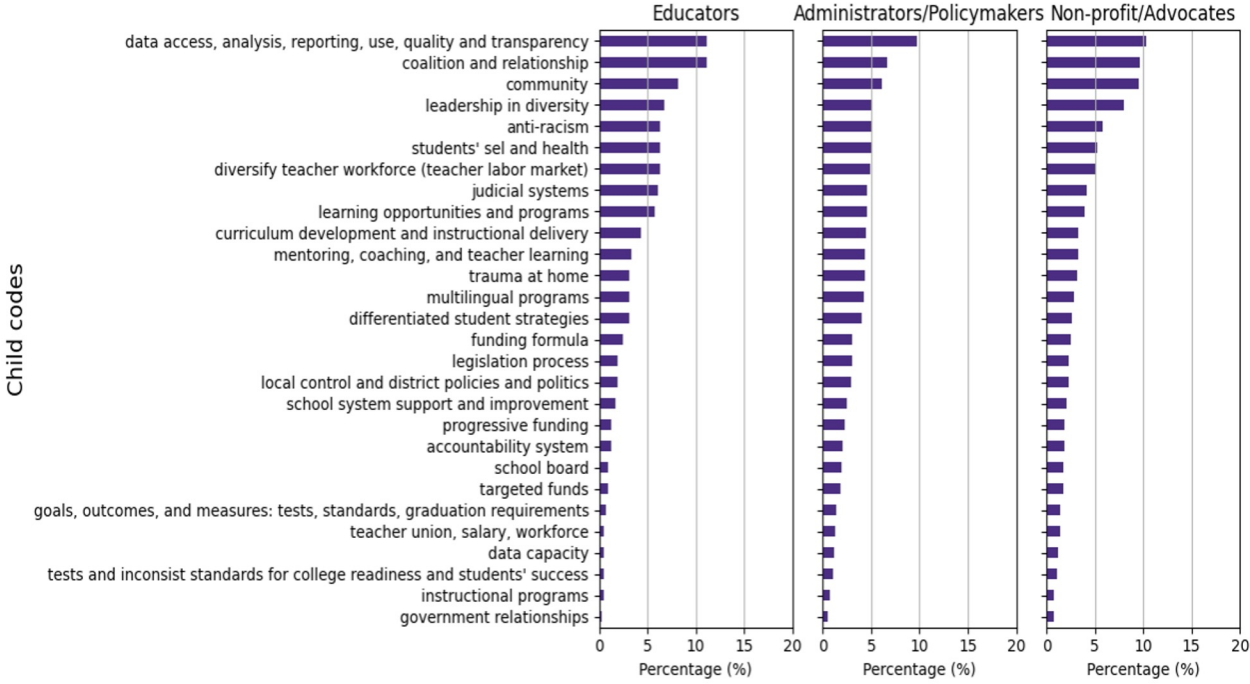
Interactive Demo (10 mins)

- Min will prepare codebook in excel.
- 2-3 interviews.
- Prompt

Results–SRQ1

What are the **key themes that WA stakeholders voiced** about K-12 public school system?





IES Institute of Education Sciences

Results– MRQ1

- How accurate and valid are GPT-4 labels of key themes when comparing to human experts' labels and traditional topic modeling results?

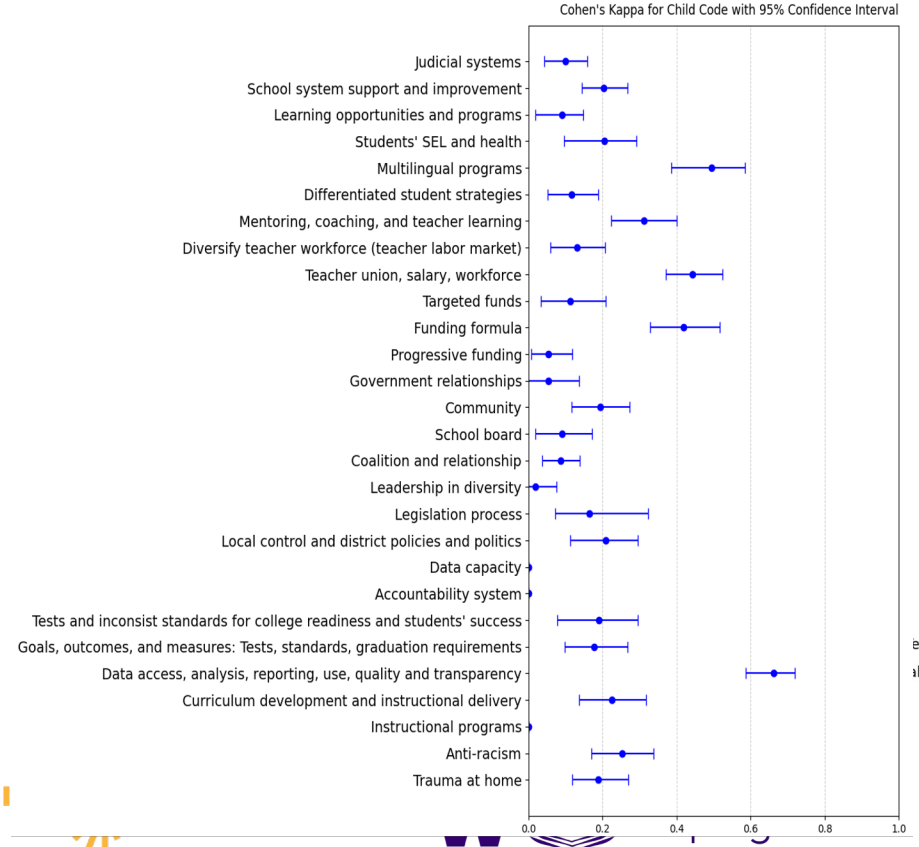
Table 2. Comparison of Agreement and Confusion Matrix Metrics

	Agreement Metrics (Child codes)					Confusion Matrix Metrics (Parent codes)				
	% Hit rates	% Shuffled hit rate	Precision	Recall	F-Score	% Hit rates	% Shuffled hit rate	Precision	Recall	F1-Score
GPT-4 vs. Human	77.89	17.89	0.33	0.63	0.42	96.02	56.67	0.52	0.87	0.62
LDA vs. Human	60.65	13.66	0.23	0.38	0.27	76.13	47.80	0.43	0.64	0.49

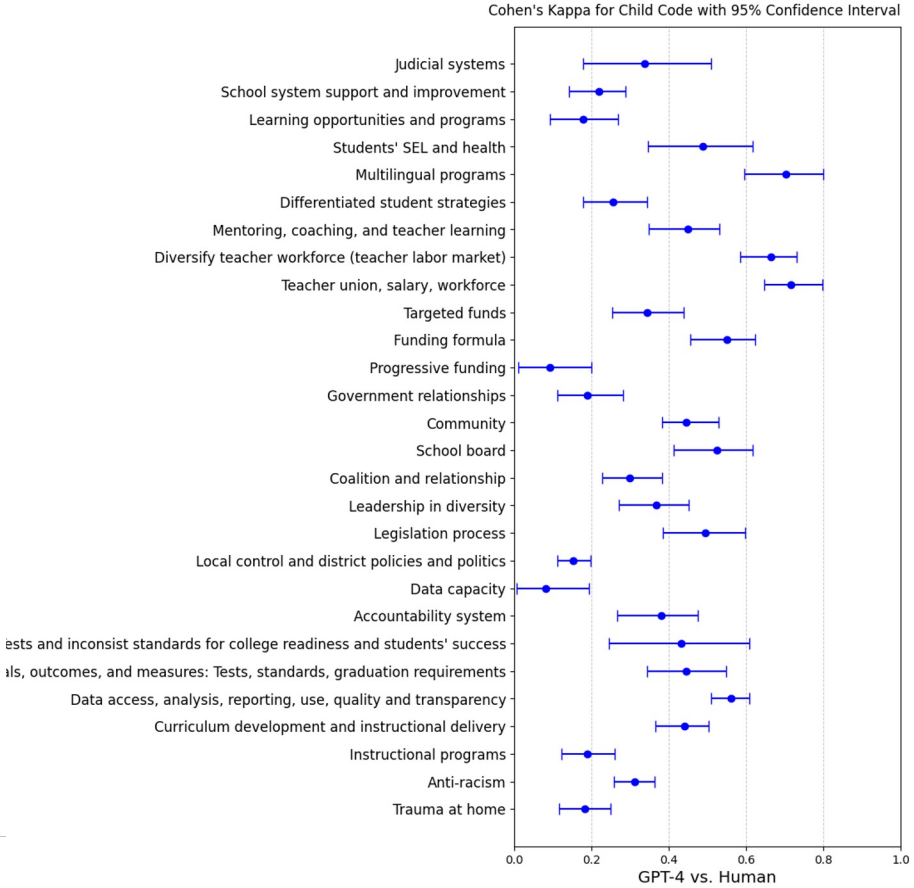
Table 3. Bootstrapped Performance Metrics

	Bootstrapped Performance Metrics (Child codes)			Bootstrapped Performance Metrics (Parent codes)		
	Accuracy	Cohen's κ	AUC	Accuracy	Cohen's κ	AUC
GPT-4 vs. Human	0.9069 (95% CI: 0.9042, 0.9088)	0.3738 (95% CI: 0.3644, 0.3758)	0.7489 (95% CI: 0.7383, 0.7596)	0.7975 (95% CI: 0.7879, 0.8053)	0.4570 (95% CI: 0.4551, 0.4605)	0.7948 (95% CI: 0.7820, 0.8059)
LDA vs. Human	0.8948 (95% CI: 0.8921, 0.8971)	0.1862 (95% CI: 0.1850, 0.1899)	0.6307 (95% CI: 0.6200, 0.6407)	0.7607 (95% CI: 0.7536, 0.7679)	0.2928 (95% CI: 0.2903, 0.2987)	0.6761 (95% CI: 0.6606, 0.6878)

GPT-Human vs LDA- Human Cohen's K by Child Code



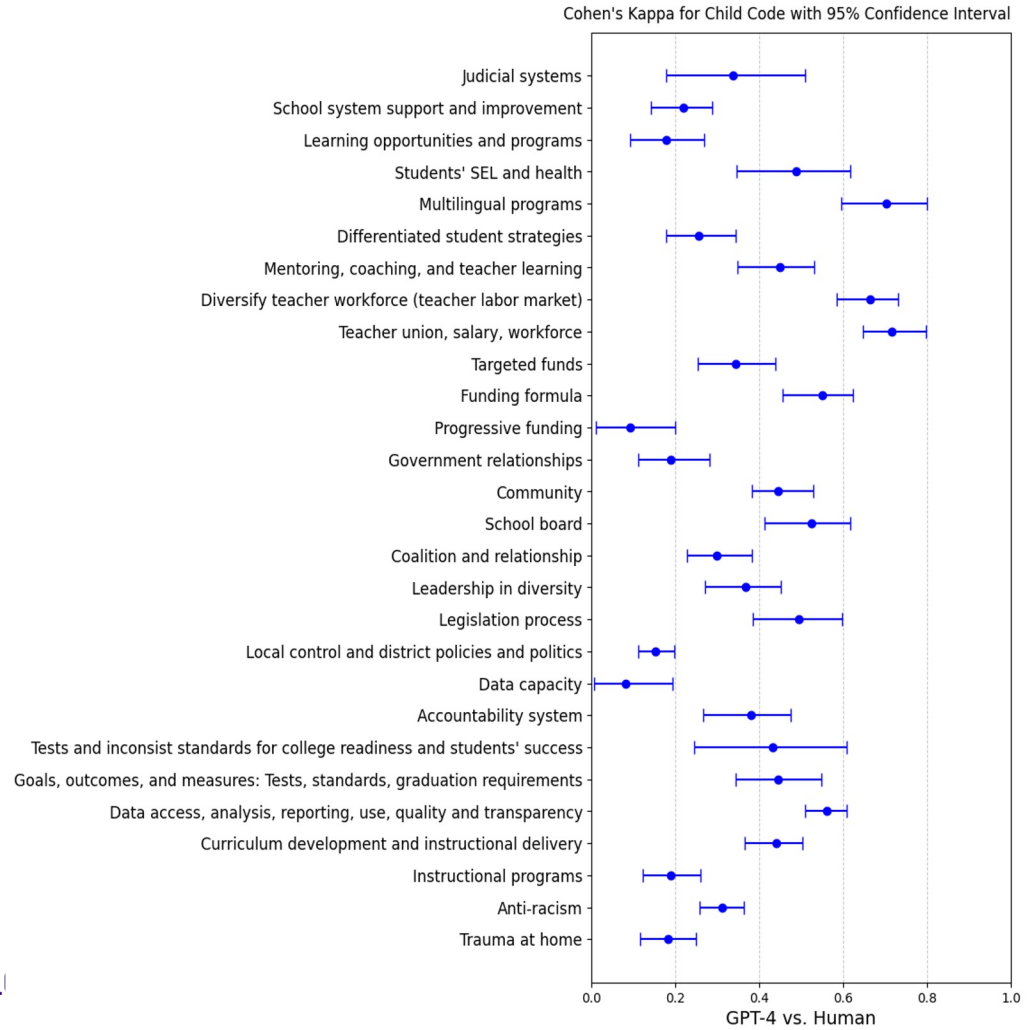
LDA vs. Human



Distribution of GPT-Human Cohen's K by Child Code

The agreement varies greatly by themes.

- Higher agreement on themes that are less domain specific, including multilingual programs; diversity teacher workforce; teacher union, salary workforce; funding formula; school board; data access, analysis, reporting, use, quality, and transparency.
- Low agreement on themes that are more domain specific themes , including progressive funding; local control and district policies and politics; data capacity; trauma at home, instructional programs.



IES Institute of Education Sciences

Discussion

LLMs' performance is sensitive to prompts.

- The utility of LLMs to assist domain-specific data analysis hinges on the integration of domain knowledge to inform prompt development
- GPT-4 classifications are more accurate and valid for themes that are less domain specific.

(LLM vs Human) compares with (Traditional NLP vs Human)

- LLM, to some degree, understand the meaning of the language and contexts, which traditional LDA or lexicon-based analysis are not able to.
- Human experts have theoretical and domain knowledge and lived experience in ed policy.

Sentiment analysis

- GPT and human have a higher agreement on either positive or negative, but lower agreement on neutral.
- Traditional lexicon-based approach couldn't capture domain-specific sentiment.

Code Demo

- STM
- LLM using API

Assignment

> See Github site

Conduct a comprehensive thematic analysis on COVID-related research article titles using topic modeling and LLM.

Data Provided: You are given a dataset comprising 50,000 titles from COVID-related research articles. To facilitate a more efficient analysis, you are recommended to use a random sample for your study. The suggested sample size is as follows:

- For Topic Modeling: Consider using a subset of less than 5,000 titles.
- For Large Language Model (LLM) Analysis: The subset size of titles for analysis may vary depending on the version of ChatGPT or a similar LLM you are utilizing. A smaller subset from the topic modeling results might be necessary.

Research Question: Identify the prevailing topics in COVID-related research articles.

Next Week

- **Classification models:**
  - **Sentiment analysis**
  - **Multi-task classifier using transformer models**
  - **Validation**
  - **RADAR for more accurate and personalized recommendation:**
    - Retrieve
    - Assess (with cosine similarity)
    - Decide (recommend, augment, or generate)
    - Augment (if necessary)
    - Regenerate (from scratch if below threshold)

Sentiment Analysis

Domain-specific definition of sentiment

```
prompt_base = f"""
Act as a policy researcher, you will classify the sentiment in the interviews of educational policy stakeholders as: “Positive”, “Negative”, or “Neutral”. Here is a statement from a policy stakeholder:
[TextGoHere]
To warrant “Positive” sentiment, the statement has to: (1) include the interviewee’s satisfaction about an educational policy (policies) and program(s), or (2) express an enhancement or potential to enhance the quality or equity of student learning or school system, or (3) identify an improvement from past practice. To warrant “Negative”, the statement describes the interviewees’ dissatisfactions, or identifies problems/issues/challenges, or suggests areas needed for further improvement. When the interviewee just states the fact without expressing either positive or negative sentiment, you can classify as “neutral”. When multiple sentiments are observed in one statement, identify the most prevailing sentiment. Explain your reasoning for your analysis."""
```

Lexical-based sentiment analysis:
nltk.sentiment.vader package in Python

Sentiment Analysis

Confusion Matrix and Performance Metrics for Sentiment Analysis

	GPT-4 →			Lexicon →			Performance Metrics		
Human ↓	Positive	Negative	Neutral	Positive	Negative	Neutral		Accuracy	Cohen's κ
Positive	218	4	20	215	11	16	GPT-4 vs.Human	0.58	0.38
Negative	71	322	162	347	151	57	LDA vs. Human	0.31	0.09
Neutral	31	31	215	405	64	43			

GPT-4 is doing much better job than lexicon-based approach.
Agarwal et al. [2019] saw $\kappa = 0.44$ for news sentiment