From Voices to Validity: Utilizing GPT-4 for Policy Stakeholder Interview Analysis

Min Sun[0000-0001-5832-1534] and Alex Liu[0000-0002-4785-1801]

University of Washington, Seattle WA 98195, USA
{misun, aleliux}@uw.edu

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.

Sun and Liu assume equal authorship.

UNIVERSITY of WASHINGTON

Introduction

- 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.
- Stakeholders' voices may be collected from **interviews**, open-ended survey responses, or texts obtained from social media posts.
- The **cost of manually** analyzing even a moderately sized text may hinder the actual use of stakeholders' voices.
- 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.
- 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:

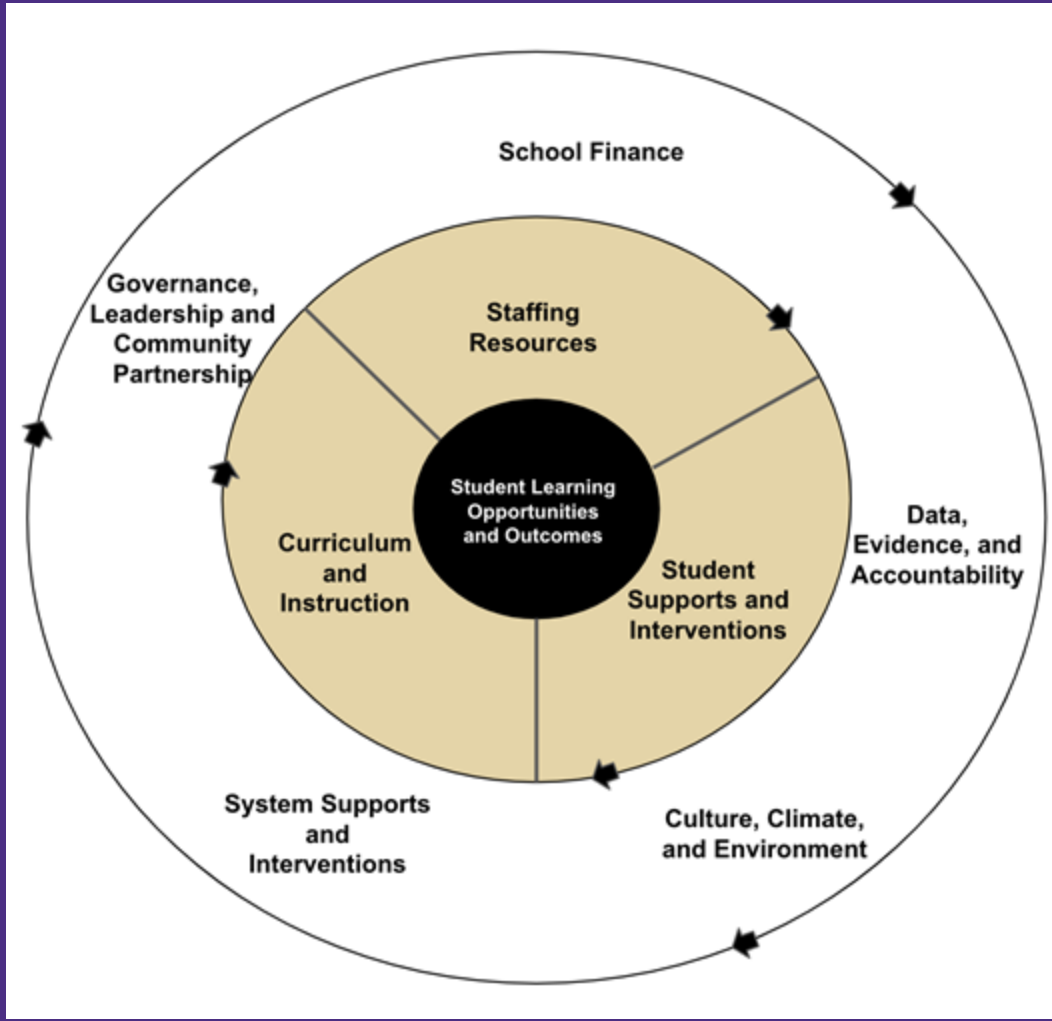
1. What are the **key themes that WA stakeholders voiced** about K-12 public school system?
2. Which themes did stakeholders **recognize as advancing educational equity (positive)**? Conversely, which areas were mentioned **as needing improvement or hinder (negative)** educational equity?

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

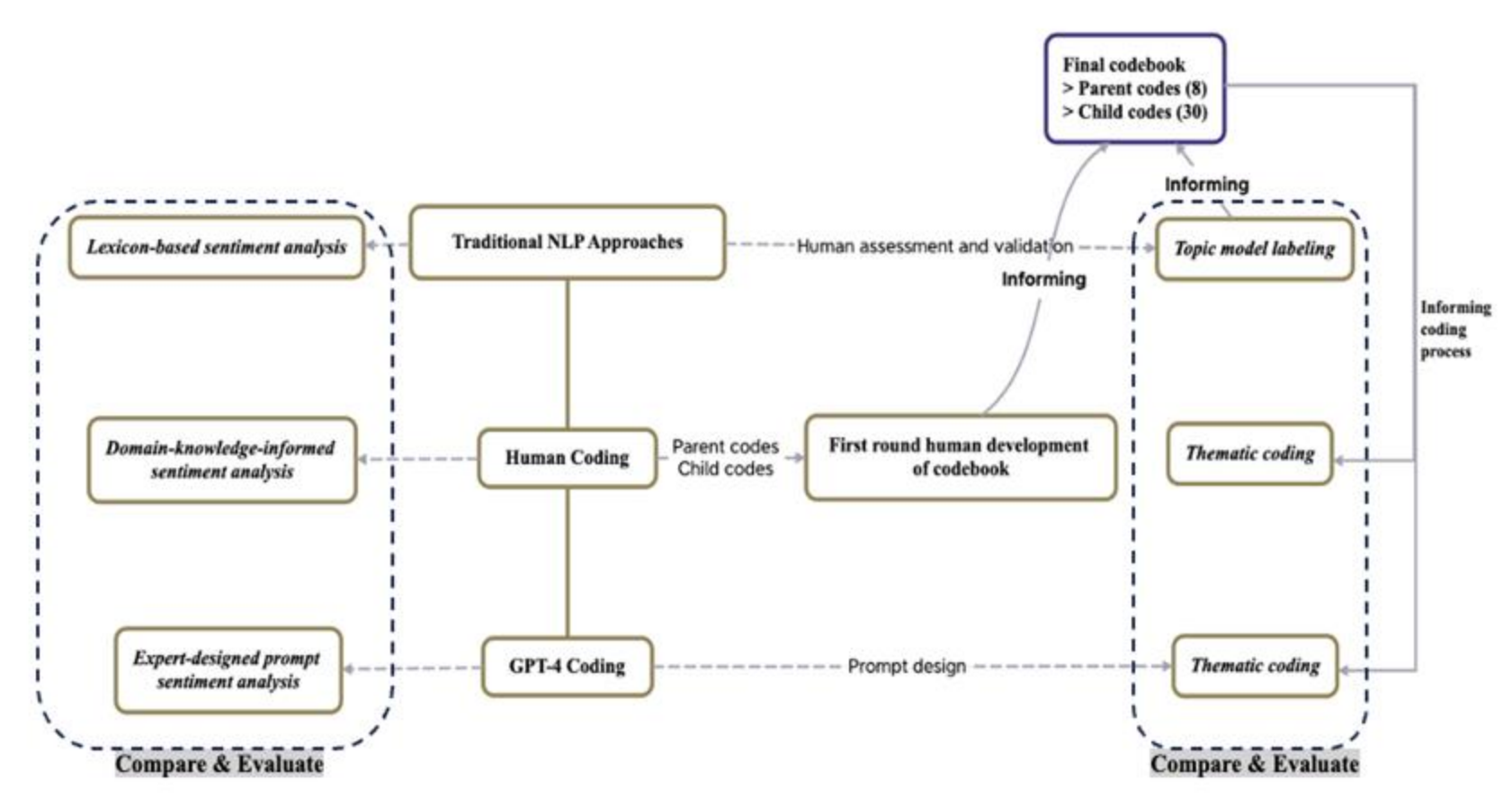
Resources Equity Framework Embedded in a Data-Informed Iterative Improvement Cycle



Data Collection and Preprocessing

- **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
- 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
- Contains interviewees’ research ID, demographics, job roles, and job location.

Methods: Human-Computer Interactive Approach



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

prompt_base2_2 = f"""
Task: As a policy researcher, you've been provided with a paragraph extracted from an interview with an education policy stakeholder. Utilize the provided Codebook (in CSV format) to code the paragraph. The Codebook comprises four columns: 'Parent', 'Child', 'Child_description', and 'Key words'.

Role, context, and overall task

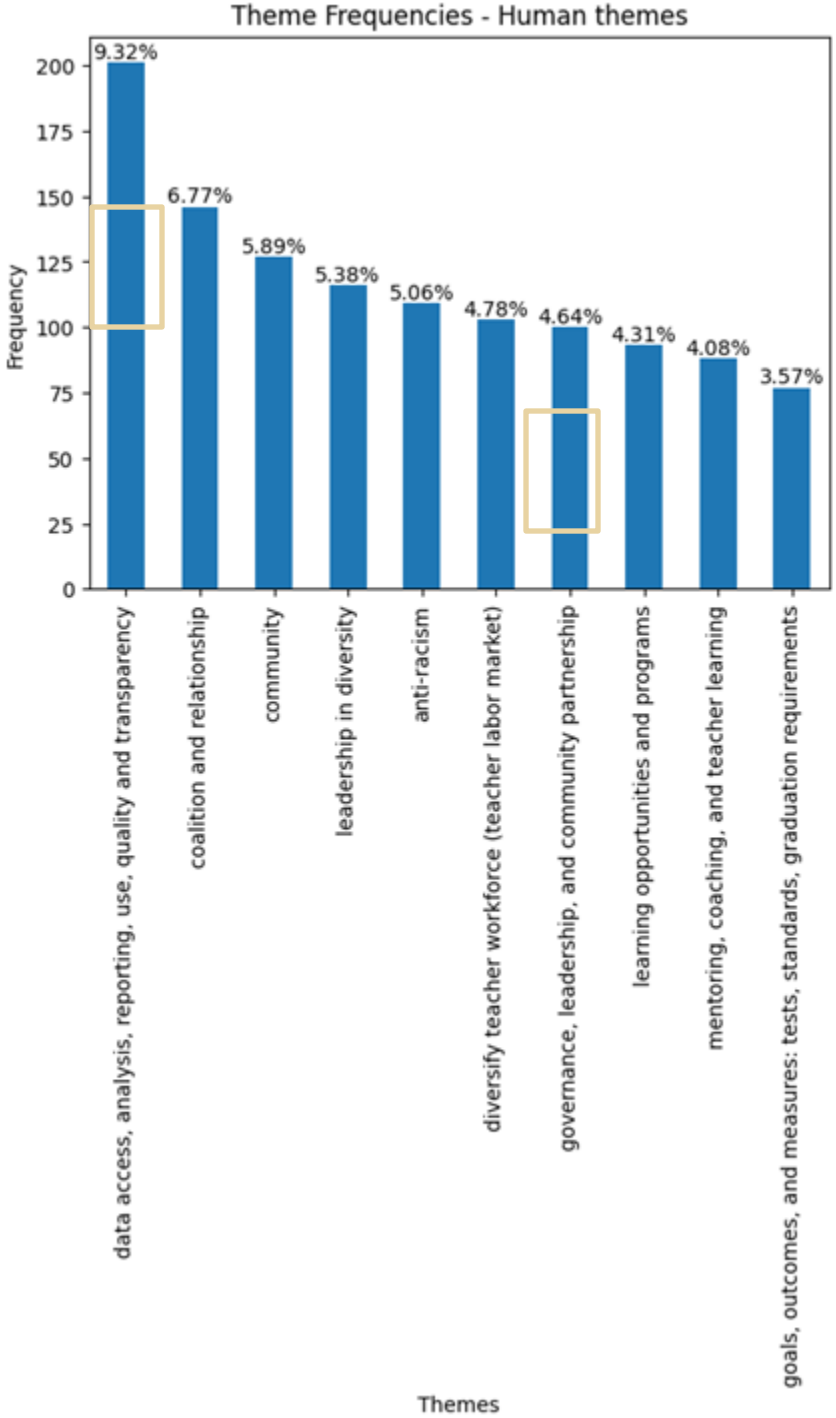
Steps:
1. Identify Salient Themes:
Understand the paragraph's content within the context of the Washington State K-12 public school system.
Refer to the 'Parent' column in the Codebook for broader thematic categories.
Pinpoint up to three salient themes from these 'Parent' categories.
These themes should highlight the most significant ideas in the paragraph.
Label the paragraph with the chosen 'Parent' themes.

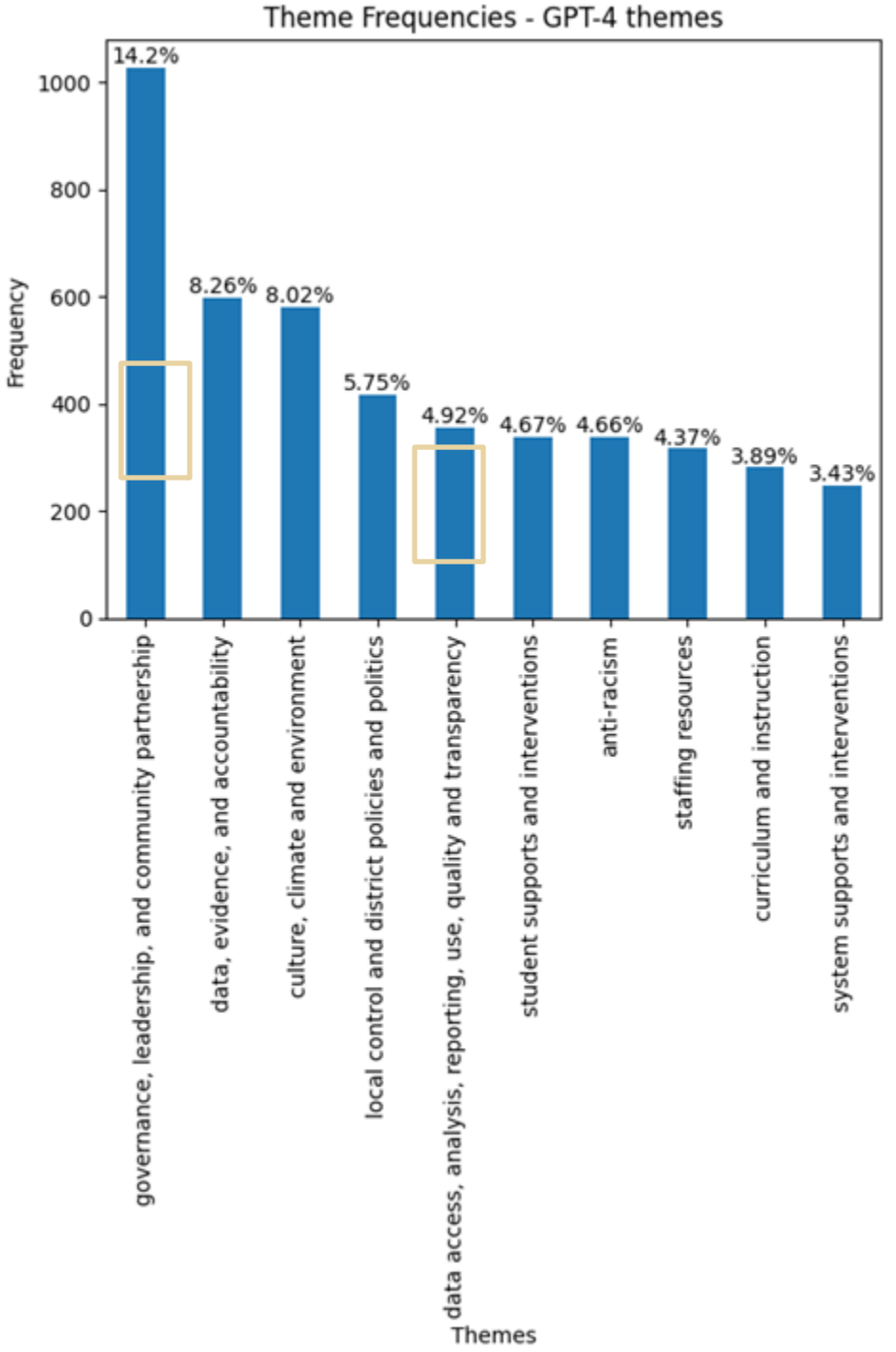
2. Dive into Child Themes:
The 'Child' column in the Codebook lists detailed thematic subcategories, which fall under the broader 'Parent' categories.
The 'Child_description' elaborates on the 'Child' categories, and the 'Key words' column lists pertinent terms for each 'Child' category.

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 apt '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:
{codebook}

Paragraph for Analysis:
[[[TEXTGOHERE]]]


Theme Frequencies - Human themes
Frequency
Themes
9.32%
6.77%
5.89%
5.38%
5.06%
4.78%
4.64%
4.31%
4.08%
3.57%
data access, analysis, reporting, use, quality and transparency
coalition and relationship
community
leadership in diversity
anti-racism
diversify teacher workforce (teacher labor market)
governance, leadership, and community partnership
learning opportunities and programs
mentoring, coaching, and teacher learning
goals, outcomes, and measures: tests, standards, graduation requirements


Theme Frequencies - GPT-4 themes
Frequency
Themes
14.2%
8.26%
8.02%
5.75%
4.92%
4.67%
4.66%
4.37%
3.89%
3.43%
governance, leadership, and community partnership
data, evidence, and accountability
culture, climate and environment
local control and district policies and politics
data access, analysis, reporting, use, quality and transparency
student supports and interventions
anti-racism
staffing resources
curriculum and instruction
system supports and interventions

Child Code Level

	Agreement Metrics		Evaluation Metrics			Bootstrapped Evaluation Metrics			
	% Hit Rates	% Shuffled Hit Rate	Precision	Recall	F1	Accuracy	MEA	Cohen's Kappa	AUC
GPT-4 vs. Human	77.89	17.89	0.33	0.63	0.42	0.9169 (95% CI: 0.9042, 0.9088)	0.0931 (95% CI: 0.0912, 0.0958)	0.3738 (95% CI: 0.3644, 0.3758)	0.7489 (95% CI: 0.7383, 0.7596)
LDA vs. Human	60.65	13.66	0.23	0.38	0.27	0.8948 (95% CI: 0.8921, 0.8971)	0.1052 (95% CI: 0.1029, 0.1079)	0.1862 (95% CI: 0.1850, 0.1899)	0.6307 (95% CI: 0.6200, 0.6407)

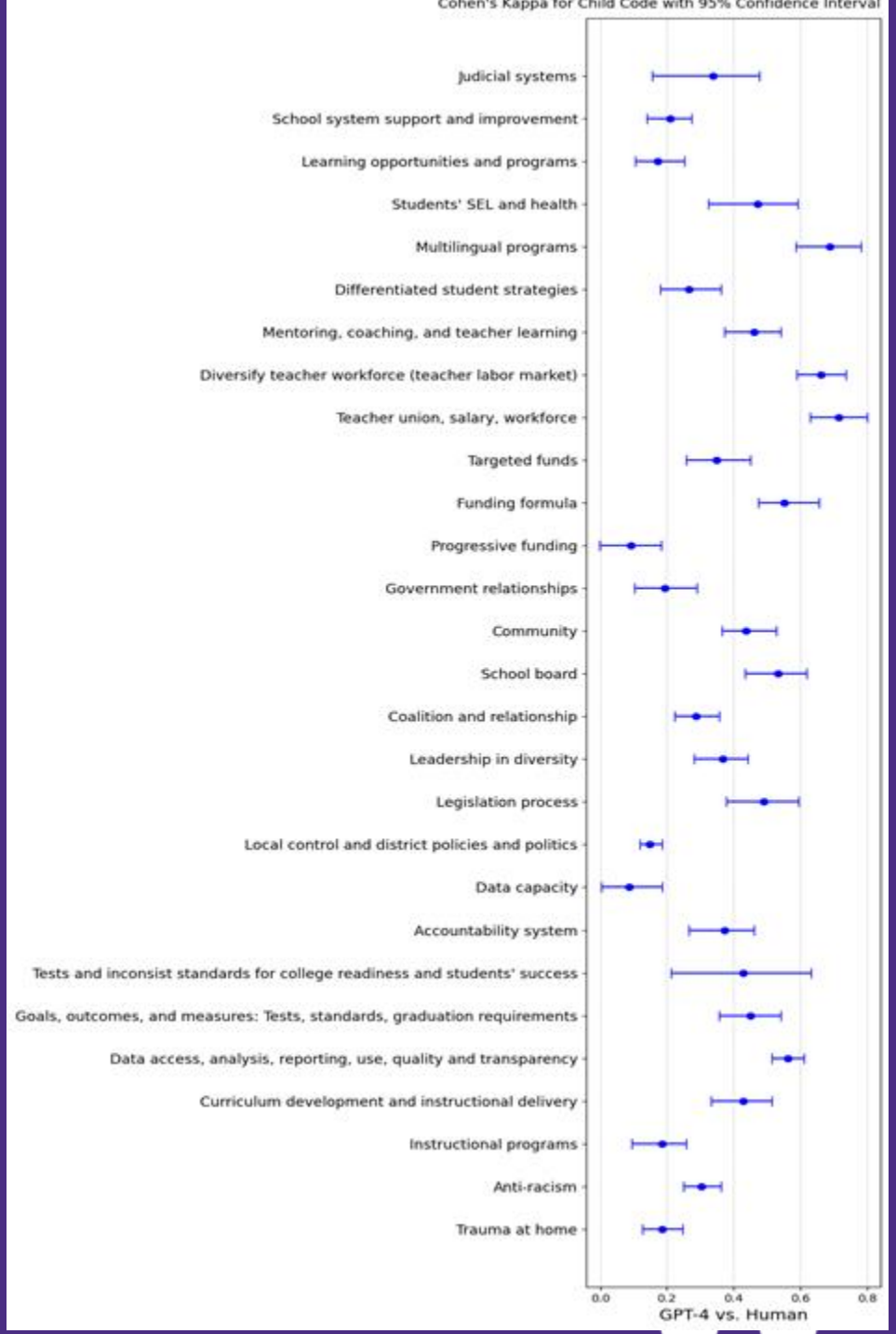
Scholer et al. [2013] reported that human assessors seeing a document for a second time only agreed with their first label 52% of the time.

Parent Code Level

	Agreement Metrics		Evaluation Metrics			Bootstrapped Evaluation Metrics			
	% Hit Rates	% Shuffled Hit Rate	Precision	Recall	F1	Accuracy	MEA	Cohen's Kappa	AUC
GPT-4 vs. Human	96.02	56.67	51.61	86.71	61.83	0.7975 (95% CI: 0.7879, 0.8053)	0.2025 (95% CI: 0.1947, 0.2121)	0.4570 (95% CI: 0.4551, 0.4605)	0.7948 (95% CI: 0.7820, 0.8059)
LDA vs. Human	76.13	47.80	42.97	63.75	48.63	0.7607 (95% CI: 0.7536, 0.7679)	0.2393 (95% CI: 0.2321, 0.2464)	0.2928 (95% CI: 0.2903, 0.2987)	0.6761 (95% CI: 0.6606, 0.6878)

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

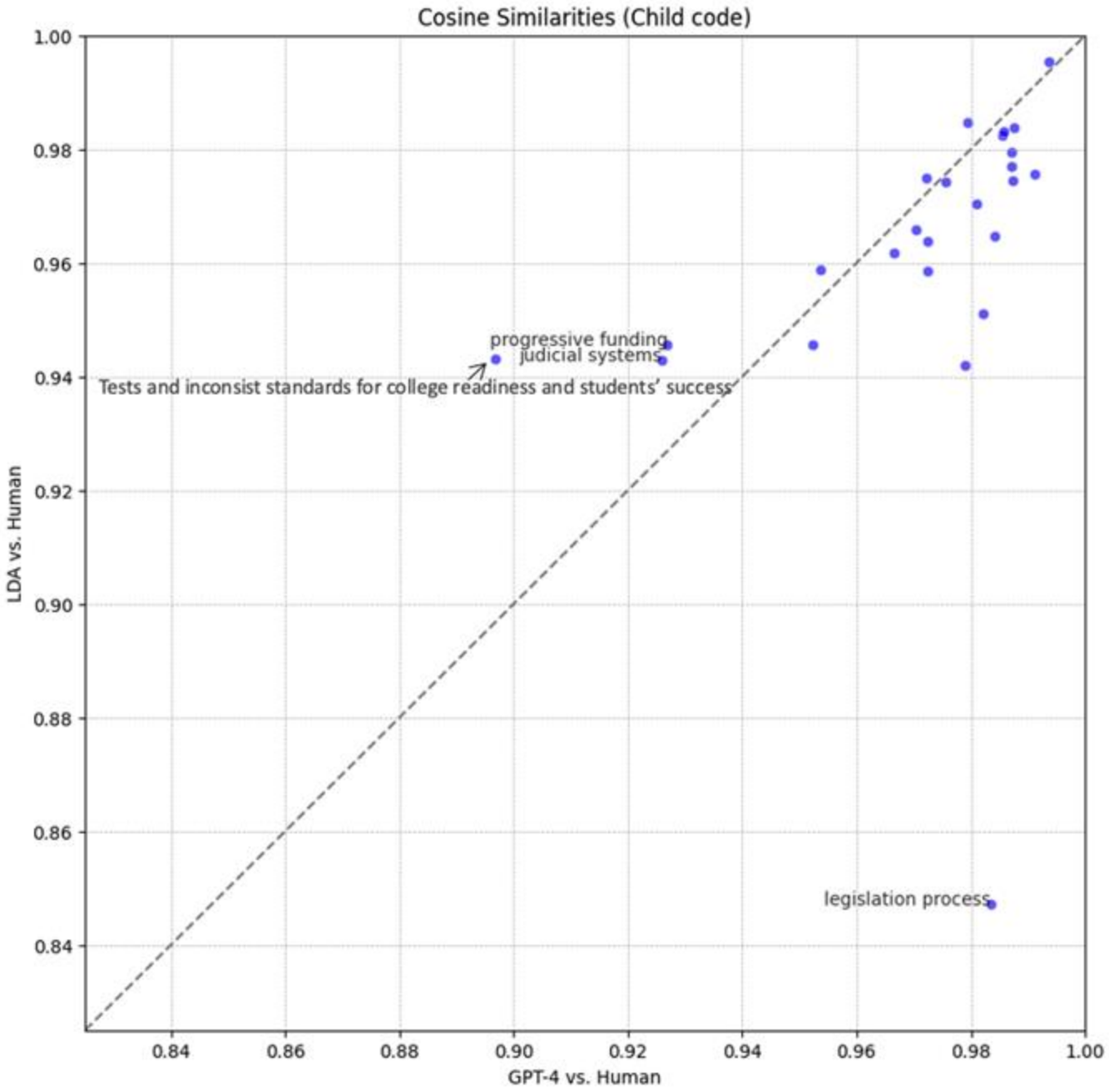
- The agreement varies greatly by themes.
  - Higher agreement on themes that are less domain specific, including multilingual programs; diversity teacher workforce; teacher union, salary workforce; funding formula; school board; data access, analysis, reporting, use, quality, and transparency.
  - Low 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.

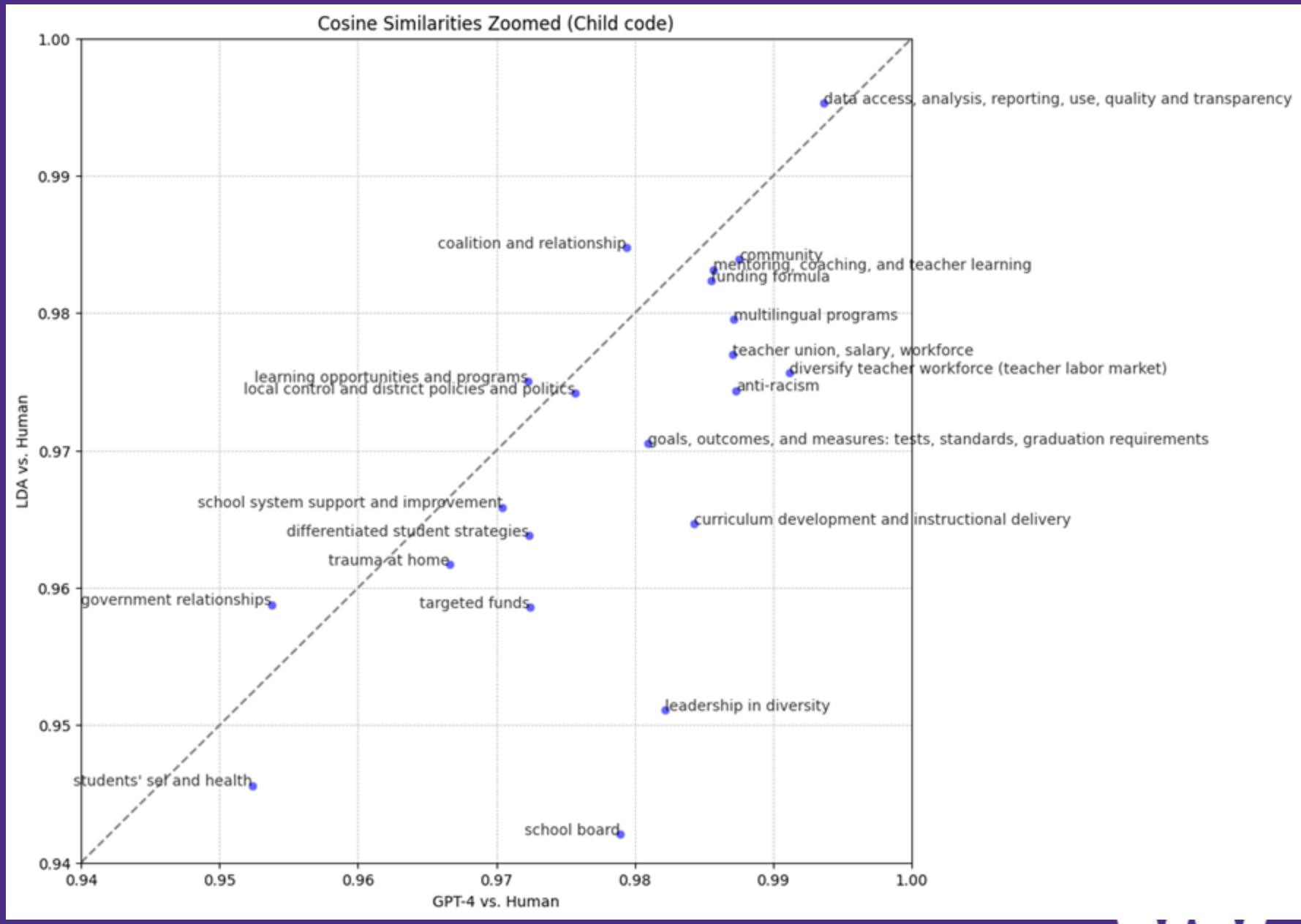


Cosine Similarity

- All "child" themes are highly correlated between machine and human labels

- Further indicating that GPT-4 and human align better, except for progressive funding; judicial system; and tests and inconsistent standards for college readiness and students' success




Cosine Similarities Zoomed (Child code)
LDA vs. Human
GPT-4 vs. Human
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.94
0.95
0.96
0.97
0.98
0.99
1.00
data access, analysis, reporting, use, quality and transparency
coalition and relationship
community
mentoring, coaching, and teacher learning
funding formula
multilingual programs
teacher union, salary, workforce
diversify teacher workforce (teacher labor market)
anti-racism
learning opportunities and programs
local control and district policies and politics
goals, outcomes, and measures: tests, standards, graduation requirements
school system support and improvement
curriculum development and instructional delivery
differentiated student strategies
trauma at home
government relationships
targeted funds
leadership in diversity
students' sel and health
school board

Sentiment Analysis

Prompt:

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: [ ]
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.

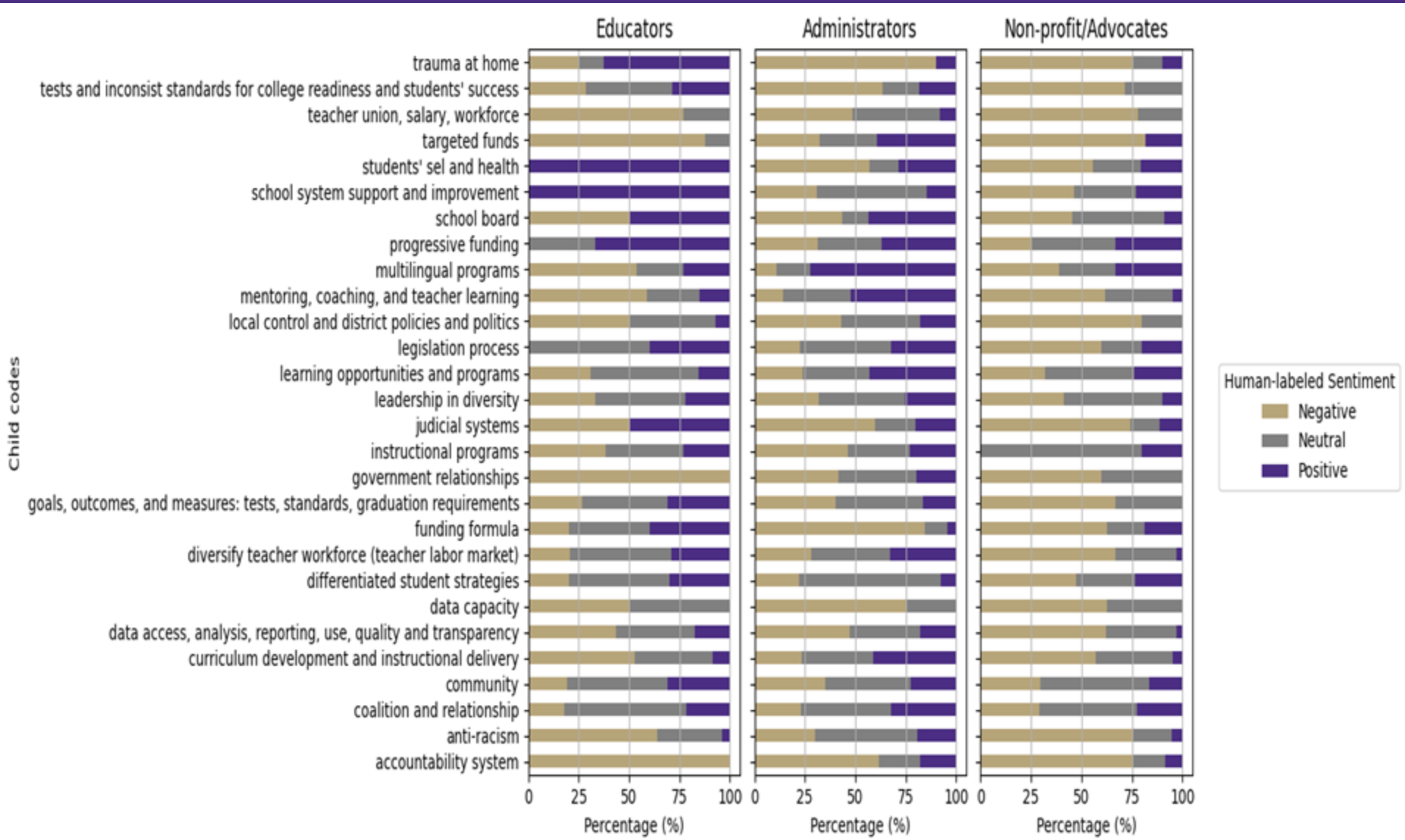
Domain-specific definition of sentiment

Sentiment Analysis

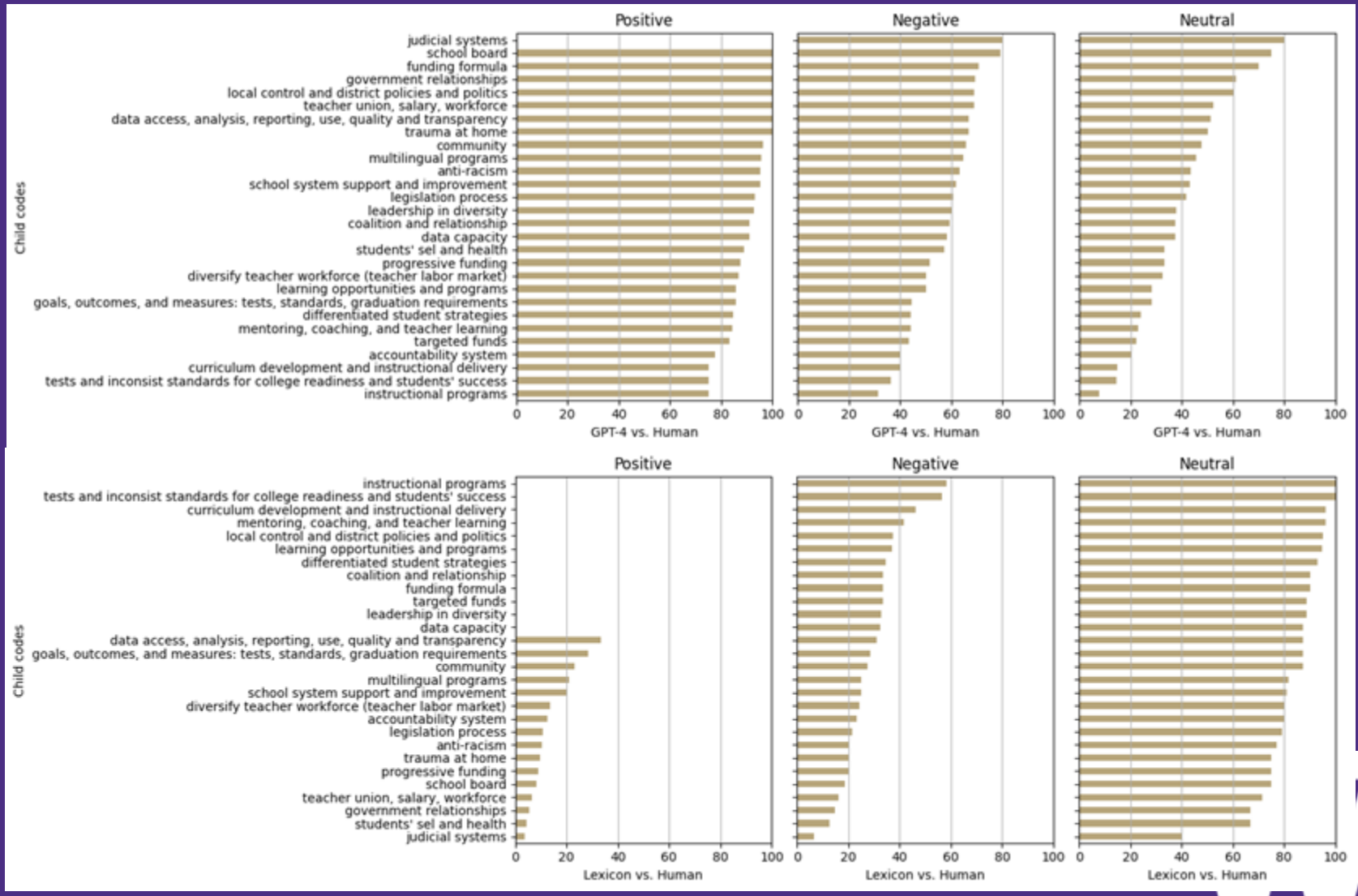
		GPT-4			Lexicon			Evaluation metrics		
		Positive	Negative	Neutral	Positive	Negative	Neutral		Accuracy	Cohen's Kappa
Human	Positive	**218**	4	20	**19**	18	205	GPT-4 vs. LDA	0.58	0.38
	Negative	71	**322**	162	18	**165**	372	LDA vs. Human	0.47	0.13
	Neutral	31	31	**215**	22	59	**431**			

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

Human Sentiment Categories at Child Code Level by Interviewees' Job Roles



% of Human Sentiment Categories Are Accurately Identified by Computer (Child Code)



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.

Back up slides

Introduction

Content validity

Definition:

The degree to which the measure is relevant to, and representative of, the targeted construct it is intended to measure.

Methods:

- Expert reviews
- Relevant conceptual framework and literature

Construct Validity

Definition:

The degree to which the multiple observable measures are related in the ways as intended:

- Convergent validity
- Divergent validity

Methods:

- Correlation

Criterion Validity

Definition:

How well LLM/ML measures perform against a set of "truth" and the utility for policy discourse and action

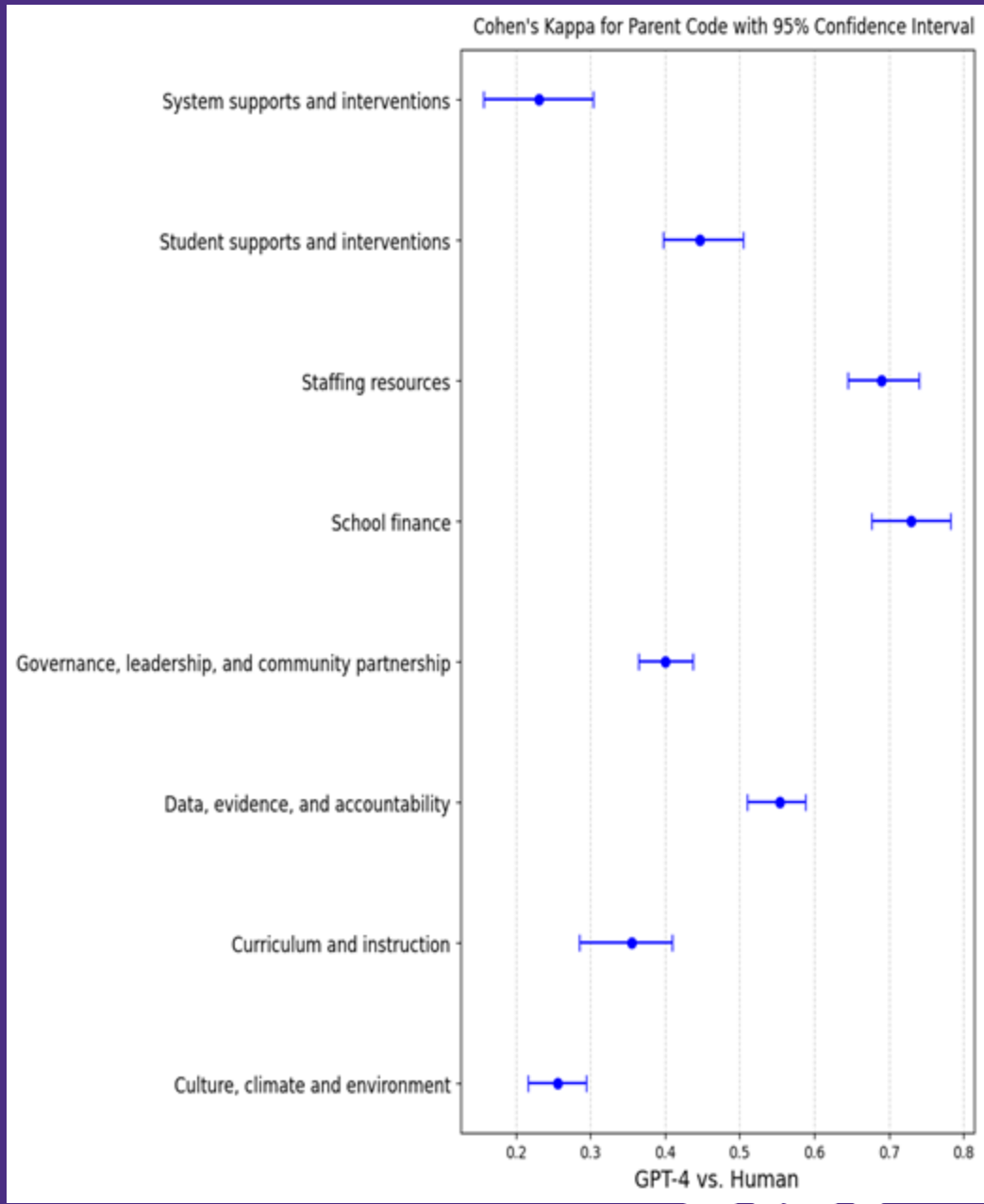
Methods:

- human expert coding
- Stakeholder's perception of the utility to interpret, communicate, and use to take actions

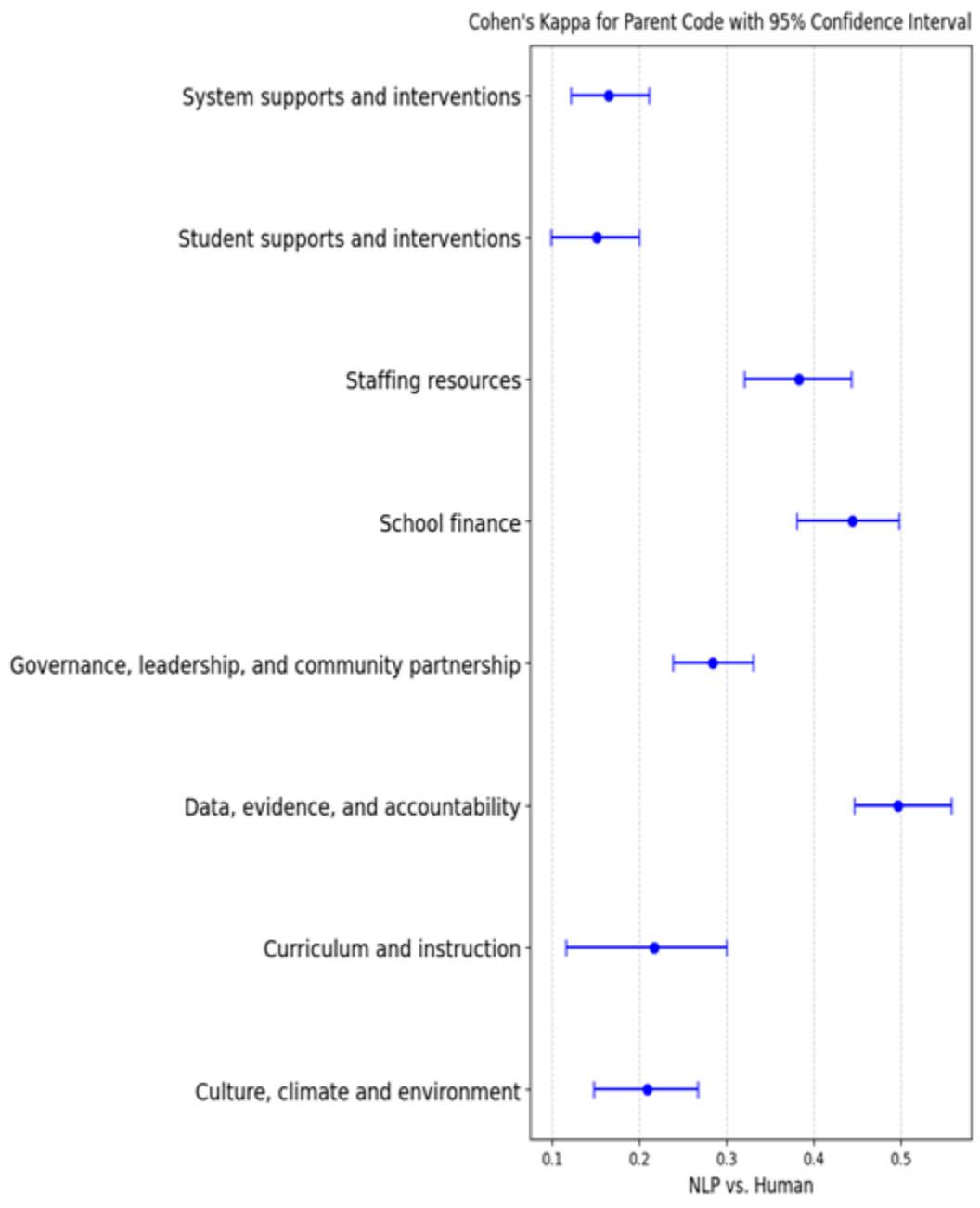
Distribution of LDA-Human Cohen's K by Child Code

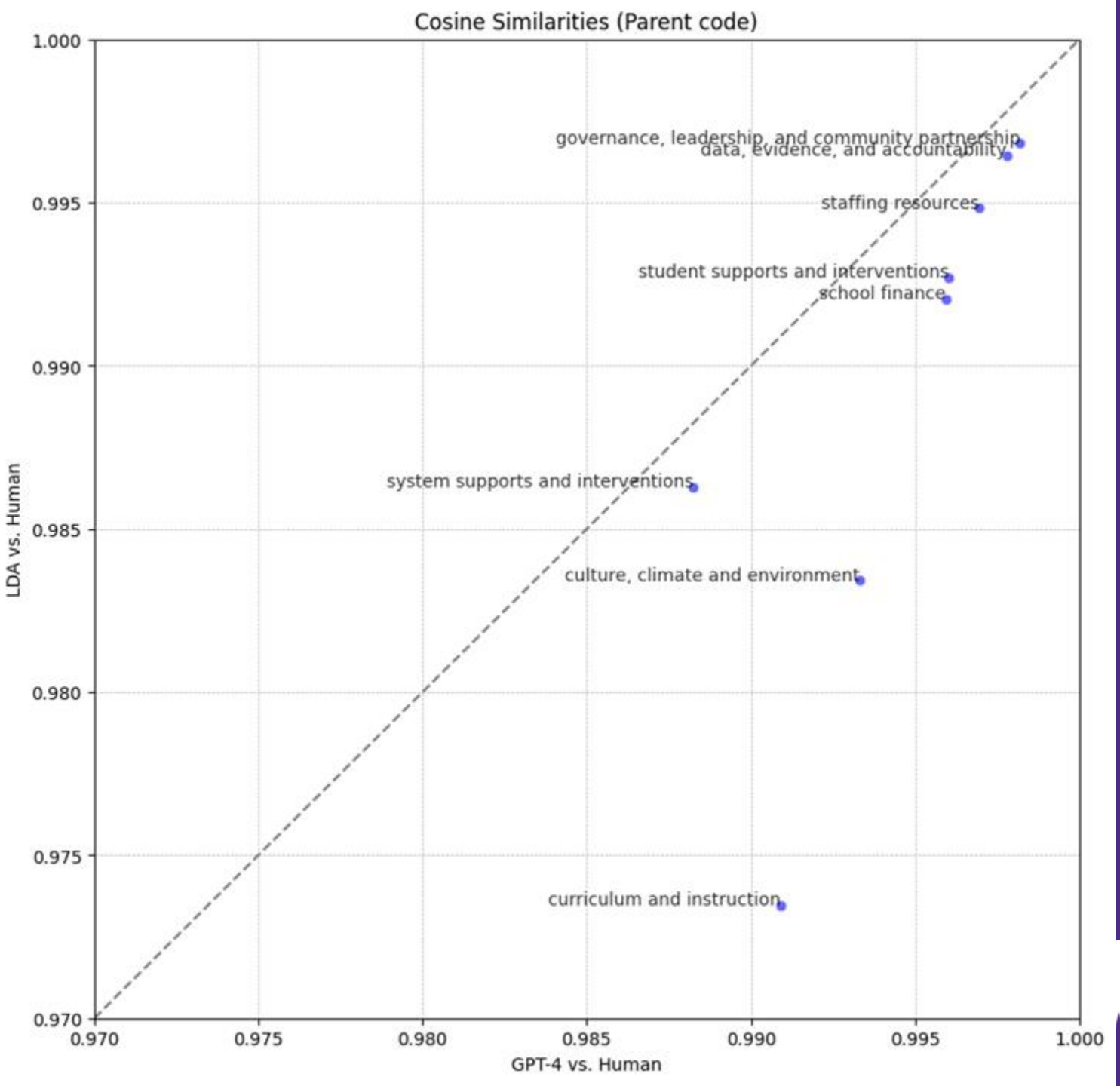
- Overall lower agreement between LDA and human
- Varies by themes:
  - Higher agreement on themes that are less domain specific, including multilingual programs; teacher union, salary, workforce; funding formula; data access, analysis, reporting, use, quality, and transparency.
  - Low agreement on themes that are more domain specific themes , including progressive funding; government relationships; leadership in diversity; data capacity, accountability system; instructional programs.

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

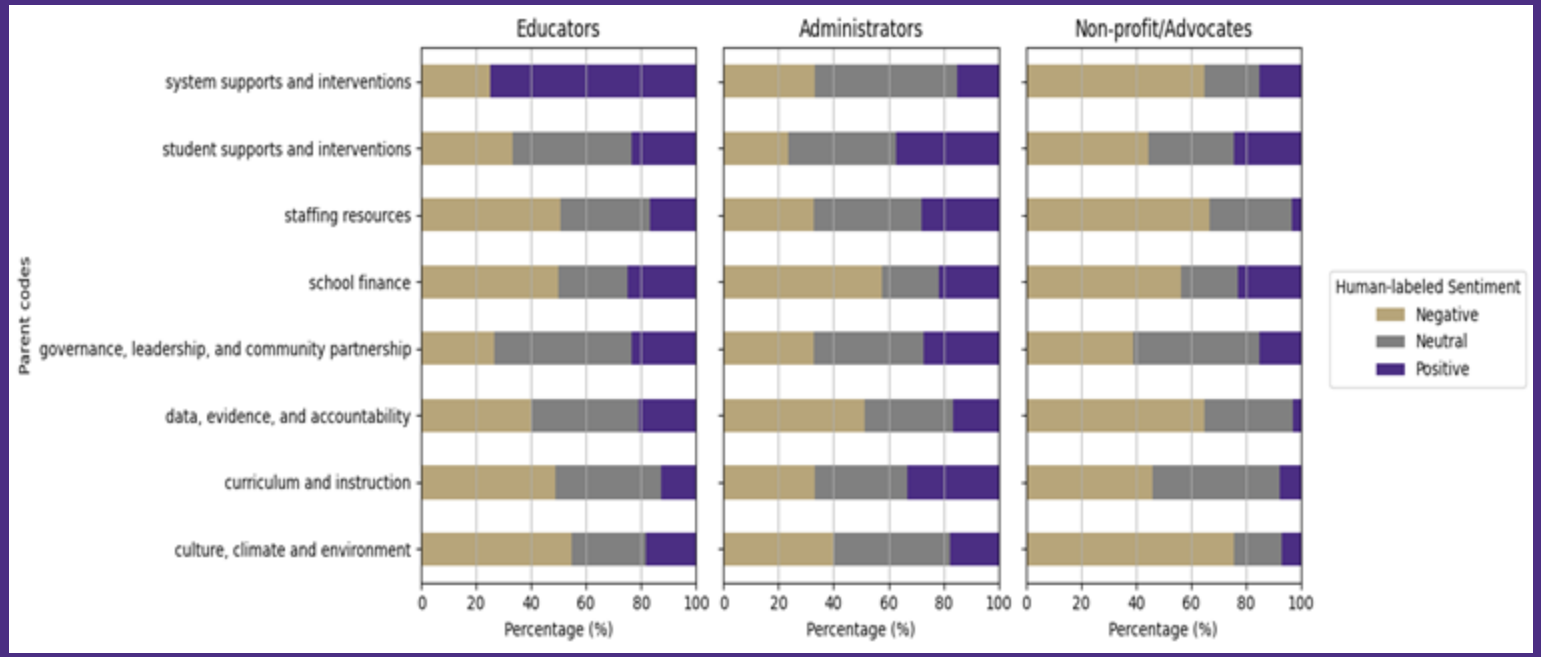


Distribution of LDA-Human Cohen's K by Child Code




Cosine Similarities (Parent code)
1.000
0.995
0.990
0.985
0.980
0.975
0.970
LDA vs. Human
governance, leadership, and community partnership
data, evidence, and accountability
staffing resources
student supports and interventions
school finance
system supports and interventions
culture, climate and environment
curriculum and instruction
0.970
0.975
0.980
0.985
0.990
0.995
1.000
GPT-4 vs. Human

Human Sentiment Classifications at Parent Code Level by Interviewees' Job Roles



% of Human Sentiment Categories Are Accurately Identified by Computer (Parent Code)