Check for updates

AERA Open
January-December 2025, Vol. 11, No. 1, pp. 1–23
DOI: 10.1177/23328584251374595
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2025. https://journals.sagepub.com/home/ero

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

Alex Liu
Min Sun

College of Education, University of Washington

Stakeholder feedback is essential for policymakers to evaluate and develop effective policies, but traditional qualitative analysis methods are often labor-intensive and time-consuming. This study investigates the use of Large Language Models (LLMs) like GPT-4 Turbo (GPT-4) with human expertise to analyze stakeholder interviews regarding K–12 education policy in a U.S. state. The research employed a mixed-methods approach where human experts developed a codebook and iterative prompts for GPT-4 to conduct thematic and sentiment analysis. Results demonstrated that GPT-4's thematic coding achieved 78% agreement with human coding at detailed levels and 96% alignment for broader themes, exceeding traditional Natural Language Processing methods by over 25%. GPT-4 also produced sentiment analysis results more closely aligned with a human expert's judgment. Our qualitative comparisons between human and GPT-4 analysis results highlight the complementary roles of human expertise and LLMs in enhancing efficiency, validity, and interpretability of educational policy research.

Keywords: *artificial intelligence, educational policy, equity, validity/reliability, textual analysis, mixed methods, content analysis, measurements, qualitative research, policy stakeholders, human–AI interaction, large language model (LLM), education policy, natural language processing (NLP), thematic analysis, sentiment analysis*

Policymakers seek reliable, valid, and meaningful evidence to support decision-making in a timely manner. An important source of policy evidence comes from stakeholders' lived experiences regarding the implementation of current policies and their suggestions for improvement (Davidson et al., 2022; Fedorowicz & Aron, 2021). These stakeholders include individuals and organizations concerned with or affected by a policy's creation, enactment, and evaluation. Insights can be gleaned from stakeholder interviews, open-ended surveys, or social media posts (Berry & Herrington, 2013; Rosenberg et al., 2021; Wallner, 2008). Qualitative data on stakeholders' insights, combined with quantitative causal analysis of policy impact, are often utilized in policy analysis. However, when decisions must be made swiftly, the cost of manually analyzing even a moderately sized corpus of text can impede the actual incorporation of stakeholders' voices (Grimmer & Stewart, 2013).

The rapid advancement of natural language processing (NLP) techniques—from traditional rule-based and statistical models to deep learning-based models, such as large language models (LLMs)—is enabling time-efficient comprehension of contents and sentiments in large text corpora. Traditional NLP approaches, including methods like Structural Topic Model (STM), have found favor in social sciences and policy studies for uncovering latent themes in extensive collections of texts, such as political speeches, news articles, and social media content (Blei et al., 2003; DiMaggio et al., 2013; Gao et al., 2023). Policy researchers are also interested in gauging stakeholders' satisfaction about a given policy, perceived intended and unintended consequences, and areas for further improvement. Lexical-based sentiment analysis has been a popular tool for exploring public reactions expressed on social media or other platforms (Abdulaziz et al., 2021; Das et al., 2021).

Recent advances in LLMs have demonstrated impressive abilities to capture textual nuances. Researchers across various fields are examining the performance of LLMs, like OpenAI's GPT-4, in content and sentiment analysis tasks (e.g., Chew et al., 2023; Wang, Xie et al., 2023), noting that model performance varies across different domains and tasks. However few studies have focused on these tools within the context of educational policy analysis. LLMs have been critiqued for algorithmic bias (Benjamin, 2019). Biases related to demographic factors, such as race and ethnicity, gender, and geographic location, are particularly problematic in the educational field which emphasizes equity (Baker & Hawn, 2022). Thus, it is crucial for education policy researchers to explore how to adapt LLMs for

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

their domain-specific needs, while maximizing their benefits. Furthermore, despite the popularity of proprietary deep learning-based models among behavior and social science researchers, given the "black box" nature of these models, researchers must remain cautious of their limitations when employing these tools to make high-stakes decisions in the public policy arena (Shlezinger et al., 2023; Tang et al., 2024; Wulff et al., 2024). Before deploying LLMs in policy analysis, it is incumbent upon us to investigate: To what extent can policymakers and researchers rely on LLMs to analyze stakeholder feedback and opinions in high-stakes, context-dependent decision-making?

This paper is part of a broader study examining policies and programs that advance or hinder educational equity in Washington State's K–12 public school system. The research was conducted in 2022, when public education was simultaneously managing post-pandemic recovery and implementing school finance reforms (SFRs) that stemmed from legislative statutes enacted in 2018–2019 following the McCleary v. State (2012) court decision, mandating changes in educational funding. For this study, educational equity is defined as the reduction of disparities in learning opportunities and outcomes for students from racial-ethnic minority groups and those from low-income backgrounds. To understand the impact of these concurrent changes, the project gathered perspectives through interviews with a diverse group of stakeholders, including state legislators, state-level policymakers, school district administrators, teacher union representatives, teachers, policy advocates, and community leaders. The study leverages cross-district variations to examine how these systemic changes affected educational equity across Washington State. Our study, guided by two sets of research questions, seeks to understand and enhance the application and performance of computer-assisted textual analysis techniques in educational policy studies:

Substantive Research Questions (SRQs)

1. What are the key themes that Washington stakeholders voiced about the 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 that hinder (negative) educational equity?

Methodological Research Questions (MRQs)

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

This study employs multiple analytical approaches, combining human coding with NLP methods including GPT-4, STM, and lexical-based analysis. For SRQ 1, we developed a Resource Equity framework and expert-derived codebook to analyze stakeholder discussions on topics such as data accessibility, governance, and diversity, revealing that theme prominence correlates with stakeholders' professional roles. SRQ 2 examines sentiment patterns regarding educational policies, highlighting areas needing improvement, while acknowledging progress in multilingual education and student support systems.

Our methodological analysis (MRQ 1) shows that GPT-4 achieved 77% alignment with human-coded themes, outperforming STM while offering unique insights despite occasional challenges with overlapping themes. For MRQ 2, GPT-4's sentiment analysis demonstrated strong alignment with human coding, although both machine-led approaches faced challenges with nuanced expressions of dissatisfaction.

In the following sections, we discuss previous studies that employed computer-assisted methods approaches for semantic discovery and coding. We outline the conceptual framework that informed the formation of our codebook, ensuring that the coding process yields results relevant to the policy issue of interest. After describing the data collection and analysis process, we summarize our findings and discuss both policy and methodological implications.

Related Work

Automated content analysis methods have made it possible to discover latent themes and understand underlying sentiments in stakeholders' narratives about given policies by systematically analyzing large text collections. Yet, the complexity of human language and domain-specific contexts suggest that automated content analysis cannot simply replace the nuanced and close reading provided by humans. The outputs of automated text analysis may be incomplete or misleading. Therefore, as the current technology stands, automatic methods should be thought of as amplifying and supplementing careful human analysis (Grimmer & Stewart, 2013). In this section, we review related prior work in GPT, traditional NLP approaches in textual analysis, their performance and limitations, and the unique contributions of our work.

GPT and Textual Data Analysis

Recent studies have investigated using GPT models for thematic and sentiment analysis of text data. These models demonstrate capability in identifying and labeling latent themes and sentiments in text, enabling automated qualitative analysis through deductive coding (Azungah, 2018), which involves either applying codebooks developed by

GPT or those created through human–GPT collaboration for text interpretation (Chew et al., 2023; Dai et al., 2023).

Sentiment analysis has also been performed using GPT. Studies of social media posts and website reviews have shown that GPT-3.5 is capable of understanding sentiment and capturing tones—including sarcasm—and has substantially increased accuracy compared to widely used lexical-based methods (Belal et al., 2023; Kheiri & Karimi, 2023).

However, GPT models tend to focus on certain aspects of the input text, resulting in errors and biases that can be problematic if these biases and errors systematically correlate with people's characteristics, such as gender, education, race and ethnicity, and socioeconomic status (Ashwin et al., 2025). If such biases are introduced during the human–AI collaboration involved in codebook development, they may compromise the validity of subsequent analyses. While prior work has shown that codebooks developed through human–GPT collaboration can achieve quality comparable to those created solely by humans, and substantially outperform GPT-only approaches (Barany et al., 2024), other studies have found that human raters who co-develop codebooks with GPT tend to evaluate passages more similarly to the model than those who were not involved in LLM-assisted codebook construction (Dai et al., 2023). This alignment suggests a potential source of bias. Notably, using a human-developed codebook in conjunction with a GPT coder has been shown to mitigate such effects (Xiao et al., 2023). Therefore, in our study, to minimize potential bias stemming from the "black-box" nature of large language models, we chose to rely on multiple rounds of human expertise without LLM assistance for codebook development.

Furthermore, LLMs like GPT may not possess all domain-specific knowledge, and their performance varies based on the structure and size of the codebook and the type of code. Inter-rater reliability varies depending on the length and complexity of passages analyzed, and the number of codes applied (Chew et al., 2023; Savelka, 2023). These weaknesses present risks for deploying GPT's rapid text processing capabilities in less studied fields to produce impactful results, which underscores the importance of validating the methods before applying them to inform policy making in areas like K–12 public education.

Since the release of OpenAI's latest models—GPT-4 and GPT-4 Turbo—multiple studies have documented significant improvements in performance and functionality compared to previous versions (GPT-3, GPT-3.5, GPT-3.5-turbo; Huang et al., 2023; Lyu et al., 2023; Sprenkamp et al., 2023). Researchers are examining the advantages and limitations of GPT-4 Turbo for deductive coding by testing various prompt types (Liu et al., 2024), finding substantial variability in inter-rater reliability across domains. Research indicates that GPT models' performance is intricately linked to prompt design and phrasing (Cheng et al., 2023; Wang, Liang et al., 2023; Zhao et al., 2021). Systematically designed prompts that specify instructions and tasks for GPT can lead to more accurate and focused responses, reducing randomness, oversimplification, and overlooked information, particularly in expertise-rich content. Carefully designed prompts may even mitigate some of the model's biases (Gao et al., 2023).

Studies on how to enhance information processing using LLMs are proliferating with unprecedented speed. This paper contributes to the current literature by expanding the investigation to a rarely studied yet high-stakes field. Methodologically, this study demonstrates a comprehensive procedure for applying GPT-4 to facilitate qualitative textual analysis, aiming to minimize latent bias and attempt to bridge inherent field-specific knowledge gaps of the model through a grounded development of the codebook by established conceptual frameworks and domain expertise in prompt designs.

Traditional NLP

Topic modeling for thematic analysis. In the realm of traditional NLP, topic modeling—like Latent Dirichlet Allocation (LDA) and STM (Structural Topic Model)—have emerged as a popular tool for the discovery stage of text analysis by extracting latent themes from texts (Jelodar et al., 2020; Leeson et al., 2019). Compared to LLMs, topic modeling maintains advantages in terms of transparency and interpretability, in addition to having well-documented operating procedures. The literature has documented a variety of strategies for validating unsupervised textual analysis results. These strategies often involve: (a) comparing the results with human expert coding of the same data, (b) juxtaposing the results with alternative data sources concerning the same phenomena, and (c) predicting criterion measures.

Researchers have used topic modeling to analyze a senator's self-presentation to their constituents (Grimmer, 2013), and to analyze school reform documents (Sun et al., 2019). In both cases comparisons of machine-derived codes compared favorably to human coding methods. Other work compared grounded theory—an interpretive qualitative method widely used in social science (Glaser et al., 1968)—and topic modeling, finding a mix of similar and complementary insights (Baumer et al., 2017).

In our study, we initially employed human qualitative coding combined with STM topic discovery. The outcomes from the STM were then used to inform the development of the final version of the codebook. This process included examining highly representative documents within topics and extracting high-frequency keywords. Additionally, we assessed the overlapping of STM's topic labels between the results of human coding, and conducted a comparative performance analysis for STM and GPT-4.

Lexical-based sentiment analysis. Lexical models represent one of the major approaches for sentiment analysis, with NLTK VADER being a popular lexical-based algorithm (Hutto & Gilbert, 2014). A recent study that applied

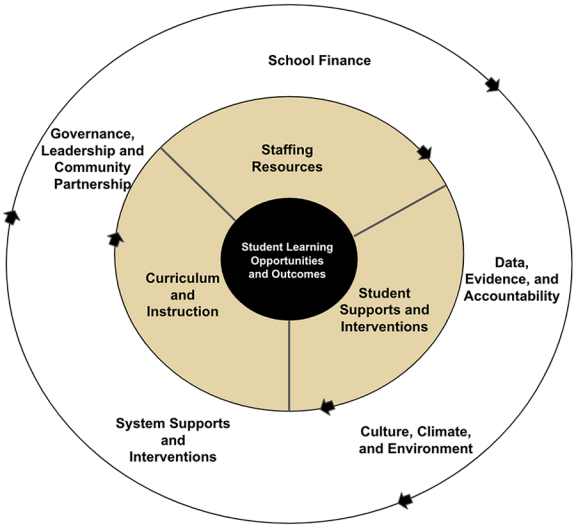

FIGURE 1. *Resources Equity Framework Situated in a Data-Informed Iterative Improvement Cycle.*

seven sentiment analysis tools—Stanford, SVC, TextBlob, Henry, Loughran-McDonald, Logistic Regression, and VADER—to process social media posts and news articles found that VADER was still capable of outperforming the others (Das et al., 2021). Nonetheless, like other rule-based classifiers, VADER exhibits a weaker ability to recognize underlying tones and sarcasm in the absence of non-textual content such as emojis or social media tags (Gosavi, 2022; Nguyen et al., 2021).

The Conceptual Framework: Resource Equity Policy

Drawing on prior research in educational equity policy (Alliance For Resource Equity, 2023), we conceptualize a Resource Equity framework that includes six essential components of educational policies that influence equitable student learning experiences and outcomes in schools (Figure 1). The "inner circle" of policy strategies that are most proximate to students and have direct impacts on student learning includes: (1) *the diversity and qualifications of school staff* (teachers and other adults) who have close interactions with students in schools (Gershenson et al., 2022; Holt & Gershenson, 2019); (2) *the curriculum and instruction* that enable teachers and students to actively engage with rigorous and culturally relevant learning content (Bonilla et al., 2021; Dee et al., 2017; Long et al., 2012); and (3) other types of *student support and intervention* programs, such as mental health and social work services, multi-tiered support systems, summer school, and tutoring, which directly support students outside and around the classroom (Cipriano et al., 2023; Fryer et al., 2020; Guryan et al., 2023).

The "outer circle" of support includes (4) *school finance* that allocates resources to schools to support the offering of educational services (Jackson et al., 2016; Morgan, 2022), (5) *school governance, leadership and community partnership* that determine school decision-making structure and power dynamics among stakeholders (Gates et al., 2019; Wu & Shen, 2022), (6) *the data and evidence* that either enable or constrain the design, implementation, and evaluation of all the previous five components, as well as evidence-based accountability for effective and equitable use of educational resources (Dixon, 2024; Grabarek & Kallemeyn, 2020), and (7) *system supports and interventions* that include system-level reforms to better support students with academic and social-emotional needs, such as whole school improvement efforts (Borman et al., 2016; Dixon, 2024; Schueler et al., 2017). In addition, we acknowledge (8) the *culture, climate, and local contexts* that constitute the environment for student and family experiences inside and outside of the school building (Belton & Brinkmann, 2024; Bryk et al., 2010; Demirtas-Zorbaz et al., 2021).

Policy and practices pertaining to each component at each level of the school system and across the hierarchy of schooling systems are embedded within a continuous cycle of improvement. In this iterative improvement cycle, it is critical to strategically incorporate stakeholder voices from diverse racial backgrounds, professional experiences, and geographic locations in the state.

Data and Sample

Data Collection: Interview Process

The data for this study comprise 24 interviews with diverse educational policy stakeholders. Our purposeful sampling strategy was designed to maximize representation based on the following criteria (Maxwell, 2004; Patton, 1990): (a) the level within public school systems, including classrooms, schools, districts, and the state; (b) geographic locations within the state, including urban, suburb, and tribal schools; (c) roles of interviewees, spanning system actors at various levels of educational systems and three branches of the government at the state level, as well as non-system actors such as community organization leaders, advocates, lobbyists, teacher union representatives, and philanthropic organizational leaders; and (d) characteristics of students and local communities in terms of race/ethnicity, socioeconomic status, language, and homeless populations.

The interviewees were categorized into three primary professional groups: policymakers and administrators (state legislators, state-level policymakers, and school district administrators), educators (teachers and those in coaching or mentoring roles), and non-profit sector participants along with advocates (including teacher union representatives, policy advocates, and community leaders). Interviews were conducted virtually via Zoom, each lasting 45 to 60 minutes.

We intentionally designed the semi-structured interview questions to be broad, facilitating the emergence of a wide

range of topics or deeper insights (Bhattacharya, 2017; see Appendix A1 in the online supplemental material for the interview protocol). Interviewees were requested to provide examples of current state and local policies that they believed most significantly enhance or limit racial and economic equity in Washington State's K–12 public education system. We further explored their reasoning and the mechanisms they proposed. Additionally, we solicited their perspectives on access to reliable data and evidence to support policy development and implementation, as well as their suggestions for iterative policy improvement at the state and local levels.

Preprocessing Interview Data

Audio recordings were transcribed into text with filler words like "um" and "you know" removed. Given the complexity and nuanced nature of our coding categories, which exceeded those of previous studies (e.g., Liu et al., 2022), we divided the interview transcripts into paragraph-length documents to enable more precise coding analysis. This preprocessing step resulted in approximately 1,400 distinct documents for analysis. Each paragraph in our analysis represents a complete thought, which may comprise one or several sentences. To maintain coding objectivity, we stored identifying information (such as interviewees' job roles and locations) separately from the text documents. Additionally, we randomly reorder the paragraphs to eliminate potential bias from sequential patterns or content similarity between adjacent texts, ensuring each paragraph's coding remains an independent decision.

Method: Human–Computer Interactive Learning

Our methods incorporate multi-stage interactions between human and computer to conduct both qualitative and quantitative analysis to examine these four research questions. Figure 2 summarizes the key aspects of the overall workflow.

Codebook Development

First round of human qualitative coding for initial codebook development. Three qualitative researchers, all with education policy research knowledge and extensive experience working in K–12 schools, participated in the first round of coding to familiarize themselves with the interview data and to generate an initial codebook. The team employed a grounded theory approach, iteratively developing a codebook informed by the Resource Equity framework as stated in the Conceptual Framework section. This initial codebook was then used by three expert coders to code the interview data and conduct qualitative analysis for another study within the larger project, which confirmed the content validity of our conceptual framework. This initial codebook also served as the baseline for development of the final codebook used by human and machine thematic coding.

Topic modeling. Concurrent with the first round of human qualitative coding, we employed STM[1] to uncover themes and patterns through latent semantic analysis of word and phrase probability distributions. The analytical process, including the selection of the optimal number of topics, topic labeling, and cross-validation, adhered to established practice guidelines (Grimmer et al., 2022, Ch. 13; detailed information on topic modeling analysis can be found in Appendix B). The topics were labeled in accordance with the Resource Equity framework. To manually validate the themes identified by the topic modeling, we developed rubrics as detailed in Appendix A2. This validation involved scrutinizing 20 documents with the highest top proportions and the 10 most frequently occurring words to interpret the topics' meaning and coherence. Out of 30 topics identified, 25 were deemed both theoretically coherent and practically relevant to the policy interests within the context of Washington State.

Code refinement and final codebook development. Subsequently, we integrated the themes from both the human-developed initial codebook and the STM findings to construct the final codebook. The initial codebook was refined with the help of STM, which provided more structured code labels and identified high-frequency keywords for the child codes. The finalized codebook contained eight broad parent codes that reflect the major themes from the stakeholder interviews and are aligned with our conceptual framework, thereby guarding the content validity of the text analysis by ensuring its results encompass the concepts pertinent to the policy issue of interest. These parent codes include (1) culture, climate, and environment; (2) curriculum and instruction; (3) data, evidence, and accountability; (4) governance, leadership, and community partnership; (5) school finance; (6) staffing resources; (7) student supports and interventions; and (8) system supports and interventions. Within these parent codes, we identified 28 child codes to represent specific topics discussed in the interviews. The final codebook (see Appendix A3) lists parent and child code labels, descriptions, and keywords.

We opted for topic modeling over LLMs in the final codebook development for several reasons. First, to avoid the potential biases inherent in LLM algorithms during the inductive coding phase (Benjamin, 2019), we sought to establish a baseline using the human-developed codebook for subsequent automatic analysis. Second, the topic classification and labeling with STM preceded the final refinements to the codebook; hence, subsequent changes to the codebook did not affect the STM outcomes. In contrast, if we incorporated GPT-informed codes into the codebook, those codes would be used for GPT-4 labeling and might have unduly favored it in thematic analysis. Furthermore, the straightforward interpretability of STM facilitated dimensionality reduction of unstructured text data, allowing for model improvement through parameter adjustments. Although STM's simplicity might overlook subtleties that

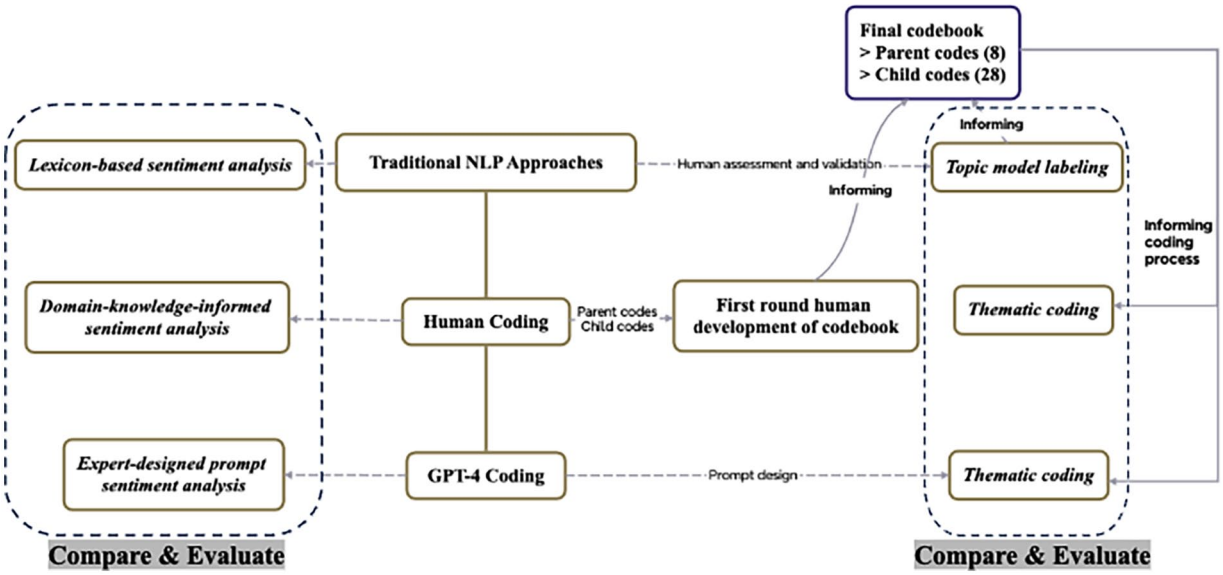

FIGURE 2. *Workflow Diagram.*
Note. This figure shows the flow of our proposed framework utilizing human expertise, GPT-4, and traditional NLP approaches for thematic and sentiment analysis.

LLMs could capture, our human coders compensated for this with their nuanced understanding and domain-specific reasoning. Conversely, LLMs could introduce unnecessary complexity or biases at this stage of pattern discovery, making STM a more appropriate choice for supporting discovery without adding undue complexity.

Thematic Annotation

Human thematic coding as ground truth. To ensure the validity and consistency of the human annotation as the benchmark for evaluating machine-generated results, we conducted two rounds of qualitative coding, involving coders without a machine learning background. Once a codebook was developed, two doctoral research assistants were trained to annotate the entire interview dataset. Both coders had substantial training in educational policy in Washington State and did not participate in the codebook development process. They were introduced to use the codebook to identify up to three most salient themes for each paragraph, from child codes. If no appropriate child code was applicable, they were to select the most fitting parent codes.

During the training phase, both coders independently annotated a shared set of 50 paragraphs. They achieved consensus on over 75% of the coding, with any inconsistencies resolved through discussions aimed at reaching consent for final code selections. Having attained a stable inter-rater reliability, they coded the entire corpus and reported no cases of paragraphs containing more than three themes. The second round of human annotation served as verification. Authors, who were familiar with the interview data and the codebook, reviewed and revised the coders' annotations to increase alignment with the codebook. This process yielded 0–3 codes for each unit of analysis, with 12% at the parent code level and 88% at the child code level. The codes identified by human coders were later used to compare with the topics assigned by the machine for each document.

We used human annotations to address substantial research questions. For SRQ 1, we analyzed the frequency of topic appearance during the interviews and further summarized the themes by stakeholders' job roles. We hypothesized that teachers and teacher mentors would discuss topics such as staffing resources, recruitment, retention, and professional development more frequently due to their direct experiences and relevant knowledge. District and state administrators were expected to focus on issues like school finance and resource allocation, whereas non-profit organizations would likely emphasize the involvement of parents and communities in policymaking. If the analysis confirmed our hypotheses, it would add credibility to the human versus machine annotation comparison.

GPT-4[2] thematic annotation. To achieve optimal results from GPT-4, we developed multiple prompts, adjusting for the number of steps, examples, and instructions. We conducted several rounds of prompt testing using different parameters, manually reviewing the outcomes and the logic of the responses.

Prompt designs. Building on recent studies in prompt engineering, we designed and tested both zero-shot and chain-of-thought (CoT) prompts to reflect our double-layered codebook structure (Kojima et al., 2023; Liu et al., 2022;

Wei et al., 2022). The instructions for GPT-4 mirrored those given to human coders, with the additions of study context and GPT's role to bridge the information gap between GPT-4 and the coders. For all prompt variations, GPT-4 was specifically instructed to consider the context of the Washington State K–12 public school system. After extensive testing, we found that CoT prompts yielded a higher alignment with human annotations (54% for zero-shot vs. 77% for CoT), consistent with findings that CoT prompts enhance GPT's performance (Kojima et al., 2023; Wei et al., 2022).

GPT analysis settings. GPT-4 was set to a temperature of 0.5. GPT temperature is a setting that controls the randomness of responses, ranging from 0 to 1 with a higher temperature allowing for more varied and creative outputs, while a lower temperature results in more predictable and conservative answers. Previous studies employing GPT for semantic annotation often selected a temperature of 0 to ensure reproducibility and low randomness (e.g., Chew et al., 2023; Dai et al., 2023; Xiao et al., 2023), albeit at the cost of limiting GPT's exploratory capabilities. A recent study in automatic qualitative coding educational themes exploring different temperature settings reported better consistency and interpretability using T = 0.5 (De Paoli, 2024). To fully utilize LLM's interpretative capacity while addressing concerns about inconsistent results at higher temperature, we employed three strategies: providing GPT with detailed task descriptions, verifying reproducibility on a subset of 50 randomly sampled paragraphs, and evaluating the randomness by comparing agreement rates with those calculated from shuffled labels. Our tests showed that over 96% of CoT results were reproducible and the shuffled agreement rates dropped substantially from 77.89% to 17.89%, indicating that the GPT labels were paragraph-specific rather than randomly assigned or "hallucinated."

We refined the prompt instructions and tested them on randomly sampled data. Upon reviewing the codes and reasoning from the prompt variants, we selected the CoT prompt format as detailed in Appendix D1. Along with the prompt, we sent the OpenAI API calling GPT-4 model the entire codebook (codebook text replaced {codebook}) and one paragraph at a time (input paragraph replaced "TEXTGOHERE"). Since the OpenAI API does not have memory for prior exchanges, each API call for coding a given paragraph is stateless and independent, which allows for controlled generation conditions. To facilitate GPT-4's step-by-step thinking, we also divided the coding task into three parts. First, in the educational context, behaving as a policy researcher, GPT-4 identified three most salient broad themes for each paragraph only using information in the "Parent" code column of the codebook (p_1, p_2, p_3). Then, GPT-4 learned columns in the codebook that related to child codes. Subsequently, GPT-4 utilized the learned information from "Child" code columns to enumerate suitable child codes under selected parent codes $(p_i c_j, i \in \{1,2,3\}, j \in \{i,...,n\})$. In this step, we requested GPT-4 to use exemplary high-frequency keywords and found that this request improved the GPT-4 performance. After attaining a collection of parent–child pairs, we asked GPT-4 to select three pairs of the most salient themes for the paragraph based on the child code description. If GPT-4 viewed that a given paragraph did not necessarily align with three Child Codes, it could return "None" to fill in the spot for the absent label. For some paragraphs, GPT-4 was unable to identify proper child codes, so it followed the instruction to choose parent codes instead, which mirrored human coders' thought processes and workflow. The output from this prompt gave us up to three parent codes and their corresponding child codes, as well as GPT-4's reasoning for selecting these codes. Appendix D2 shows GPT-4's thought process along this guidance. Here is an example of GPT-4 output:

{"Theme 1": "Accountability system",
"Theme 2": "Instructional programs",
"Theme 3": "Teacher union, salary, workforce",
"Reasoning": "The paragraph discusses the need for accountability in ensuring equal delivery of instruction across different regions (north and south side), which corresponds to the 'Accountability system' theme. It also mentions the delivery of instruction, which fits the 'Instructional programs' theme. Finally, the speaker refers to potential solutions that the union might not accept, which relates to the 'Teacher union, salary, workforce' theme."}

Topic modeling thematic analysis. The STM results, including topic proportions and topic labels, were obtained prior to the finalization of the codebook (see section Method/Codebook Development/Topic Modeling). To be consistent with the structure of human and GPT-4 coding, we assigned the three highest topic proportions labels as the topic modeling codes to each paragraph. These codes are at child code labels as the STM did not directly assign parent codes. Because our topic modeling only produced 25 valid labels that could be matched to child codes in the final codebook, three child codes[3] from the codebook did not appear in the STM results.

Parent and child code level analysis. The codebook's hierarchical structure led to an evaluation of machine performance at both broad (parent code) and specific (child code) theme levels. This involved standardizing annotation outputs, mapping child codes to parent codes, and removing duplicates, resulting in 0–3 codes at each level per paragraph. The methodological research questions were addressed by analyzing outcomes using both child and parent codes, where "original" labels refer to unmodified human or machine-assigned

codes, and "parent codes" refer to all parent-level codes including those mapped from their child codes.

Sentiment Annotation

Human sentiment annotation. Authors conducted sentiment annotations and labeled the sentiment expressed in each paragraph as "Positive," "Neutral," or "Negative." To ensure objectivity and content validity, we closely followed established definitions. "Positive" labels were assigned when interviewees expressed satisfaction with a policy or practice, demonstrated improvement from past practices, or identified policies or practices that have enhanced or have the potential to enhance educational equity. Conversely, "Negative" labels were used when interviewees expressed dissatisfaction, identified issues or challenges, or demanded improvements. When a paragraph merely describes the fact without expressing either "Positive" or "Negative," we code it "Neutral." After establishing common understanding of the definitions, the authors coded the interviews independently and then cross-verified each other's coding. To maintain consistency throughout the coding process, we reached consensus through discussing any paragraph that we were unsure of or that had differential coding or understanding. Specific child codes in the codebook, such as "Progressive funding" and "Tests and inconsistent standards for college readiness and student success," were recognized as inherently carrying sentiment in the context of studying policies advancing educational equity. Hence, we hypothesized a lower frequency of negative labels for "Progressive funding" and a higher occurrence of negative labels associated with "Tests and inconsistent standards for college readiness and students' success" in the sentiment analysis results. This hypothesis serves as a validity check for the human sentiment analysis results.

GPT-4 sentiment annotation. The prompt for GPT-4 was carefully designed to reflect the task performed by the human coding, following the same guidelines (see Appendix D3). GPT-4 received and coded one paragraph at a time. The output included a sentiment label and a brief reasoning for assigning that label. Here is an example output:

> {"Sentiment": "Negative",
> "Reasoning": "The interviewee expresses dissatisfaction with the tendency for teachers to delay teaching linear equations until February, resulting in a three-month gap in the curriculum. They believe that kids could learn them anytime with proper scaffolding, implying that the current practice is not effective or efficient. This statement highlights a problem or challenge in the educational system, warranting a negative sentiment classification."}

Lexical-based sentiment analysis. We utilized the nltk.sentiment.vader package in Python for lexical-based sentiment analysis. This tool is one of the commonly used, best-performing lexical-based tools for sentiment analysis (Abdulaziz et al., 2021; Belal et al., 2023; Das et al., 2021). The interview text was input into the pre-computed algorithm, which then associated words with sentiment scores and compounded the overall sentiment scores for paragraphs, taking into account negations and intensifiers. These compound sentiment scores, ranging from -1 (most negative) to +1 (most positive), were then classified into "Positive," "Negative," and "Neutral" sentiment labels based on default cutoffs.[4] These cutoffs were designed to be generally effective across various texts, as determined through extensive testing and validation by the developers of the tool.

Comparing GPT and Traditional NLP to Human Coding

We applied semantic analysis to explore the underlying themes and sentiments in the interview data and discovered unique patterns that aligned with the interviewees' job roles and responsibilities. The patterns were consistent with our hypotheses, which were derived from education policy literature and authors' understanding of Washington State policy contexts. These findings cast credibility to the human annotation serving as the baseline for assessing the validity of computer-assisted analysis results.

To examine the validity of GPT-4's thematic and sentiment analysis in the context of Washington State K–12 public education, we conducted systematic evaluations across multiple dimensions. We assessed thematic agreement using overlap metrics, such as the hit rate that calculates the proportion of GPT-4-identified codes matching to human-assigned codes out of the total number of human codes (Mathis et al., 2024; Xu et al., 2025). We further compared GPT-4's coding with human coding results using confusion metrics. To ensure that a high overlap between GPT-4 and human codes was not due to overgeneralization or randomness, we also calculated a shuffled hit rate by randomly shuffling GPT-4 labels. A significantly lower shuffled hit rate would indicate meaningful, non-random coding.

Next, we evaluated topic-level alignment between machine and human codings using bootstrapped reliability measures to address class imbalance in the data[5] (Gwet, 2016) and cosine similarity based on TF-IDF to capture text-level similarity (Grimmer et al., 2022, Chapter 7). To adapt binary classification measures for our multi-label classification scenario, we applied one-hot encoding (Dahouda & Joe, 2021), assigning a value of 1 if a paragraph was associated with a given code, and 0 otherwise. Using 1,000 bootstrapped iterations—sampling with replacement to generate datasets of the same size ($n = 100$)—we computed Cohen's κ to account for chance agreement (Cohen, 1960), along

with 95% confidence intervals for each code by comparing human and machine annotations. Additionally, we calculated accuracy (the percentage of correct annotations) and AUC values, which assess a classifier's ability to distinguish between classes and have been shown to be effective evaluation metrics in educational research contexts (Bowers & Zhou, 2019; Çorbacıoğlu & Aksel, 2023; Gilardi et al., 2023). Together, these diagnostic measures provided a comprehensive evaluation of the construct and criterion validity of GPT-4 for text analysis within our domain context.

Results

This section presents the key findings organized by each research question.

SRQ 1. What Are the Key Themes That Washington Stakeholders Voiced About the K–12 Public School System?

Human-annotated code frequencies (original labels) revealed the primary areas of the Washington K–12 public school system as highlighted by stakeholders during their interviews. Table 1 shows stakeholders' extensive focus on data related topics, such as data collection and access, and use of data to help practitioners improve practices and inform policy making. This also includes issues of data sharing, reporting, transparency, and quality ("Data access, analysis, reporting, use, quality and transparency"). Beyond data-related concerns, inclusive governance in the K–12 public education system is one of the central topics, with four of the top five frequent themes related to it. Stakeholders emphasized the importance of meaningful engagement, such as building relationships and centering voices of the marginalized youth and families ("Coalition and relationship") and collaborating with their communities ("Community"). In addition to bringing marginalized communities actively into the conversation, they also advocated for increasing representation of these communities in leadership roles ("Leadership in diversity") to enhance diversity, inclusivity, and anti-racism in the system. Given that the interviews were situated in the context of the K–12 public school system, it was not surprising that "Governance, leadership, and community partnership" emerged as a salient topic.

When we disaggregated our analysis by stakeholders' job roles,[6] we developed a more nuanced understanding of the themes. As illustrated in Figure 3, the human annotation results largely indicated the same patterns of thematic distribution among the three types of stakeholders' job roles: administrators and policymakers, educators, and non-profit advocates. These three types of stakeholders did voice unique concerns related to their daily work, expertise, and lived experience. For example, educators—including teachers and their mentors—concentrated more on "Staff resources," such as teacher education for diversifying the teacher workforce ("Diversifying the teacher workforce and teacher labor market"), and "Student supports and interventions," including learning opportunities in schools and access to curriculum programs ("Learning opportunities and programs"). Administrators and policymakers discussed "School finance" related themes, including "Targeted funds," "Progressive funding" practices, "Funding formula" revisions, along with their work around bills and legislative process for educational policies ("Legislation process"). Nonprofit advocates highlighted governance and community relationships.

SRQ 2. Which Themes did Stakeholders Recognize as Advancing Educational Equity (Positive)? Conversely, Which Areas Were Mentioned as Needing Improvement or Hindering (Negative) Educational Equity?

The study found that very few areas in the Washington state K–12 public education system received a majority of positive sentiment, which might highlight a need for improvement across many aspects of the system. Figure 4 illustrates the proportion of positive, negative, or neutral sentiments that Washington state K–12 public education stakeholders expressed within each topic area, as delineated by child or parent codes.

A larger proportion of positive sentiment was expressed about progressive funding. Additionally, stakeholders commonly acknowledged the efforts and preliminary positive results in student support and interventions, particularly appreciating the reform of multilingual programs. The reform switched previously adopted late-exit programs, which focused on transitioning ELL students from home languages to English, to the multilingualism programs that committed to honor students' cultural and linguistic heritage. As a local administrator illustrated:

> I think the thing that we have been able to do in that shifting, in that transition, is really clarify our commitment to bilingualism. And especially for our families who are second language [speakers], around that is your heritage language, that is the language of your ancestors. That is what connects us to who we are and the generations who came before us, and how important that is.

Such transitions demonstrated WA's K–12 system's commitment to equitable learning opportunities, differentiated support for students with diverse needs, and culturally responsive program design.

Stakeholders also pinpointed persistent issues such as inadequate student social-emotional learning and health support, attributed to lack of funding and insufficient staff resources—issues that became pronounced during the remote learning mode of the COVID-19 pandemic. A representative from a statewide nonprofit organization indicated the dire state of student mental health resources: "It is important to note the lack of necessary mental health resources for

TABLE 1
Human Label Frequency

Parent Codes (Aspects in Parent Codes That Are Not Covered by Child Codes)	Parent Code Frequency (%)	Child Codes	Child Code Frequency (%)
Data, evidence, and accountability	0.32	Data access, analysis, reporting, use, quality and transparency	9.32
Governance, leadership, and community partnership	4.64	Coalition and relationship	6.77
Governance, leadership, and community partnership	4.64	Community	5.89
Governance, leadership, and community partnership	4.64	Leadership in diversity	5.38
Culture, climate and environment	1.16	Anti-racism	5.06
Staffing resources	0.19	Diversify teacher workforce (teacher labor market)	4.78
Student supports and interventions	1.11	Learning opportunities and programs	4.31
Staffing resources	0.19	Mentoring, coaching, and teacher learning	4.08
Data, evidence, and accountability	0.32	Goals, outcomes, and measures: Tests, standards, graduation requirements	3.57
System supports and interventions*		School system support and improvement	3.53
Curriculum and instruction*		Curriculum development and instructional delivery	3.34
School finance	1.35	Funding formula	3.06
Governance, leadership, and community partnership	4.64	Legislation process	2.78
Staffing resources: Any mention related to how state and districts decide or influence the ways staff, train, retain, and support any teacher, paraeducator, school or district leader.	0.19	Teacher union, salary, workforce	2.74
Governance, leadership, and community partnership	4.64	Local control and district policies and politics	2.64
Data, evidence, and accountability	0.32	Accountability system	2.55
Student supports and interventions	1.11	Differentiated student strategies	2.5
Student supports and interventions	1.11	Multilingual programs	2.27
Governance, leadership, and community partnership	4.64	Government relationships	2.18
School finance	1.35	Targeted funds	2.18
Student supports and interventions	1.11	Students' SEL and health	1.99
Culture, climate and environment	1.16	Trauma at home	1.76
Governance, leadership, and community partnership	4.64	School board	1.67
School finance	1.35	Progressive funding	1.58
System supports and interventions*		Judicial systems	1.58
Curriculum and instruction*		Instructional programs	1.44
Data, evidence, and accountability	0.32	Tests and inconsistent standards for college readiness and students' success	1.16
Data, evidence, and accountability	0.32	Data capacity	1.11

Note. All parent code frequencies were measured for original labels rather than converted parent codes from child codes. Parent codes indicated by * were not directly assigned to a paragraph by human coders.

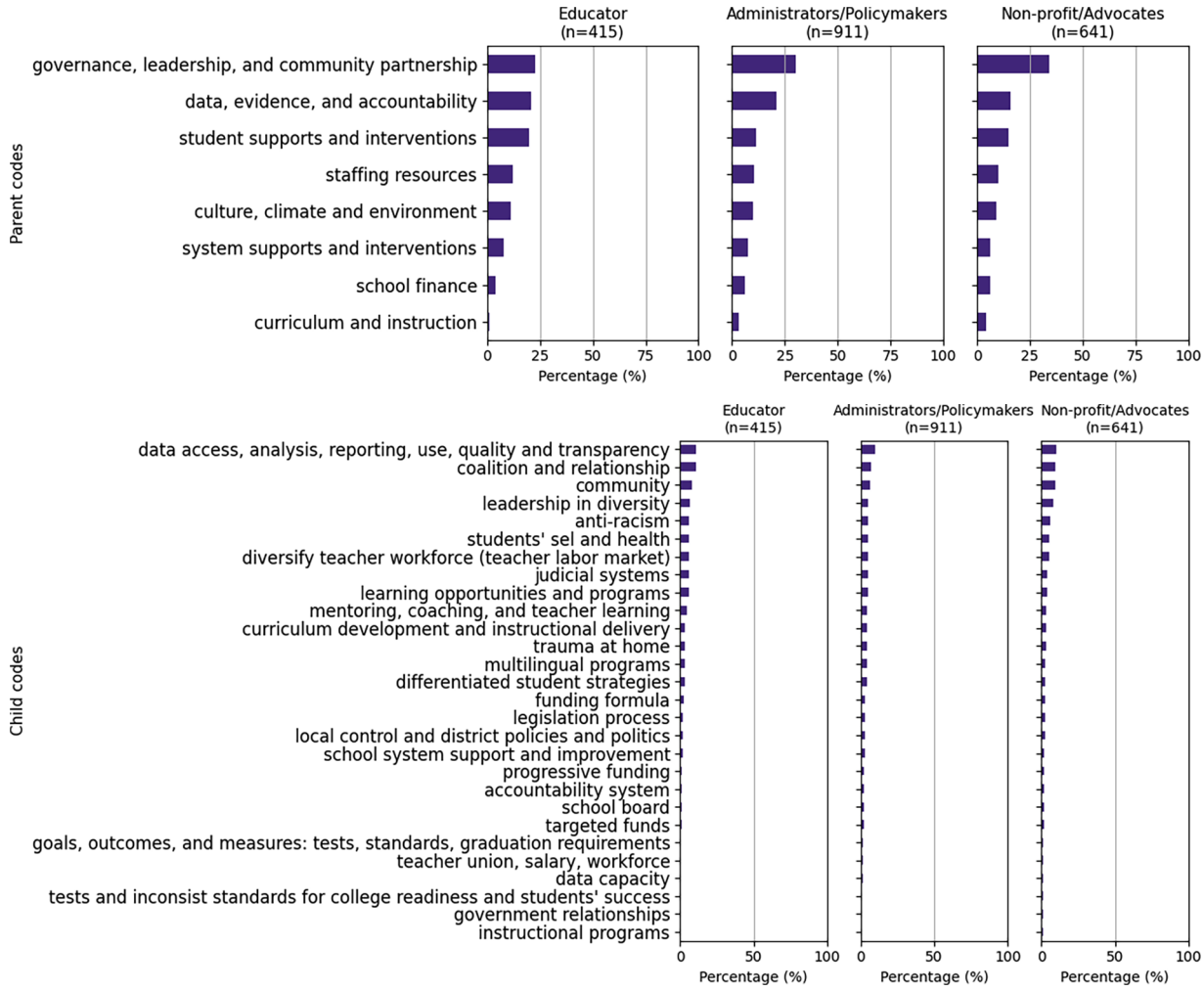

FIGURE 3. *Human Labeled Theme Frequency by Stakeholders' Job Roles.*

Note. Among 24 interviewees, five are educators, 10 are administrators and policymakers, and nine are non-profit advocates. Note that *n* indicates the number of labels assigned by human coders, drawn from interviews with individuals in the specified job roles. For example, out of 415 labels assigned to paragraphs from educator interviews, approximately 24% were coded either directly with the parent theme or with one of its child themes under "governance, leadership, and community partnership." Note that *n* may exceed the total number of paragraphs for a given role, as coders could assign multiple labels to a single paragraph.

students around the state . . . and COVID-19 has raised the impact that the mental health crisis have been having on our students due to the lack of resources available."

In addition to staff resources, the state's data, evidence, and accountability system was critiqued for several commonly recognized concerns. Many stakeholders spoke positively about the volume of data collected and available within the state data system. However, they highlighted a distinction between data collection and data accessibility, critiquing the system's complexity for creating barriers to those lacking data capacity and network connections. Furthermore, inconsistent standards and outcome measures, coupled with an absent accountability system, led to confusion and gaps in feedback loops across various aspects of the system. For instance, this lack of clarity adversely affected educators' ability to deliver an equitable curriculum, as one stakeholder noted: "There should be some accountability that a kid on the north side is going to get the same delivery of instruction as on that south side. And I think that is still a gray area for schools." Additionally, the absence of unified standards and accountability posed challenges in monitoring funding allocations for state administrators, with one commenting, "When you are looking at funding, we have put in almost $10 billion in the last 10 years, but there is really no accountability system to that. We still have 23 different accounting systems that feed into Office of Superintendent of Public Instruction [OSPI], and then they have to figure it out."

Stakeholders' attention to various areas reflected their roles and responsibilities, leading to divergent perceptions of the same issues. Compared to their counterparts in other roles, stakeholders from non-profit organizations and advocacy groups expressed general dissatisfaction with the

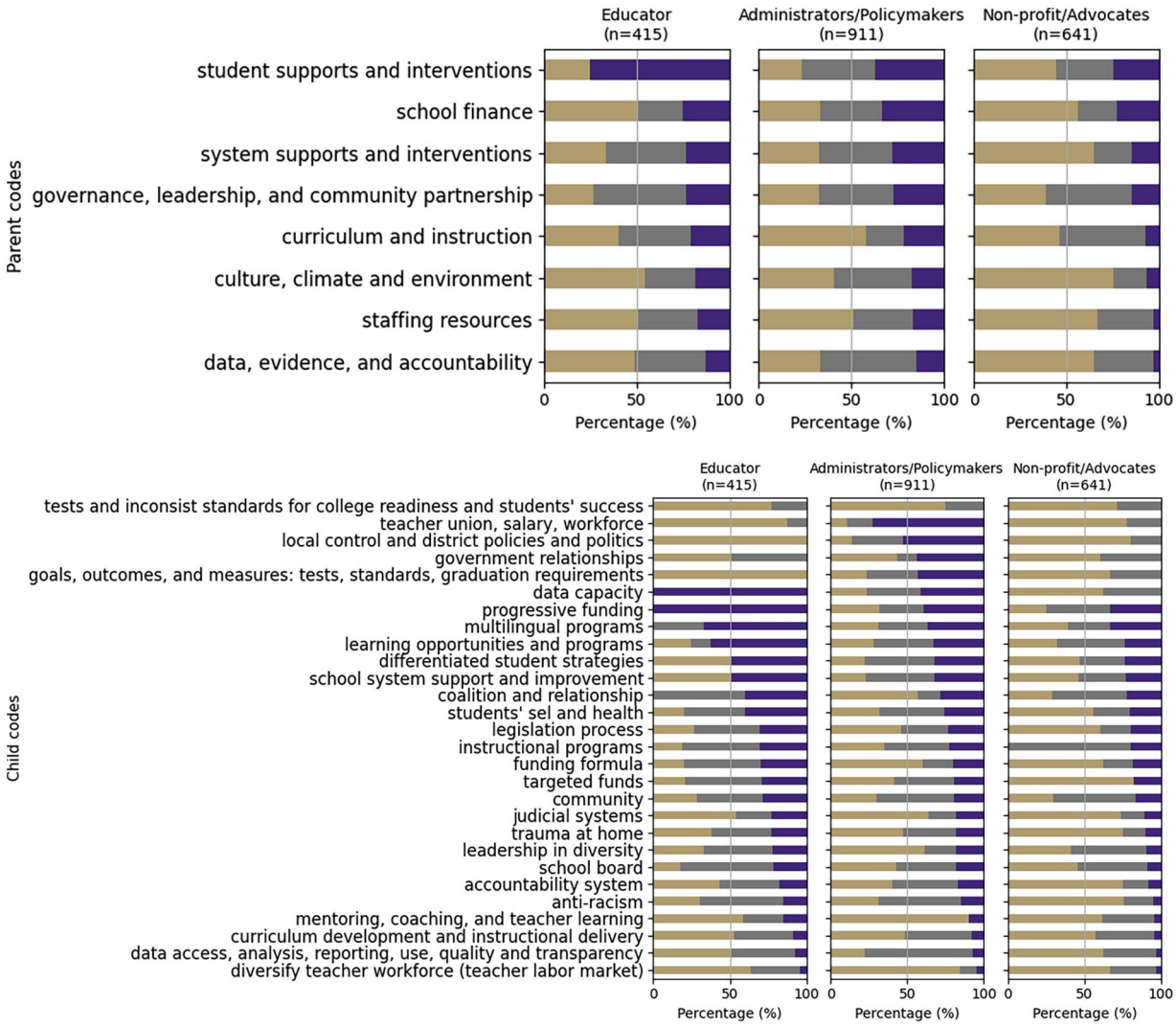

FIGURE 4. *Human Labeled Sentiment by Stakeholders' Job Roles.*

Note. Among 24 interviewees, five are educators, 10 are administrators and policymakers, and nine are non-profit advocates. Note that *n* may exceed the total number of paragraphs for a given role, as coders could assign multiple labels to a single paragraph. For example, out of 415 labels assigned to paragraphs from educator interviews, based on Figure 2, approximately 24% were coded either directly with the parent theme or with one of its child themes under "governance, leadership, and community partnership." Among those paragraphs (415; 24%), this Figure 4 shows approximately 25% expressed positive sentiment, 25% negative sentiment, and the remaining 50% were neutral.

current practices and policies, and called for improvements in numerous areas. This dissatisfaction often stemmed from the system's inequitable treatment of marginalized students, families, and communities. A non-profit representative shared her observation of racial stratification within the system, recounting an instance where Black parents moving into a White community faced discouraging and demeaning challenges. She recalled a parent–teacher conference where the question posed to the parents was a telling one: "What makes you qualify?" She stressed that "we cannot just ask Black and brown people to enter spaces like that without some pretty intense incentive."

These barriers impacted not only community and family access to the system but also influenced decisions related to the recruitment and retention of teachers of color. This caught the attention of state administrators aiming to diversify the teacher workforce and of representatives of teachers of color. Institutional and systemic racism led to a situation where, as an administrator acknowledged,

> There are very few teachers that stay for longer than 5 years. And a lot of them are exhausted. And a lot of the folks that we talk to, they leave because they do not want to be in the racial trauma of being in the school building, educating our kids and trying to decolonize the educational curriculum at the same time.

TABLE 2
Agreement Metrics and Evaluation Metrics for Thematic Analysis

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

Note. In our methodology, both GPT-4 and human coders were instructed to assign up to three codes per text segment. However, a notable difference emerged in how this instruction was implemented. GPT-4 consistently applied all three possible codes, drawing from the full range of thematic possibilities in the data. Human coders, by contrast, typically selected fewer than three codes, often stopping after identifying the most salient themes. This difference in coding behavior contributes to GPT-4's higher hit rate, as it increases the likelihood of overlapping with codes selected by human coders. At the same time, GPT-4's broader coding approach results in lower precision and recall scores, as many of its additional codes—though relevant—do not always align with the more selective human annotations. Importantly, GPT-4 provided reasoning to indicate these additional codes are not necessarily incorrect; rather, they reflect GPT-4's broader interpretive scope and its ability to capture nuances that may be overlooked by humans. We emphasize the hit rate metric in our analysis because it best reflects GPT-4's effectiveness in identifying themes that align with human judgment. This measure illustrates how GPT-4 can complement human qualitative analysis by surfacing a wider array of relevant themes that enrich the interpretation of complex textual data.

Educators recognized the limitations of policies concerning teacher unions, salaries, and the workforce. They also articulated a need for policies and practices to be designed and implemented with the aim of supporting students' academic and social-emotional learning. Furthermore, practitioners suggested that the goals of such policies should be evaluated during implementation, rather than simply requiring compliance under the umbrella of local district control or complex government relationships. As one educator put it,

> I know that a couple of years ago the district came out with an equity policy. I think what is very interesting is how that actually plays out in leadership in classrooms. It felt like a very formal or not formal, but just like a checkbox. We did the thing, we wrote the thing we are going to abide by these policies that are very vague and not very specific.

MRQ 1. How Accurate and Valid are GPT-4 Labels of Key Themes When Compared to Human Experts' Labels and Traditional Topic Modeling Results?

Evaluation for GPT-4 thematic analysis. To evaluate the overlap between machine and human coding, we calculated hit rates, which measured the percentage of machine-labeled themes that corresponded with human-labeled themes. Table 2 illustrates that, on average, approximately 78% of GPT-4 annotated child codes matched those annotated by humans for each paragraph. Given that up to three child codes were assigned to each paragraph, these results indicate a significant overlap at child code level about detailed themes, with GPT-4 and human coders identifying at least two identical themes per paragraph. The STM approach also demonstrated a high rate of overlap, aligning with human annotations for more than half of the child codes. Hit rates for both computer-assisted methods increased when comparing the broad themes (parent codes). In particular, GPT-4 parent code labels nearly fully cover the human experts' coded parent codes.

To address potential random matches in GPT-4 outputs when the temperature setting is higher than 0, we calculated shuffled hit rates by comparing machine annotations for a paragraph with human annotations from another randomly selected paragraph. As shown in Column 3 of Table 2, the decrease in shuffled hit rates suggest that both GPT-4 and STM were capable of discerning paragraph content and assigning labels based on specific content. To test the robustness of our evaluation to different evaluation metrics, we also calculated Szymkiewicz-Simpson coefficients, Sørensen-Dice coefficients, and Jaccard similarity indices. The results of these measures, presented in Appendix C Table C1, are consistent with our findings in Table 2, that GPT-4 surpassed STM in accuracy for both child and parent level annotations.

Furthermore, we assessed GPT-4's annotation performance using measures derived from the confusion matrix. Notably, at both the child and parent code levels, GPT-4 and STM exhibited high recall rates compared to precision. This implies that themes identified by human coders were likely to be recognized by the machines, but not all themes suggested by the machines were confirmed by the human coders. The relatively low precision could result from the different output formats in that human coders could choose between zero to three labels per paragraph flexibly; in contrast, the algorithms constrained the STM to always assign

TABLE 3
Bootstrapped Performance Metrics

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

Note. This table presents the overall evaluation metrics. We also provide code-wise evaluation metrics in Appendix C (Figures C6–C9), which illustrate variations in performance across different topics. These figures highlight differences in GPT-4's alignment with human coding depending on the thematic category.

three themes and GPT-4[7] to identify three salient themes in most cases. Therefore, machine annotations naturally included more labels, leading to an increase in false positives. Nevertheless, the high recall rates confirmed claims from agreement metrics, illustrating that GPT-4 was adept at capturing human-identified themes and outperformed STM in this respect. Additionally, the low precision might indicate that GPT-4 was able to uncover nuances in text analysis that had not been detected by human coders.

Performance varied across thematic contents. As indicated in Table 3, GPT-4 generally outperformed STM on average, at individual parent code level, and for the majority of child codes (see Appendix C, Figures C2–C5). The average Cohen's κ at the child level was comparable to deductive coding using an expert-informed codebook in linguistics (Xiao et al., 2023) and was better than deductive coding using a GPT-informed codebook for an open-access dataset (Dai et al., 2023).

Our parent-level statistics showed better performance. When comparing GPT-4 to human annotations for each code, Cohen's κ exceeded 0.4 for five parent codes and surpassed 0.7 for two parent themes "Staffing resources" and "School finance." Although Cohen's κ for STM and human annotations for child codes ranged between 0–0.5, the agreement on the theme of "Data access, analysis, reporting, use, quality, and transparency" exceeded 0.6, outperforming the agreement between GPT-4 and human for this theme. For the remaining child codes, GPT-4 demonstrated higher concordance with human annotations than STM, especially for themes such as "Multilingual programs," "Diverse teacher workforce (teacher labor market)," and "Teacher union, salary, and workforce." Similar patterns were observed in the codewise AUC, which measures a machine's ability to differentiate true positives from false positives for a theme and is unaffected by data imbalance (see Appendix C, Figures C6–C9).

In Appendix C10–C12 we present comparisons of human and machine coding at the text level and find high cosine similarities for both GPT-4 and STM. Summarizing all the evaluations, it appears that GPT-4 is capable of identifying themes from context-rich interview data, despite the complexity of the tasks posed by the hierarchical structure of our codebook and the large number of codes. GPT-4 recognized the same themes that were identified by human coders and also picked up on themes that were not selected by human coders. These deviations could potentially enrich the human interpretation of the text, adding nuance to the semantic analysis. However, agreement varied significantly across different themes. In general, GPT-4 performed better at identifying broader themes (parent codes) than more specific themes[8] (child codes), and was more adept at recognizing less domain-specific themes than more domain-specific ones.[9] Across both child and parent code levels, GPT-4 substantially outperformed STM on average, although STM might be more suitable for certain specific themes.

Qualitative review for differences between GPT-4 and human thematic analyses. For thematic analysis using GPT-4, we categorized themes by stakeholders' job roles and summarized the findings. Unlike the uniform patterns of topic distribution among these three types of stakeholders' job roles revealed by human coding as shown in Figure 3, the GPT-4 results presented in Figure 5 displayed a distinct topic distribution for educators. Specifically, educators focused extensively on themes such as "Anti-racism" and "Trauma at home" within the "Culture, climate, and environment" category, as well as on "Community," "Student supports," and "School supports." These prevalent themes were consistent with our hypothesized focal areas based on educators' daily responsibilities, which further cast confidence in the utility of GPT-4 coding.

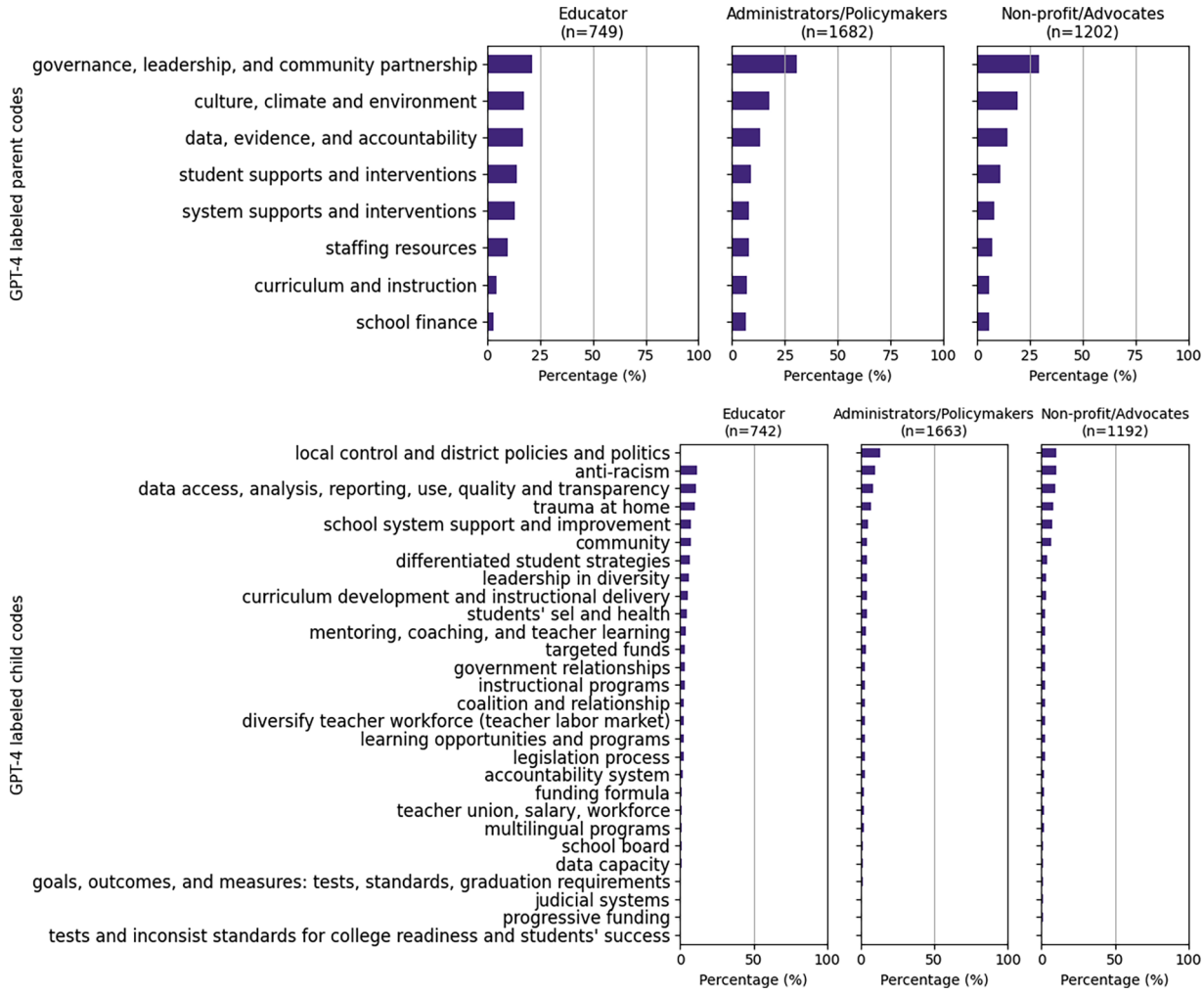

FIGURE 5. *GPT-4 Labeled Theme Frequency by Stakeholders' Job Roles.*

Note. Among 24 interviewees, five are educators, 10 are administrators and policymakers, and nine are non-profit advocates. Note that *n* indicates the number of labels assigned by human coders, drawn from interviews with individuals in the specified job roles. For example, out of 749 parent labels assigned to paragraphs from educator interviews, approximately 24% were coded with the parent theme "governance, leadership, and community partnership" and out of 742 child labels assigned to paragraphs from educator interviews, approximately 0% were coded with the child theme "local control and district policies and politics." Note that *n* may exceed the total number of paragraphs for a given role, as GPT-4 could assign multiple labels to a single paragraph. The *n* values for parent and child codes within a given role are not necessarily the same. In particular, the *n* for parent codes may exceed that of child codes, as GPT-4 might identify only a parent code for some paragraphs without assigning a corresponding child code.

However, the theme "Local control and district policies and politics" was notably absent from educators' narratives. Considering that the parent category of this child theme remained a significant topic among educators, we propose three potential explanations for this omission. First, the informal expertise that human coders utilized during coding may not have been fully captured in the child code description in the codebook. Second, the definition of this child code, which encompassed district-level decision-making and engagement with partners, as well as accountability and compliance under local control, might have been too nuanced for the LLM to discern, particularly in differentiating it from other child codes under the same parent category. Third, the plethora of codes provided may have overwhelmed GPT-4, preventing it from consistently attending to all codes. For instance, one educator's statement clearly fell under this code and its definition but was not accurately labeled by GPT-4:

> In terms of local policies, something that can be limiting is the emphasis on compliance. I was a teacher in the district, and then I left for several years, and then I came back as a professional-development specialist around English learners.

Moreover, compared to the frequency of human coding, GPT-4 identified the "Culture, climate, and environment" category more frequently. Given that our interviews focused centrally on the evidence for equity in Washington K–12

TABLE 4
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	STM vs. Human	0.31	0.09
Neutral	31	31	215	405	64	43			

public education, it is not surprising that themes such as "Anti-racism" were prominent in the discussions. GPT-4 seemed to detect nuances that had been overlooked by human coders. For example, one interviewee discussed changes in student demographics in programs for highly capable and gifted students, hinting at increased opportunities for non-White-middle-class students. GPT-4 recognized the inherent anti-racism tone in this statement. Nonetheless, GPT-4 also tended to overgeneralize broader themes because it lacked the ability to prioritize themes selectively, as human coders do by choosing fewer than three themes. Additionally, GPT-4 struggled to distinguish between codes with overlapping meanings under the same parent categories, resulting in some instances where themes human coders identified as "Progressive funding" were labeled as "Targeted funding" by GPT-4, inflating the latter's frequency.

In conclusion, GPT-4 demonstrated proficiency in identifying underlying themes in our interview data. The results also suggest that human coders and LLMs can be complementary, with humans providing the priority, specificity, and expertise needed for detailed analysis and judgment, while GPT-4 uncovers embedded meanings that may be overlooked by human analysis.

MRQ 2. How Accurate and Valid are GPT-4 Sentiment Classifications When Compared to Human Experts' and Lexicon-Based Sentiment Analysis?

Evaluation for sentiment analysis. The confusion matrices in Table 4 indicate that GPT-4 aligned more closely with human sentiment labels than the lexical-based VADER. VADER notably overstated the positive sentiment in the text. Although the default settings of VADER, which were used, had been tested and validated by its developers for a variety of contexts, they may not have adapted well to the domain-specific language in educational policy, and failed to identify the underlying dissatisfactory sentiments embedded in stakeholders' descriptions of programs and policies. Evidently, GPT-4 performed better at identifying the sentiment, aligning with human judgment in 58% of the corpus, especially for "understanding" the expressions of satisfaction, the potential for enhancing equity, and compliments on improvements. However, GPT-4 tended to overestimate sentiment. A significant divergence was observed between human perceptions and GPT-4 classifications when distinguishing "Negative" sentiment from "Neutral," which impeded the machine's overall performance. Humans might be more adept at detecting challenges and demands for improvement that were conveyed in the narratives. The inter-rater reliability between human coder and GPT-4 was moderate as measured by Cohen's κ in the last column of Table 4. Cohen's κ for sentiment analysis showed variability in previous studies, with agreement levels differing even among human coders. For example, Takala et al. (2014) reported that some pairwise agreement varied from 0.62 to 0.90 and Cohen's κ varied from 0.41 to 0.80 between experienced human annotators using the three-category sentiments on a common set of economy news stories.

Qualitative review for differences between GPT-4 and human sentiment analysis. Similar to the findings from thematic analysis, the agreement between GPT-4 and humans varied by theme. Figure 6 suggests that areas with more direct connections to equity tended to have higher agreement levels, as stakeholders could more clearly articulate their opinions on whether a given practice or policy was equity-enhancing or limiting. The more explicit illustrations resulted in easier sentiment capture by the machine.

GPT-4 struggled to distinguish mixed feelings, leaning toward neutral rather than negative. For instance, an educator's statement on learning standards was as follows:

> I do not even know where I stand, really, on common core standards. I like the idea of standards. I think that does help deliver a more equitable curriculum to students. but I think they are still fuzzy enough that it just simply does not happen. I know it does not happen."

Human coders labeled the sentiment as "Negative," noting the gap between intent and implementation expressed in the statement. In contrast, GPT-4 recognized the mixed emotions but gave a "Neutral" classification with an annotation:

> While they like the idea of standards and believe it can deliver a more equitable curriculum, they also mention that the standards are still fuzzy and not effectively implemented. The statement does not clearly lean towards a positive or negative sentiment, making it neutral.

Although the analyses from human coders and GPT-4 were similar, human coders utilizing their domain knowledge could discern the underlying negative emphasis by understanding the logic in the paragraph.

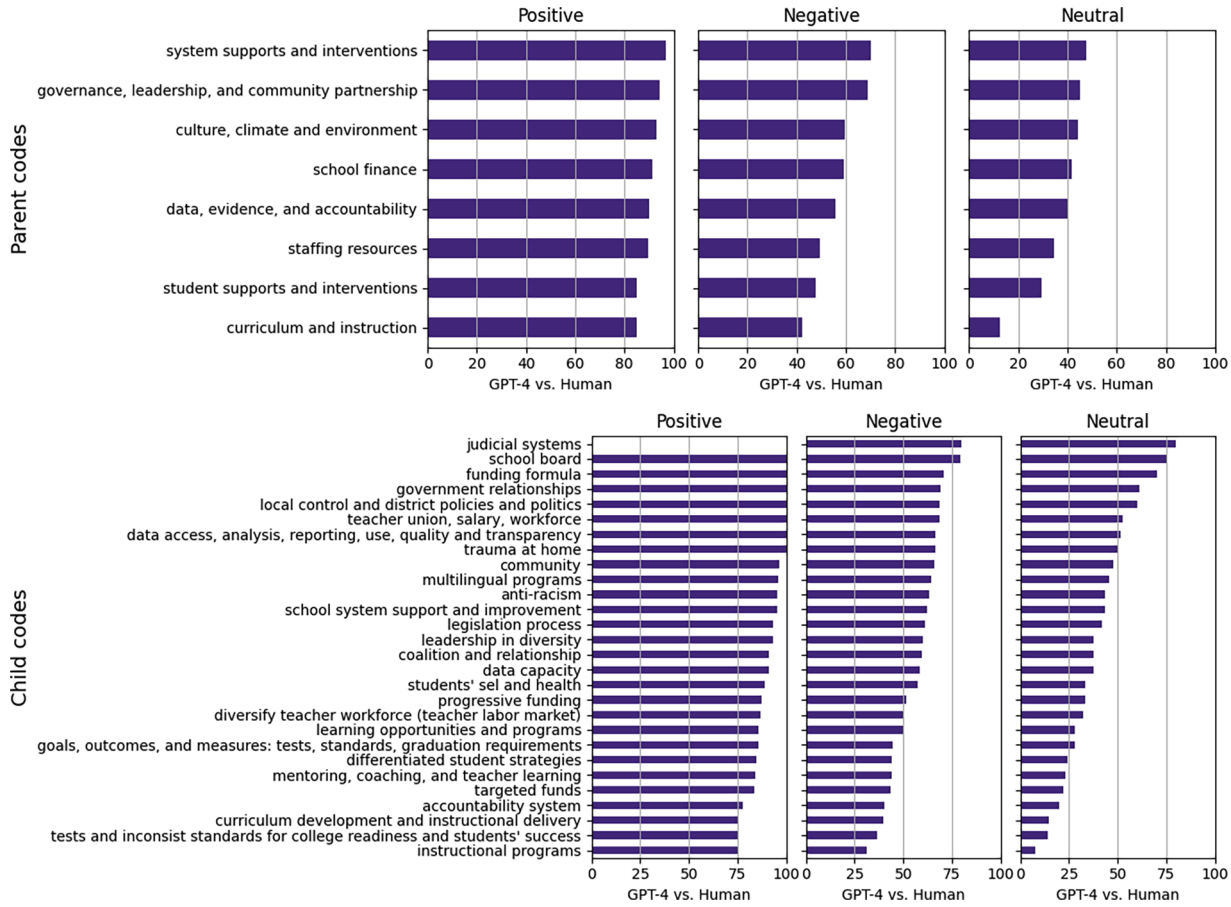

FIGURE 6. *Percentage of GPT-4 Sentiment Annotation That Agrees With Human Annotation, by Themes.*
Note. Among 24 interviewees, five are educators, 10 are administrators and policymakers, and nine are non-profit advocates. Among all paragraphs labeled as "system supports and interventions" (human labels are used for this analysis), over 90% of those coded as "positive" by human coders were also identified as "positive" by GPT-4. However, fewer than 50% of the paragraphs labeled as "neutral" by human coders were matched with the same neutral sentiment by GPT-4. The 0% match in positive sentiment for the child code "judicial system" under the educator role is due to the absence of any "positive" labels assigned by human coders when educators discussed this topic.

In conclusion, while human coders are better at logical reasoning for differentiating more nuanced sentiment intensity, GPT-4 also has its unique advantage of independent evaluation for each paragraph and uncovering implicit emotions, particularly in descriptive narratives. These distinct capabilities suggest that human and GPT-4 can complement each other in sentiment analysis to combine domain expertise with fresh perspectives.

Discussion and Limitations

This is one of the first studies to examine the potential of LLMs—represented by GPT-4—to facilitate highly domain-specific and context-dependent textual data analysis to facilitate high-stake decision-making. Our dual substance and methodological inquiries in this study have several implications for both educational policy for advancing racial and economic equity, and the potential promises and pitfalls of using LLMs.

Different stakeholder groups—educators, administrators and policymakers, and non-profit advocates—showed unique thematic focuses in the realm of educational policy. This diversity in thematic emphasis reflects the unique perspectives and priorities of each group. Educators predominantly concentrated on diversifying the teacher workforce and enhancing student supports, underscoring a direct engagement with the educational process. School administrators and policymakers, on the other hand, were more concerned with school finance and legislative processes, indicating a focus on the structural and regulatory aspects of education. Non-profit representatives brought attention to governance and community issues, highlighting the broader social context in which education operates. The varied thematic focuses suggest the need for a multi-faceted approach to educational policy analysis that takes into account the diverse priorities and perspectives of all stakeholders involved.

Our findings in sentiment analysis show that positive sentiments were predominantly directed towards "Progressive

funding" and reforms in student support, particularly highlighting the value of "Multilingual programs" in preserving students' cultural and linguistic heritage. This positive outlook underscores a growing recognition of the importance of cultural inclusivity in education. Conversely, the analysis also identified significant challenges, especially in areas like social-emotional and mental health support, which have been further exacerbated by the COVID-19 pandemic. Another notable concern raised in the analysis is the need for policymakers to prioritize diverse voices in decision-making processes. This includes balancing resource allocation, enhancing inclusivity, and extending community engagement to address the unique needs and perspectives of various groups within the education system.

Methodologically, GPT-4 demonstrates the large potential of LLMs to assist with analyzing large corpora of data to facilitate domain-specific decision-making in educational policy, especially in recognizing broader themes (parent codes). Besides its excellent performance in identifying themes with clear focal points, like "Multilingual programs" and "Data access, analysis, reporting, and use," GPT-4 was capable of capturing nuances within broader themes like "Legislation process." Compared to traditional NLP approaches, such as STM for thematic analysis and lexical-based sentiment analysis, LLMs have performed exceedingly well in various aspects.

GPT-4 can complement human expertise in that (a) it can discover nuances overlooked by human coders and (b) keep objective and consistent coding schema without being biased by human's implicit and unconscious influence from their lived experiences, prior knowledge, or cognitive tiredness during the coding. For example, GPT-4 captured nuances in stakeholder discussions about equity related to themes of "Anti-racism" and "Trauma at home," which were not initially picked up by human coders. This observation is consistent with prior research (De Paoli, 2024; Liang et al., 2024)

Additionally, LLMs offer a significant advantage in terms of time efficiency. Human qualitative coding can be labor-intensive; in the initial round, three expert coders devoted approximately 720 total hours over 3 months to complete all coding tasks. This process is also susceptible to the conceptual and subjective biases of the researchers. Collaboration with computational tools markedly reduces the time required for coding. In the subsequent round of human coding with an NLP-informed codebook, two research assistants completed the coding in a combined total of 42 hours. Including codebook development, the second stage took approximately 60 hours, a mere fraction of the time spent in the first stage. The use of LLM, specifically GPT-4, further decreased the time required, completing the coding in just 6 hours and introducing new insights into thematic discovery. Therefore, while recognizing the strengths and limitations of LLMs, a synergistic approach between LLMs and human expertise in textual analysis can enhance both efficiency and accuracy.

However, the performance of LLMs is sensitive to the nature of prompts, varying with domain intensity and the clarity of code descriptions (Liu et al., 2024). Prompts that incorporate extensive domain knowledge can help bridge the information gaps evident in current LLMs. The content validity of LLMs in aiding domain-specific data analysis relies on integrating domain knowledge into prompt development, which includes the codebook being incorporated into the prompt. Although humans' domain knowledge and lived experience may lead to bias and inconsistent implementation of coding schema, they ensure and verify the validity of LLMs and serve as an invaluable form of informal expertise that enriches LLMs' judgment.

Moreover, human experts are adept at capturing subtleties that LLMs may overlook. GPT-4 showed a tendency to overgeneralize themes, lacking the selective prioritization that human experts exhibit. Notably, GPT-4 performed well in identifying broader themes (parent codes) with an F1 score of 62%, which is higher than previous studies using similar technologies. However, its performance was less effective at the more specific child code level, struggling to differentiate between closely related sub-themes within the same broader categories. This issue suggests the need for more detailed descriptions in the codebook or prompts to improve LLMs' precision. We recommend future work to explicitly direct the model to consider all available options before producing results, add more iterations with examples, or conduct multiple analyses on subsets of the data or codebook (e.g., identify the top layer [parent codes] then bottom layer [child codes]).

Despite our efforts to integrate AI and human coding approaches, several limitations warrant consideration. (1) This study is limited by its scope, focusing on participants within a single state. For larger-scale studies involving broader geographic areas and more participant roles, such as students and families from more diverse backgrounds, algorithmic biases in LLMs may have more significant impacts, which needs to be addressed when applying automated qualitative analysis. (2) While the AI-computer partnership offers flexibility in theme granularity through codebook adjustments, this adaptability may introduce inconsistency in analysis. For example, more specific codebooks may add more challenges to LLM coding. (3) Additionally, questions remain about the optimal LLM selection and parameter settings. We call for evaluation on different models' performance for specific domain applications to ensure analytical rigor in the future study. Furthermore, we encourage a deeper conversation around the broader implications of adopting LLMs, particularly concerning the use of copyrighted content and the environmental impact of large-scale model deployment.

Conclusion

This study explored the potential of LLMs as a tool for uncovering themes and sentiments embedded in the narratives from stakeholders in Washington K–12 public education. Stakeholders emphasized the critical importance of inclusive governance, data transparency, and community partnerships within the Washington K–12 public school system, highlighting the need for stronger representation of marginalized voices. While they acknowledged some progress—such as reforms in multilingual education and progressive funding—concerns remained around insufficient student support, limited accountability systems, and persistent racial and systemic inequities across policy, practice, and leadership. The thematic and sentiment analyses have demonstrated GPT-4's potential in educational policy studies to analyze stakeholders' lived experiences and inform policymaking. Despite these promising results, LLMs are not yet capable of performing such analyses independently. Human domain-specific expertise remains crucial throughout the process for guidance and as a quality checker, since the risks—such as neglect, difficulty in making fine distinctions, and a tendency for overgeneralization—still need to be addressed. Policymakers and researchers should be cognizant of the limitations of LLMs as analytical tools, especially in terms of capturing the specificities of domain-specific meanings. Therefore, a balanced approach that combines the efficient thematic analysis capabilities of LLMs with the nuanced understanding of human coders can lead to more comprehensive and inclusive educational policy analysis that attends the varied needs and priorities of all stakeholders.

Author Contributions

Liu and Dr. Sun assume equal authorship.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is supported by Ballmer Group, 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.

Open Practices

The data and analysis files for this article can be found at https://www.openicpsr.org/openicpsr/project/237139/version/V1/view.

ORCID iDs

Alex Liu https://orcid.org/0000-0002-4785-1801
Min Sun https://orcid.org/0000-0001-5832-1534

Notes

1. We used STM's R package (stm), without specifying prevalence or content formulas.

2. As of November, 2023, the GPT-4 model available via OpenAI's API (and in ChatGPT with the GPT-4 label) is now a variant called GPT-4-turbo. OpenAI has not announced any architectural changes or version updates to GPT-4-turbo since its release in November 2023.

3. Child codes—including data capacity, accountability system, and instructional programs—were not identifiable during STM topic labeling.

4. Positive: If the compound score is $> = 0.05$. Negative: If the compound score is $< = -0.05$. Neutral: If the compound score is between - 0.05 and 0.05.

5. For both Cohen's Kappa and AUC, we applied the following procedure: (1) From the full set of annotated documents ($N \approx 1400$), we randomly sampled—with replacement—a new dataset of the same size: 100. This resampled set may include duplicate documents and omit others. (2) For Cohen's Kappa, we calculated the agreement between coding approaches on the resampled dataset using parent codes and child codes. For AUC, we computed the area under the ROC curve based on predicted probabilities and true labels from the resampled set. (3) Steps 1 and 2 were repeated 1,000 times to generate an empirical distribution for each statistic. (4) From this distribution, we estimated the mean, standard deviation, and 95% confidence intervals, yielding robust estimates even in the presence of class imbalance or noisy labels (Gwet, 2016).

6. Our data includes three broad job role categories: (1) Administrators/Policymakers: state legislators, other state-level policymakers, school district administrators; (2) Non-Profit and Advocates: teacher union representatives, policy advocates, and community leaders; and (3) Educators: teachers, teacher coaches or mentors.

7. Our prompt instructed GPT-4 to identify the three most salient themes, except in cases where there were not enough suitable themes available.

8. 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.

9. 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," and "Instructional programs."

Note. This manuscript was accepted under the editorial team of Kara S. Finnigan, Editor in Chief.

References

Abdulaziz, M., Alotaibi, A., Alsolamy, M., & Alabbas, A. (2021). Topic based sentiment analysis for COVID-19 tweets. *International Journal of Advanced Computer Science and Applications*, *12*(1), 626–636. https://doi.org/10.14569/IJACSA.2021.0120172

Alliance For Resource Equity. (2023, September 14). *Dimensions of equity*. Education Resource Strategies & The Education Trust. https://educationresourceequity.org/dimensions-of-equity/

Ashwin, J., Chhabra, A., & Rao, V. (2025). Using large language models for qualitative analysis can introduce serious bias.

Sociological Methods & Research. Advance online publication. https://doi.org/10.1177/00491241251338246

Azungah, T. (2018). Qualitative research: Deductive and inductive approaches to data analysis. *Qualitative Research Journal*, *18*(4), 383–400. https://doi.org/10.1108/qrj-d-18-00035

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., . . . Baker, R. S. (2024, July). *ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development* [Conference session]. International conference on artificial intelligence in education (pp. 134–149). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64299-9_10

Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, *68*(6), 1397–1410. https://doi.org/10.1002/asi.23786

Belal, M., She, J., & Wong, S. (2023). Leveraging ChatGPT as text annotation tool for sentiment analysis. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2306.17177

Belton, D., & Brinkmann, J. L. (2024). The relationship between school climate and student achievement in reading in public elementary schools in Virginia, USA. *Educational Planning*, *31*(1), 7–25. https://eric.ed.gov/?q=source%3A%22Educational+Planning%22&ff1=eduElementary+Education&id=EJ1416289

Benjamin, R. (2019). Assessing risk, automating racism. *Science*, *366*(6464), 421–422. https://doi.org/10.1126/science.aaz3873

Berry, K. S., & Herrington, C. D. (2013). Tensions across federalism, localism, and professional autonomy: Social media and stakeholder response to increased accountability. *Educational Policy*, *27*(2), 390–409. https://doi.org/10.1177/0895904812466171

Bhattacharya, K. (2017). *Fundamentals of qualitative research: A practical guide*. Routledge.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://dl.acm.org/doi/10.5555/944919.944937

Bonilla, S., Dee, T. S., & Penner, E. K. (2021). Ethnic studies increases longer-run academic engagement and attainment. *Proceedings of the National Academy of Sciences*, *118*(37), Article e2026386118. https://doi.org/10.1073/pnas.2026386118

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2016). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, *73*(2), 125–230. https://doi.org/10.3102/00346543073002125 (Original work published 2003)

Bowers, A. J., & Zhou, X. (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk*, *24*(1), 20–46. https://doi.org/10.1080/10824669.2018.1523734

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. University of Chicago Press. https://consortium.uchicago.edu/publications/organizing-schools-improvement-lessons-chicago

Cheng, L., Li, X., & Bing, L. (2023). Is GPT-4 a good data analyst? In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 9496–9514). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.637

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2306.14924

Cipriano, C., Strambler, M. J., Naples, L. H., Ha, C., Kirk, M., Wood, M., Sehgal, K., Zieher, A. K., Eveleigh, A., McCarthy, M., Funaro, M., Ponnock, A., Chow, J. C., & Durlak, J. (2023). The state of evidence for social and emotional learning: A contemporary meta-analysis of universal school-based SEL interventions. *Child Development*, *94*(5), 1181–1204. https://doi.org/10.1111/cdev.13968

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Çorbacioglu, S. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine*, *23*(4), 195–198. https://doi.org/10.4103/tjem.tjem_182_23

Dahouda, M. K., & Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. *IEEE Access*, *9*, 114381–114391. https://doi.org/10.1109/access.2021.3104357

Dai, S.-C., Xiong, A., & Ku, L.-W. (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 9993–10001). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.669

Das, N., Gupta, S., Das, S., Yadav, S., Subramanian, T., & Sarkar, N. (2021, September). *A comparative study of sentiment analysis tools* [Conference session]. 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES) (pp. 1–7). IEEE.

Davidson, P., Arndt-Bascle, C., Liedekerke, M-G. de, & Reyes, R. (2022). *Improving stakeholder engagement and evidence-based policy making*. The Regulatory Review. Retrieved January 14, 2023, from https://www.theregreview.org/2022/12/07/davidson-improving-stakeholder-engagement/

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, *42*(4), 997–1019. https://doi.org/10.1177/08944393231220483

Dee, T. S., & Penner, E. K. (2017). The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. *American Educational Research Journal*, *54*(1), 127–166. https://doi.org/10.3102/0002831216677002 (Original work published 2017).

Demirtas-Zorbaz, S., Akin-Arikan, C., & Terzi, R. (2021). Does school climate that includes students' views deliver academic achievement? A multilevel meta-analysis. *School Effectiveness and School Improvement*, *32*(4), 543–563. https://doi.org/10.1080/09243453.2021.1920432

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on

culture: Application to newspaper coverage of US government arts funding. *Poetics*, *41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Dixon, C. (2024). *School turnaround strategies: A literature review on effective practices and barriers*. Hoover Institution. https://www.hoover.org/research/school-turnaround-strategies-literature-review-effective-practices-and-barriers

Fedorowicz, M., & Aron, L. Y. (2021). *Improving evidence-based policymaking: A review*. Urban Institute. Retrieved January 14, 2023, from https://www.urban.org/sites/default/files/publication/104159/improving-evidence-based-policymaking-a-review.pdf

Fryer, R. G., Jr., & Howard-Noveck, M. (2020). High-dosage tutoring and reading achievement: Evidence from New York City. *Journal of Labor Economics*, *38*(2), 421–452. https://doi.org/10.1086/705882

Gao, J., Guo, Y., Li, T. J.-J., & Perrault, S. T. (2023). CollabCoder: A GPT-powered workflow for collaborative qualitative analysis. In *Companion publication of the 2023 conference on computer supported cooperative work and social Computing* (pp. 354–357). Association for Computing Machinery. https://doi.org/10.1145/3584931.3607500

Gates, S., Baird, M., Master, B., & Chavez-Herrerias, E. (2019). *Principal Pipelines: A feasible, affordable, and effective way for districts to improve schools*. RAND Corporation. https://doi.org/10.7249/RR2666

Gershenson, S., Hart, C. M. D., Hyman, J., Lindsay, C. A., & Papageorge, N. W. (2022). The long-run impacts of same-race teachers. *American Economic Journal: Economic Policy*, *14*(4), 300–342. https://doi.org/10.1257/pol.20190573

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(30), Article e2305016120. https://doi.org/10.1073/pnas.2305016120

Glaser, B. G., Strauss, A. L., & Strutzel, E. (1968). The discovery of grounded theory; strategies for qualitative research. *Nursing Research*, *17*(4), 364. https://doi.org/10.1097/00006199-196807000-00014

Gosavi, S. R. (2022). *Transformer based detection of sarcasm and its sentiment in textual data* [Doctoral dissertation, National College of Ireland]. https://norma.ncirl.ie/6129/1/shubham-ramgosavi.pdf

Grabarek, J., & Kallemeyn, L. M. (2020). Does teacher data use lead to improved student achievement? A review of the empirical evidence. *Teachers College Record*, *122*(12), 1–42. https://doi.org/10.1177/016146812012201201

Grimmer, J. (2013). Appropriators not position takers: The distorting effects of electoral incentives on congressional representation. *American Journal of Political Science*, *57*(3), 624–642. https://doi.org/10.1111/ajps.12000

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences* (pp. 70–78). Princeton University Press. https://press.princeton.edu/books/paperback/9780691207551/text-as-data?srsltid=AfmBOoq-VwiJ8kU_K2J2T1BtkVC3sVevQqHCkuDTMXKaQBCYYmYOkIOJ

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. https://doi.org/10.1093/pan/mps028

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M. V., Dodge, K., Farkas, G., Fryer, R. G., Mayer, S., Pollack, H., Steinberg, L., & Stoddard, G. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review*, *113*(3), 738–765. https://doi.org/10.1257/aer.20200647

Gwet, K. L. (2016). Testing the difference of correlated Agreement Coefficients for Statistical Significance. *Educational and Psychological Measurement*, *76*(4), 609–637. https://doi.org/10.1177/0013164415596420

Holt, S. B., & Gershenson, S. (2019). The impact of demographic representation on absences and suspensions. *Policy Studies Journal*, *47*(4), 1063–1093. https://doi.org/10.1111/psj.12229

Huang, Y., Gomaa, A., Semrau, S., Haderlein, M., Lettmaier, S., Weissmann, T., Grigo, J., Tkhayat, H. B., Frey, B., Gaipl, U., Distel, L., Maier, A., Fietkau, R., Bert, C., & Putz, F. (2023). Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: Potentials and challenges for AI-assisted medical education and decision making in radiation oncology. *Frontiers in Oncology*, *13*, 1265024. https://doi.org/10.3389/fonc.2023.1265024

Hutto, C., & Gilbert, E. (2014, May). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225.https://doi.org/10.1609/icwsm.v8i1.14550

Jackson, C. K., Johnson, R. C., & Persico, C. (2016). The effects of school spending on educational and economic outcomes: Evidence from school finance reforms. *The Quarterly Journal of Economics*, *131*(1), 157–218. https://doi.org/10.1093/qje/qjv036

Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2733–2742. https://doi.org/10.1101/2020.04.22.054973

Kheiri, K., & Karimi, H. (2023). Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2307.10234

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners. *arXiv Preprint*. https://arxiv.org/abs/2205.11916

Leeson, W., Resnick, A., Alexander, D., & Rovers, J. (2019). Natural language processing (NLP) in qualitative public health research: A proof of concept study. *International Journal of Qualitative Methods*, *18*, 1–9. https://doi.org/10.1177/1609406919887021

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, *1*(8), eAIoa2400196. https://doi.org/10.1056/AIoa2400196

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z., & Tang, J. (2022). P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 61–68). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.8

Liu, X., Zhang, J., Barany, A., Pankiewicz, M., & Baker, R. S. (2024). Assessing the potential and limits of large language models in qualitative coding. In Y. J. Kim, & Z. Swiecki (Eds.), *International conference on quantitative ethnography* (pp. 89–103). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-76335-9_7

Long, M. C., Conger, D., & Iatarola, P. (2012). Effects of high school course-taking on secondary and postsecondary success. *American Educational Research Journal*, *49*(2), 285–322. https://doi.org/10.3102/0002831211431952 (Original work published 2012).

Lyu, Q., Tan, J., Zapadka, M. E., Ponnatapura, J., Niu, C., Myers, K. J., Wang, G., & Whitlow, C. T. (2023). Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: Results, limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art*, *6*(1), 9. https://doi.org/10.1186/s42492-023-00136-5

Mathis, W. S., Zhao, S., Pratt, N., Weleff, J., & De Paoli, S. (2024). Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine*, *255*, Article 108356. https://doi.org/10.1016/j.cmpb.2024.108356

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, *33*(2), 3–11. https://doi.org/10.3102/0013189x033002003

McCleary v. State, 269 P.3d 227 (Wash. 2012). Retrieved from https://law.justia.com/cases/washington/supreme-court/2012/84362-7-0.html

Morgan, I. (2022). *Equal is not good enough: An analysis of school funding equity across the U.S. and within each state* (Report No. ED626472). Education Trust. https://eric.ed.gov/?id=ED626472

Nguyen, H., Moon, J., Paul, N., & Gokhale, S. S. (2021, September). *Sarcasm detection in politically motivated social media content* [Conference session]. 2021 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom) (pp. 1538–1545). IEEE. https://doi.org/10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00207

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications. https://uk.sagepub.com/en-gb/eur/qualitative-research-evaluation-methods/book232962

Rosenberg, J. M., Borchers, C., Dyer, E. B., Anderson, D., & Fischer, C. (2021). Understanding public sentiment about educational reforms: The next generation science standards on Twitter. *AERA Open*. Advance online publication. https://doi.org/10.1177/23328584211024261

Savelka, J. (2023). *Unlocking practical applications in legal domain: Evaluation of GPT for zero-shot semantic annotation of legal texts* [Conference session]. Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, (447–451), Association for Computing Machinery. https://doi.org/10.1145/3594536.3595161

Schueler, B. E., Goodman, J. S., & Deming, D. J. (2017). Can states take over and turn around school districts? Evidence from Lawrence, Massachusetts. *Educational Evaluation and Policy Analysis*, *39*(2), 311–332. https://doi.org/10.3102/0162373716685824

Shlezinger, N., Whang, J., ESTMr, Y. C., & Dimakis, A. G. (2023). Model-based deep learning. In *Proceedings of the IEEE* (Vol. 111, pp.465–499). https://doi.org/10.1109/JPROC.2023.3247480

Sprenkamp, K., Jones, D. G., & Zavolokina, L. (2023). Large language models for propaganda detection. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2310.06422

Sun, M., Liu, J., Zhu, J., & LeClair, Z. (2019). Using a text-as-data approach to understand reform processes: A deep exploration of school improvement strategies. *Educational Evaluation and Policy Analysis*, *41*(4), 510–536. https://doi.org/10.3102/0162373719869318

Takala, P., Malo, P., Sinha, A., & Ahlgren, O. (2014). *Gold-standard for topic-specific sentiment analysis of economic texts* (Vol. 2014, pp. 2152–2157). LREC.

Tang, R., Chuang, Y. N., & Hu, X. (2024). The science of detecting LLM-generated text. *Communications of the ACM*, *67*(4), 50–59.

Wallner, J. (2008). Legitimacy and public policy: Seeing beyond effectiveness, efficiency, and performance. *Policy Studies Journal*, *36*(3), 421–443. https://doi.org/10.1111/j.1541-0072.2008.00275.x

Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., . . . Zhou, J. (2023). Is chatgpt a good NLG evaluator? a preliminary study. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2303.04048

Wang, Z., Xie, Q., Ding, Z., Feng, Y., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2304.04339

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)* (Article 1800, pp. 1–14). Curran Associates Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

Wu, H., & Shen, J. (2022). The association between principal leadership and student achievement: A multivariate meta-meta-analysis. *Educational Research Review*, *35*, Article 100423. https://doi.org/10.1016/j.edurev.2021.100423

Wulff, D. U., Hussain, Z., & Mata, R. (2024). The behavioral and social sciences need open LLMs. *OSF Preprints*. https://doi.org/10.31219/osf.io/ybvzs

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces* (pp. 75–78). Association for Computing Machinery. https://doi.org/10.1145/3581754.3584136

Xu, H., Yi, S., Lim, T., Xu, J., Well, A., Mery, C., Zhang, A., Zhang, Y., Ji, H., Pingali, K., Leng, Y., & Ding, Y. (2025). TAMA: A human-AI collaborative thematic analysis framework using multi-agent LLMs for clinical interviews. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2503.20666

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 12697–12706). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v139/zhao21c.html

Authors

ALEX LIU is a PhD candidate in the College of Education at the University of Washington, Seattle, WA, USA; email: alexliux@uw.edu. Her research focuses on bridging AI/ML, teacher learning, and strategic teacher engagement to advance instructional quality, strengthen teaching effectiveness, and improve student outcomes.

Dr. MIN SUN is a professor in the College of Education at the University of Washington, Seattle, WA, USA, and Director of the AmplifyLearn AI Center; email: misun@uw.edu. Her research focuses on teacher learning, the school and policy contexts that support teacher learning, and AI/ML research and application developments in education.