Costs and Benefits of AI-Enabled Topic Modeling in P-20 Research: The Case of School Improvement Plans

Syeda Sabrina Akter[1], **Seth B. Hunter**[2], **David S. Woo**[3], **Antonios Anastasopoulos**[1,4]
[1]Department of Computer Science, George Mason University
[2]Department of Educational Leadership and Policy, George Mason University
[3]Department of Educational Leadership and Policy, The University of Utah
[4]Archimedes, Athena Research Center, Greece
sakter6@gmu.edu

Abstract

As generative AI tools become increasingly integrated into educational research workflows, large language models (LLMs) have shown substantial promise in automating complex tasks such as topic modeling. This paper presents a user study that evaluates AI-enabled topic modeling (AITM) within the domain of P-20 education research. We investigate the benefits and trade-offs of integrating LLMs into expert document analysis through a case study of school improvement plans, comparing four analytical conditions. Our analysis focuses on three dimensions: (1) the marginal financial and environmental costs of AITM, (2) the impact of LLM assistance on annotation time, and (3) the influence of AI suggestions on topic identification. The results show that LLM increases efficiency and decreases financial cost, but potentially introduce anchoring bias that awareness prompts alone fail to mitigate.[1]

1 Introduction

Educators are increasingly turning to artificial intelligence to streamline research and administrative workflows, particularly within P-20 contexts, which cover education from Pre-K through graduate levels and workforce training. It has sparked considerable interest in the potential of generative AI tools to tackle complex analytical tasks (Kasneci et al., 2023; Xu et al., 2024). Among these applications, topic modeling (TM)—a method for uncovering hidden themes in unstructured text—has become a prominent technique in P-20 research over the past decade (Brookes and McEnery, 2019; Daenekindt and Huisman, 2020; Sun et al., 2019; Wang et al., 2017). In contrast to conventional text data analysis (CTDA), which often requires substantial human input and can be constrained by its labor-intensive nature, subjectivity, and potential for inconsistency, Artificial Intelligence-enabled Topic Modeling (AITM), driven by sophisticated LLMs like GPT-4 (OpenAI et al., 2024), holds the potential for significant improvements in efficiency and scalability by automating or assisting with these demanding procedures (Dell'Acqua et al., 2023; Grossmann et al., 2023).

Implementing AITM offers several key benefits, notably a reduction in the time required for analysis and the potential for more consistent and thorough topic identification. These efficiencies can significantly influence research productivity and, importantly, may lead to qualitatively different research findings compared to CTDA due to variations in identified themes. Nonetheless, the rapid adoption of AITM raises important concerns about potential drawbacks, such as financial costs and environmental impacts associated with substantial computational resource utilization. At present, there is a lack of empirical research that compares these costs to those of traditional methods, especially within the field of K12 educational research.

Another critical but underexplored concern with AITM is the psychological phenomenon known as anchoring bias—the tendency for humans to rely excessively on initially presented information when making subsequent judgments or decisions (Nagtegaal et al., 2020). In contexts where humans interact with AI-generated insights, anchoring bias may skew human analysts' judgments, thus, affecting the final research outcomes (Zhao et al., 2024; Choi et al., 2024).

Given these critical gaps, we investigate the financial, environmental, cognitive, and analytical trade-offs of integrating AITM into P-20 research. Our case study focuses on principal-written school improvement plans (henceforth "Plans") from a formal field-based principal evaluation sys-

[1]Code available here. The data are not publicly available due to privacy restrictions but can be requested through the Network for Educator Effectiveness (NEE) at the University of Missouri and the Missouri Department of Elementary and Secondary Education (DESE).

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 460–476
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

tem in hundreds of K12 districts in the Midwest USA. We systematically evaluate four analytic conditions: `AI-Only` , `Human-Only`, `AI-Human`, and `AI-Human-Deanchoring`. Through this comparative analysis, we address three research questions:

- **RQ1:** What are the marginal financial and environmental costs of implementing AITM in P-20 research?
- **RQ2:** What are the causal effects of different analytic approaches on analysis time?
- **RQ3:** What are the causal effects of these analytic approaches on the topics identified?

Preliminary findings suggest that AI analysis significantly reduces costs and analysis time per document compared to human analysis, although AI-assisted methods vary slightly in terms of speed. Additionally, when humans and AI were provided with pre-specified topic lists, only minor differences emerged in the topics identified. Through a thorough evaluation of these aspects, we aim to offer an empirical understanding of AITM's value proposition for P-20 educational research.

2 Related Work

The field of topic modeling has seen significant advancements, moving from traditional probabilistic methods to more contemporary AI-driven techniques. Early models, such as Latent Dirichlet Allocation (LDA; Blei et al., 2003), conceptualized documents as combinations of topics, with each topic characterized by a distribution of words. While widely adopted, LDA and similar approaches often required substantial manual interpretation, as they yielded clusters of words without clear semantic labels (Gao et al., 2024b). Subsequent neural network-based models, like BERTopic (Grootendorst, 2022), improved the coherence of topics by leveraging transformer embeddings that capture richer contextual meaning. More recently, frameworks leveraging large language models (LLMs), such as TopicGPT (Pham et al., 2024), have further enhanced the accessibility and interpretability of topic modeling by generating human-readable topic labels and summaries (Overney et al., 2024; Gao et al., 2024a).

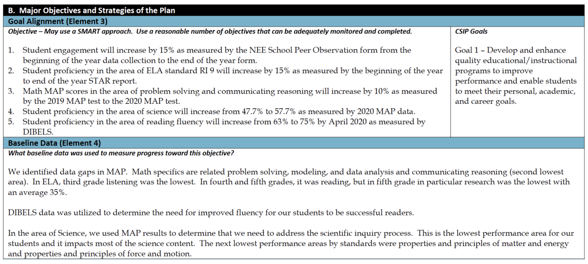
Within educational research, topic modeling has proven to be a powerful tool for analyzing large-scale textual data, such as curricula, school improvement plans, and scholarly literature. Studies have applied topic modeling to uncover latent themes in educational leadership, policy discourse, and reform strategies (Wang et al., 2017; Sun et al., 2019; Daenekindt and Huisman, 2020). These methods claim to significantly reduce the labor associated with traditional qualitative coding, making large-scale analysis more scalable and helping to address a fundamental impediment to research use by educators: the amount of time it takes to conduct research (Drahota et al., 2016; Asmussen and Møller, 2019).

As AI tools, particularly LLMs, become more prominent in education research and practice, they are being increasingly adopted for tasks such as writing content, analyzing student responses, or synthesizing research findings (Liu and Wang, 2024; Cambon et al., 2023; Jaffe et al., 2024). However, effective adoption in educational contexts requires addressing the environmental and financial costs of model training and inference (Strubell et al., 2019; Hershcovich et al., 2022), challenges around the reliability and interpretability of model outputs (Mittelstadt et al., 2016; Sahoo et al., 2024), and cognitive pitfalls such as automation and anchoring bias that may skew human judgment during analysis (Goddard et al., 2012; Koo et al., 2024; Echterhoff et al., 2024). This is particularly concerning in high-stakes domains like education, where premature reliance on AI-suggested outputs can limit critical thinking, reduce analytical diversity, and ultimately affect the integrity of findings (Al-Zahrani, 2024; Sallam, 2023).

Furthermore, bias mitigation remains a pressing challenge. LLMs have been shown to inherit and sometimes amplify social and cultural biases (Resnik, 2024). Interestingly, emerging research suggests that strategies such as structured group discussions and collaborative review can counteract some of these effects, promoting more balanced and reflective decision making in AI-assisted workflows (Horst et al., 2019; Rachael A. Hernandez and Teal, 2013; Michaelsen et al., 2002).

3 Data

We use a proprietary dataset from the Network for Educator Effectiveness (NEE), an educator evaluation system widely implemented across K–12 school districts in Missouri. This dataset spans the academic years 2005–2006 through 2022–2023 and comprises de-identified, text-based portfolios authored by school principals. These documents, formally known as *Building Improvement Plans*



B. Major Objectives and Strategies of the Plan

Goal Alignment (Element 3)

Objective – May use a SMART approach. Use a reasonable number of objectives that can be adequately monitored and completed.	***CSIP Goals***
1. Student engagement will increase by 15% as measured by the NEE School Peer Observation form from the beginning of the year data collection to the end of the year form. 2. Student proficiency in the area of ELA standard RI 9 will increase by 15% as measured by the beginning of the year to end of the year STAR report. 3. Math MAP scores in the area of problem solving and communicating reasoning will increase by 10% as measured by the 2019 MAP test to the 2020 MAP test. 4. Student proficiency in the area of science will increase from 47.7% to 57.7% as measured by 2020 MAP data. 5. Student proficiency in the area of reading fluency will increase from 63% to 75% by April 2020 as measured by DIBELS.	Goal 1 – Develop and enhance quality educational/instructional programs to improve performance and enable students to meet their personal, academic, and career goals.

Baseline Data (Element 4)

What baseline data was used to measure progress toward this objective?

We identified data gaps in MAP. Math specifics are related problem solving, modeling, and data analysis and communicating reasoning (second lowest area). In ELA, third grade listening was the lowest. In fourth and fifth grades, it was reading, but in fifth grade in particular research was the lowest with an average 35%.

DIBELS data was utilized to determine the need for improved fluency for our students to be successful readers.

In the area of Science, we used MAP results to determine that we need to address the scientific inquiry process. This is the lowest performance area for our students and it impacts most of the science content. The next lowest performance areas by standards were properties and principles of matter and energy and properties and principles of force and motion.

Figure 1: The two elements extracted from the Building Improvement Plans (BIPs) used in our goal-based study.

(BIPs) or *School Improvement Plans*, are submitted annually as part of a standardized evaluation process and are structured around seven performance criteria (referred to as *elements*) evaluated by principal supervisors using a consistent rubric.

For our study, we randomly selected 23 BIPs and focused on two specific elements from each plan: (1) the major objectives stated for school improvement, and (2) the data principals planned to use to measure progress toward those objectives (Figure 1). These elements are highly relevant to evaluating strategic goal-setting and progress tracking in educational leadership and K12 school improvement. The documents are entirely text-based and machine-readable, making them ideal for qualitative analysis via topic modeling.

4 Experimental Design

To investigate the integration of LLMs into educational research, we have adapted our methodology from the user study conducted by Choi et al. (2024), which examined the efficiency and precision of LLMs in specialized tasks through a structured user study focused on human-LLM interactions. Their findings showed that while LLMs significantly increased task speed, they also led users to anchor on AI-provided suggestions. Informed by their findings on anchoring bias, we expand on their experimental framework by adding a novel treatment condition: **`AI-Human-Deanchoring`**. This condition is designed to reduce the over-reliance on LLM by making participants explicitly aware of potential anchoring effects in LLM-generated suggestions (see Figure 2).

Our study is structured in two stages:

- **Stage 1: Topic Discovery**, in which participants identify and curate a list of topics from a shared set of BIPs.
- **Stage 2: Topic Assignment**, in which participants apply those topics to a new set of documents under controlled conditions.

Document	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**
Stage 1						
D1–D11	T2	T3	T4	T2	T3	T4
Stage 2						
D12–D15	T2	T3	T4	T2	T3	T4
D16–D19	T3	T4	T2	T3	T4	T2
D20–D23	T4	T2	T3	T4	T2	T3

Table 1: Document assignments for Stages 1 and 2. In Stage 1, each analyst analyzed the full set of 11 documents (D1–D11) under a single assigned condition; experimental conditions are defined as T2: `Human-Only`, T3: `AI-Human`, and T4: `AI-Human-Deanchoring`. In Stage 2, analysts analyzed documents D12–D23, assigned in a balanced design across all experimental conditions to ensure multiple annotations per document.

We designed the following four treatment conditions:

1. **`AI-Only`:** Tasks were performed solely by the LLM without human intervention, providing a benchmark for AI performance.
2. **`Human-Only`:** Participants performed tasks without any AI assistance, serving as the baseline for human performance.
3. **`AI-Human`:** Participants received suggestions from an LLM before performing tasks, allowing us to assess the influence of AI assistance.
4. **`AI-Human-Deanchoring`:** Participants were presented with LLM-generated suggestions with explicit instructions to be skeptical of them due to potential anchoring bias. By encouraging participants to thoughtfully evaluate and adjust AI-generated recommendations, we aim to improve the trustworthiness and credibility of AI-generated results.

To assign treatment conditions in the 12 school improvement plans (BIPs) in stage 2, we used a Latin square design (Montgomery, 2017). Each of the six human participants was assigned a specific sequence of treatment conditions across different plans, ensuring a balanced and systematic distribution of the `Human-Only`, `AI-Human`, and `AI-Human-Deanchoring` settings (see Table 1). Analysts proceeded in the order of conditions T2 → T3 → T4 in stage 2. Participants in the AI-assisted settings (T3, T4) were provided with LLM-generated topic annotations, while those in the `Human-Only` setting (T2) worked independently without any AI input. Analysts were unaware of the condition until they accessed the designated docu-

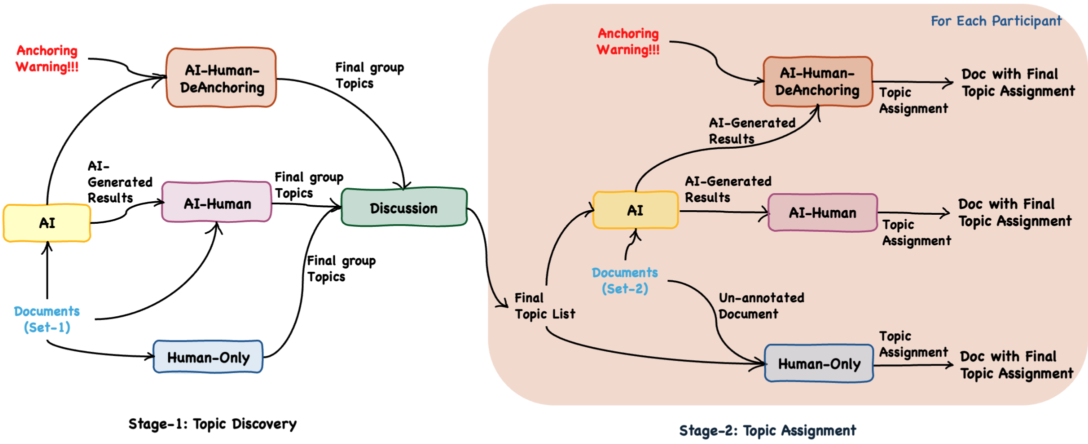


Figure 2: Overview of our Topic Modeling workflow and experimental settings. Stage-1: Topic Discovery involves discovering latent topics within documents. Team discussion occurred at the end of Stage-1 in order develop the Final Topic List. Stage-2 involves assigning topics to a different set of documents in all treatment conditions.

ment in Label Studio (Tkachenko et al., 2020) that we used to conduct our user study. Each analyst was asked to indicate whether each topic t from the Final Topic List appeared in each paragraph/field f of the assigned BIPs.

4.1 Stage 1: Topic Discovery

We used the data from Stage 1 to examine how topic lists were generated across different analytic conditions. First, all six participants analyzed the same set of 11 BIPs, working individually under one of three assigned conditions: **`Human-Only`**, **`AI-Human`**, or **`AI-Human-Deanchoring`**, with two participants per condition.

Participants in the `AI-Human` and `AI-Human-Deanchoring` conditions were provided with an LLM-generated topic list before beginning their analysis, while only the latter were explicitly warned about potential anchoring effects (see Appendix A for the full instruction). Analysts in the `Human-Only` condition received no AI input.

Each analyst independently reviewed all 11 BIPs and recorded a preliminary list of topics. After this individual phase, participants met in their respective condition groups for a 30-minute discussion to consolidate their findings into a group-specific topic list. Finally, all six analysts engaged in a 60-minute cross-condition discussion to synthesize the **Final Topic List**, which was later used as the reference framework in Stage 2. All individual and group topic lists are included in Appendix B.

Results In Stage 1, we collected three topic lists: from the `Human-Only` group, `AI-Human` group, and `AI-Human-Deanchoring` group. The analysts unanimously curated a final list of 13 topics after reviewing all three.

Despite differences in conditions, we observed moderate overlap: six of the 13 final topics (46%) appeared in all three lists, though not always as exact matches. For instance, some themes were phrased differently across settings, such as *Educational Technology* from the `AI-Human-Deanchoring` list and *Technology Integration* from the `Human-Only` list which were conceptually merged into a single topic, *Technology Use/Integration*, in the final list. This highlights how interpretive nuance plays a role in topic curation.

Comparing each user-generated list with the `AI-Only` list revealed systematic differences. The `AI-Only` list included 8 topics. The `Human-Only` list had 15, with 4 overlapping (26.67%), while `AI-Human` and `AI-Human-Deanchoring` lists identified 8 and 11 topics with 3 (37.5%) and 7 (63.64%) overlapping, respectively. The `Human-Only` list had a more granular set of topics tailored to the dataset, whereas the `AI-Only`, `AI-Human`, and `AI-Human-Deanchoring` lists tended to include broader, more generic themes that echoed the LLM's original suggestions. This suggests that the presence of LLM suggestions may have influenced annotators to propose fewer, more AI-aligned topics. In contrast, the final topic list—compiled after collaborative review—shared only 4 of 13 topics (30.77%) with the `AI-Only` list. This divergence suggests that discussion among annotators helped complement AI outputs by adding nuanced topics that the model did not generate.

Comparing Final Topic List and AI-Only List	# of Topics
Exact topic matches between Final Topic List & AI-Only List	4
Topics Present (or Discovered) in the Final Topic List, but not in the AI-Only List	5
Two or more topics from Final Topic List subsumed under one broader AI-Only topics[2]	4
Topics completely discarded by the annotators from AI-Only List	2
Total topics in Final Topic List	**13**

Table 2: The comparison of the AI-Only List with respect to the Final Topic List shows that there are few topics that the model has failed to cover in its overall topic generation task.

Metric	Human-Only	AI-Human	AI-Human-Deanchoring
Avg Precision	0.68	0.84	0.83
Avg Recall	0.55	0.69	0.67
Avg Annotation Speed (words/min)	73.75	71.15	89.91
Avg Annotator Agreement with AI-Only (%)	54.64	73.44	71.41
Avg Inter-Annotator Agreement (κ)	0.57	0.71	0.69

Table 3: Summary of Stage 2 results across the three settings. Metrics include annotation speed (words per minute), agreement with LLM outputs (%), and inter-annotator agreement (Cohen's κ). See Appendix C for detailed results and metrics definitions.

However, 5 of the 13 topics in the final topic list were not present in the AI-Only list at all (see Table 2). These "missing" topics—such as *Classroom Environment* and *Attendence*—often represented context-specific or nuanced areas that the LLM failed to surface.

Additionally, annotators explicitly discarded two LLM topics, *Education* and *School Improvement Planning*, as overly broad. This further illustrates a recurring pattern: while LLMs are helpful in identifying broad thematic content, they may struggle with generating the fine-grained, action-relevant topics that human experts prioritize in education policy contexts. These findings are consistent with prior work by Choi et al. (2024), which similarly highlighted LLMs' limitations in capturing nuanced, context-specific insights.

4.2 Stage 2: Topic Assignment

Participants used the Final Topic List to annotate a new set of 12 BIPs, each segmented into three paragraph-level fields. Participants were randomly reassigned to one of the three human-in-the-loop conditions. Those in AI-Human and AI-Human-Deanchoring received LLM-generated topic suggestions; those in Human-Only did not. We recorded the time spent on each document to facilitate an efficiency analysis.

Results Following Stage 2, we analyzed expert annotations across three conditions: Human-Only, AI-Human, and AI-Human-Deanchoring. Each paragraph in the dataset was represented as a 14-element vector—13 corresponding to topics from the Final Topic List established in Stage 1, and one for "None"—indicating whether annotators assigned relevant topics. This structure allowed us to assess the impact of LLM suggestions on annotation behavior.

Participants used the **Final Topic List** to annotate a new set of 12 school improvement plans (BIPs), each segmented into three paragraph-level fields. Each field was annotated independently by five human analysts, resulting in 12 plans $\times$ 3 fields $\times$ 5 analysts = 180 annotations. Additionally, each field was annotated once under the AI-Only condition, yielding 36 more entries, for a total of 216 topic-field-annotator combinations.

We evaluated the LLM's ability to replicate expert topic assignments using precision and recall, with the Human-Only condition treated as ground truth[3]. The AI-Only treatment achieved an average precision of 0.68 and recall of 0.55 when compared to Human-Only annotations, suggesting that while AI outputs are often accurate, they miss nearly half of expert-identified topics.

Annotators were significantly faster in the AI-Human-Deanchoring condition (89.91 words / min) than in the Human-Only (73.75 words/min) or AI-Human (71.15 words/min) conditions. This may reflect a tendency to anchor on LLM-generated suggestions, even when warned, leading to faster—but potentially *less critical*—annotation behavior.

Annotator agreement with AI-Only treatment was highest in the AI-Human condition (73.44%), followed by AI-Human-Deanchoring (71.41%), and lowest in Human-Only (54.64%). These findings suggest that LLM suggestions strongly influ-

[2]Multiple Final Topic List entries (e.g., "Academic Assessments" and "Academic Goals") were grouped under a single LLM topic (e.g., "Student Assessment and Achievement")

[3]We consider the Human-Only annotations as the ground truth because, typically, experts work independently without AI-assistance. This makes the annotations the closest representation of real-life expert results in our study.

Source	Reported Cost in the paper	Standardized Cost (per 100 tokens)
Walther (2024)	\$0.001 per 100 input, \$0.003 per 100 output	\$0.004 roundtrip
DeepLearning.AI (2024)	\$4 per million tokens (GPT-4o); \$2 per million tokens (Batch API)	\$0.0002–\$0.0004
Chen et al. (2023)	\$0.20–\$300 per 10M tokens (GPT-J to GPT-4 Turbo)	\$0.000002–\$0.003
Irugalbandara et al. (2024)	5×–29× cost reduction over GPT-4	\$0.00014–\$0.0008
Samsi et al. (2023)	3–4 Joules per token (LLaMA-65B)	0.000083–0.000111 kWh
Husom et al. (2024)	0.000083–0.0023 kWh per query (2B–70B)	0.000083–0.0023 kWh
Calma (2023)	>10× increase in energy per query	Relative 10× increase (qualitative only)

Table 4: Reported and standardized LLM inference costs from recent sources. All values in the third column are standardized to cost per 100 tokens—monetary in USD and environmental in kilowatt-hours (kWh).

ence annotator decisions, and simple warnings are not sufficient to mitigate anchoring effects.

Pairwise agreement (Cohen's κ; Cohen, 1960) between annotators was highest when both had access to LLM suggestions (`AI-Human`: 0.71, `AI-Human-Deanchoring`: 0.69), and lowest in the `Human-Only` condition (0.57), reflecting a possible anchoring effect in which annotators align more closely—not with each other independently—but around the AI-provided suggestions.

5 RQ1: Estimating AI Inference Costs

Methodology To evaluate the marginal cost of using LLMs in our topic modeling workflow, we synthesized pricing and energy consumption data from peer-reviewed literature, arXiv preprints, and blog sources. For environmental costs, we reviewed the literature that estimates kilowatt-hour (kWh) usage and dollar-converted emissions per LLM inference. To enable comparison across studies with differing units and assumptions, we standardized all monetary costs to U.S. dollars per 100 tokens and converted energy-related figures to kilowatt-hours (kWh) per 100 tokens using a conversion factor of $1,\text{kWh} = 3.6 \times 10^6$ joules. While we do not report pretraining costs—since our study involves only inference—we present a plausible range of energy costs based on similar LLM use cases.

Results We synthesized recent estimates of both the monetary and environmental costs of LLM inference by reviewing peer-reviewed publications, technical reports, and industry analyses. Table 4 summarizes the most relevant findings.

Our analysis shows that LLM inference costs range from \$0.0002 to \$0.004 per 100-token roundtrip, depending on the model, pricing tier, and batching strategy (Walther, 2024; DeepLearning.AI, 2024; Chen et al., 2023). Models like GPT-4 Turbo average around \$0.004 per inference, while batching can further reduce costs to as low as \$0.0002. Open-source alternatives offer additional savings, with some deployments reporting cost reductions of up to 29$\times$ (Irugalbandara et al., 2024). Although not directly reporting numeric costs, theoretical analyses from Aryan et al. (2023) further support these findings by emphasizing significant potential for cost optimization through efficient deployment strategies.

Environmental costs also scale significantly with model size and usage. For example, generating 100 tokens with LLaMA-65B consumes approximately $8.3 \times 10^{-5} - 1.1 \times 10^{-4}$ kWh (Samsi et al., 2023), while inference across commercial models ranging from 2B to 70B parameters consumes between 8.3×10^{-5} and 2.3×10^{-3} kWh per 100 tokens (Husom et al., 2024). Although these values may appear small in isolation, they accumulate rapidly at scale. As Calma (2023) note, the widespread integration of LLMs, such as their integration into search platforms, could increase the energy footprint per query by more than tenfold, underscoring the need for energy-efficient deployment strategies.

To contextualize these findings, we also consider the cost of human-led topic modeling, which is approximately \$48 per document per analyst (Carrell et al., 2016; Dernoncourt et al., 2017). Compared to this baseline, LLMs offer dramatic reductions in marginal financial cost per query. However, these monetary savings come with trade-offs: unlike human labor, LLM usage incurs measurable environmental impact that scales rapidly with deployment.

Moreover, since **our analysis draws from a diverse and evolving set of sources, both cost and energy estimates should be viewed as approximate benchmarks rather than fixed values**. These results underscore the importance of balancing cost-efficiency with sustainability when adopting AITM in educational research.

Setting	Coef. (s)	Std Err	z	p-value
Intercept (Human-Only)	383.7	94.8	4.1	<0.001
AI-Human	-1.6	132.8	-0.01	0.99
AI-Human-Deanchoring	-126.4	132.8	-0.95	0.34
AI-Only	**-382.7**	**156.7**	**-2.4**	**0.015**
Random Effects (Annotator):	Variance = 849.67			

Table 5: Linear mixed-effects model predicting annotation time (in seconds) across LLM support conditions with `Human-Only` as the reference category. The `AI-Only` condition significantly reduced annotation time, while partial AI support (`AI-Human`, `AI-Human-Deanchoring`) showed no statistically significant speed gains.

6 RQ2: Measuring Impact on Annotation Time

Methodology We used the data from stage 1 of the study to analyze annotation time.

For each human analyst, the total annotation time is calculated as:

$$\texttt{time}_a = \sum_{p=1}^{11} \texttt{time}_{ap} + 90 \text{ minutes}$$

Here, $\texttt{time}_{ap}$ denotes the time spent by analyst a on Plan p, and the additional 90 minutes accounts for two structured group discussions—one 30-minute within-treatment session and one 60-minute cross-treatment session.

In total, we collected 77 person-by-document entries: 6 human analysts $\times$ 11 Plans = 66 human entries, plus 11 entries from the `AI-Only` condition (1 AI $\times$ 11 Plans). To estimate the impact of treatment on time-on-task, we fit a linear mixed-effects model:

$$\texttt{time}_{ap} = \texttt{Treatment}_{ap} + \phi_a + \varepsilon_{ap}$$

where, $\texttt{time}_{ap}$ is the annotation time recorded by analyst a for Plan p, $\texttt{Treatment}_{ap}$ is a fixed effect with four levels: `Human-Only`, `AI-Only`, `AI-Human`, and `AI-Human-Deanchoring`, with `Human-Only` as the reference category. ϕ_a is a random intercept for each analyst (6 humans + 1 AI), which accounts for analyst-specific baseline differences and increases the precision of estimates, helping us isolate the impact of the treatment more reliably. ε_{ap} is the residual error term.

Results Table 5 presents the results of this analysis. The baseline annotation time in the `Human-Only` condition was approximately 384 seconds. The `AI-Human` condition showed virtually no difference in speed (Coef = -1.6 s, p = 0.99) relative to the `Human-Only` condition. The `AI-Human-Deanchoring` condition was faster by about 126 seconds relative to the `Human-Only` condition, but this difference was not statistically significant (p = 0.341). Notably, the `AI-Only` condition led to a statistically significant reduction of approximately 383 seconds (p = 0.015), representing a 6.4-minute decrease relative to the `Human-Only` condition. The random effect variance for annotators was estimated at 849.67, suggesting meaningful variability in baseline annotation speed between individuals. Some annotators were consistently faster or slower than others, regardless of treatment condition.

The **`AI-Only`** condition significantly reduces annotation time compared to **`Human-Only`**, suggesting that full AI support accelerates expert decision-making. However, **partial AI support** (i.e., `AI-Human` or `AI-Human-Deanchoring`) does not lead to statistically significant time savings. This indicates that the participants may have spent additional time reviewing and deliberating on the suggestions generated by the LLM. Rather than simply accepting AI outputs, Annotators have reportedly felt compelled to cross-check or validate these suggestions against their own judgment, leading to more careful and possibly slower decision-making. **This extra layer of comparison may have introduced hesitation or cognitive load, offsetting any potential efficiency gains from having AI support**. In contrast, participants in the Human-Only condition could rely solely on their intuition and expertise, resulting in a more streamlined workflow. This indicates that annotators may not gain measurable speed advantages unless they fully offload the task to the AI.

7 RQ3: Measuring Impact on Topic Identification

Methodology To evaluate how treatment condition influenced topic identification, we analyzed the Stage 2 annotation dataset described in Section 4. Each observation is a binary outcome indicating whether topic t was assigned to field f of plan p by annotator a. We fit the following multilevel linear probability model:

$$\Pr(\texttt{topic}_{fpat} = 1) = \texttt{treatment} + \eta_p + \varepsilon_{fpa}$$

Outcome	**Human-Only** (reference) Coef (SE)	**AI-Only** Coef (SE)	**AI-Human** Coef (SE)	**AI-Human-Deanchoring** Coef (SE)	**Joint Test of Treatments** (p-value)	**Plan RE** Variance (SE)
Academic Assessments	0.1979 (0.0715)	-0.0313 (0.0770)	0.0114 (0.0673)	0.0615 (0.0673)	0.6496	0.0344 (0.0171)
Academic Goals	0.3541 (0.0776)	-0.0763 (0.0906)	-0.0519 (0.0793)	-0.0105 (0.0793)	0.8054	0.0349 (0.0185)
Attendance	0.3098 (0.0877)	-0.1153 (0.0769)	-0.0360 (0.0674)	-0.0266 (0.0674)	0.5063	0.0654 (0.0297)
Behavioral Goals	0.1824 (0.0668)	-0.0435 (0.0717)	-0.0148 (0.0628)	0.0176 (0.0628)	0.8538	0.0301 (0.0150)
Classroom Management	0.0326 (0.0159)	-0.0326 (0.0242)	-0.0337 (0.0210)	-0.0140 (0.0210)	0.3577	0.0004 (0.0005)
College and Career Readiness	0.0505 (0.0394)	0.0051 (0.0408)	0.0078 (0.0357)	-0.0093 (0.0357)	0.9682	0.0110 (0.0054)
Curriculum	0.1067 (0.0556)	-0.0233 (0.0588)	-0.0090 (0.0515)	0.0390 (0.0515)	0.7016	0.0213 (0.0105)
Graduation	0.0167 (0.0159)	-0.0167 (0.0243)	-0.0009 (0.0212)	0.0009 (0.0212)	0.8886	0.0004 (0.0005)
Instruction	0.0674 (0.0335)	-0.0674 (0.0439)	-0.0246 (0.0383)	0.0056 (0.0383)	0.3472	0.0047 (0.0029)
Parent/Community Engagement	0.1982 (0.0699)	-0.0871 (0.0669)	-0.0418 (0.0585)	-0.0193 (0.0585)	0.6048	0.0383 (0.0179)
Professional Development	0.2266 (0.0786)	-0.0877 (0.0732)	-0.0348 (0.0641)	0.0550 (0.0641)	0.2369	0.0497 (0.0231)
Technology Use Integration	0.0756 (0.0636)	-0.0478 (0.0257)	-0.0223 (0.0226)	0.0122 (0.0226)	0.0935	0.0455 (0.0189)
Classroom Environment or Culture	0.1059 (0.0463)	-0.0503 (0.0510)	-0.0185 (0.0446)	-0.0491 (0.0446)	0.6521	0.0139 (0.0070)

Table 6: Coefficients (with SEs) from multilevel linear probability models estimating the impact of treatment on topic identification, relative to the Human-Only baseline. Joint tests assess whether all AI-based treatments collectively differ from the baseline. No statistically significant differences were observed across any treatment, indicating that topic identification remained stable despite varying levels of AI assistance.

Here, $\texttt{topic}_{fpat}$ is 1 if topic t was identified by analyst a in field f of plan p, and 0 otherwise. The model includes treatment as a fixed effect (with Human-Only as the reference condition) and η_p as a random intercept for each plan. This structure captures the hierarchical nature of the data while accounting for differences in topic prevalence across plans. ε_{fpa} accounts for the residual error.

We tested several alternative model specifications, including crossed and nested analyst effects, but these did not improve model fit or alter the results meaningfully. Thus, we retained the simpler formulation, which allows us to isolate the effect of treatment condition on topic identification behavior across annotators.

Results The results of the regression is given in Table 6. We used Human-Only as the reference condition and computed coefficients for each AI-based treatment: AI-Only , AI-Human, and AI-Human-Deanchoring. Each row in Table 6 presents the estimated probability of a topic being identified under each treatment, along with standard errors and joint significance test results.

For the topic *Academic Assessments*, the baseline Human-Only coefficient is 0.1979. Compared to this, the AI-Only coefficient is about 3 percentage points lower, the AI-Human coefficient is 1.1 percentage points higher, and the AI-Human-Deanchoring coefficient is 6.2 percentage points higher, respectively.

When comparing the Human-Only and AI-Human conditions reveals minimal differences across topics, with coefficients typically within ± 5 percentage points and no statistically significant deviations. This suggests that introducing AI support does not substantially shift topic identification patterns, and expert judgments remain largely consistent with the Human-Only baseline.

Next, examining the AI-Only and AI-Human conditions relative to the Human-Only baseline, we find that human analysts working with AI suggestions tend not to diverge far from the original AI-Only outputs. Instead, the AI-Human estimates tend to fall between the AI-Only and Human-Only values, implying that humans may be partially influenced— or anchored— by AI suggestions in their decision-making.

A similar pattern holds when comparing AI-Human and AI-Human-Deanchoring, each relative to the Human-Only baseline. Despite the presence of explicit deanchoring warnings, the estimates in these two conditions show minimal deviation from each other when considered through their differences from the baseline. In some cases, the deanchoring estimates are numerically closer but not statistically different to the AI-Human ones than to the Human-Only baseline. This indicates that, in this context, explicit instructions to critically evaluate AI suggestions had limited observable effect.

However, the joint significance test ($p = 0.6496$) does not indicate statistically significant differences between the treatment groups. This pattern holds across most topics. Joint significance tests across all 13 outcomes yielded p-values greater than 0.05, suggesting that the combination of effects from the three AI-based treatments does not reflect a systematic deviation from the Human-Only condition. In other words, there was no consistent pattern across the three AI conditions that significantly distinguished them from the Human-Only baseline.

The findings suggest that **while human annotators may incorporate AI input into their judgments, they are not significantly over-relying on it compared to the Human-Only condition.** Deanchoring prompts offered limited additional benefit in mitigating potential anchoring effects. **Topic identification remained stable across all treatment conditions, indicating that different approaches to incorporating AI did not produce meaningful divergence in these results.**

8 Conclusion

This study examined how AI-enabled topic modeling (AITM) can be integrated into educational research workflows, focusing on its financial, environmental, cognitive, and analytical trade-offs. Our findings show that while LLMs provide clear efficiency benefits, especially by speeding up annotation and lowering costs, these gains come with important risks. In both stages of human-in-the-loop annotation, we found evidence of anchoring bias: human analysts who saw LLM suggestions were more likely to stick with them, even when explicitly cautioned. However, when we looked at topic-level outcomes, we did not find statistically significant differences in which topics were identified across the treatment conditions. This suggests that while anchoring may shape how annotators approach the task, for example, in how quickly they work or how much they agree with AI, it doesn't necessarily change the final set of topics they choose.

As institutions consider scaling up AI-based analysis, the trade-off between speed and depth becomes harder to ignore. AI can definitely help efficiency and cost reduction, but human judgment is still crucial, especially for subtle, context-specific details that models tend to miss. Relying only on AI might make things more efficient, but it also risks losing the kinds of insights that matter most for real-world decisions. A balanced approach, where AI helps with the heavy lifting, but humans stay in the loop, seems like the best way to get both speed and substance.

Limitations

While this study offers important insights into the use of LLMs for topic modeling in educational research, it is essential to acknowledge its limitations. First, our analysis is based on a relatively small sample of 23 school improvement plans from a single state, which may limit the generalizability of our findings to other contexts. Second, our study focused on a specific type of text document. While these documents are relevant to educational leadership and policy, the findings may not be directly transferable to other forms of educational text, such as student essays, teacher evaluations, or policy documents. Third, our investigation of anchoring bias relied on a single de-anchoring intervention. While this allowed us to isolate the effect of such prompts, future research could explore the efficacy of other de-biasing techniques, such as structured protocols or collaborative decision-making strategies. Finally, the rapidly evolving nature of LLM pricing and energy consumption means that these figures of our cost analysis should be interpreted as indicative rather than definitive.

Acknowledgments

We are grateful to our study participants and anonymous reviewers for their valuable feedback, and also thank NEE, DESE for sharing their data and Dr. Thomas Hairston for his helpful input. This work was supported in part by a grant from the Minerva Research Initiative of the Department of Defense (Award No: FA9550-22-1-0171) and by the National Science Foundation under award IIS-2327143.

References

Abdulrahman M Al-Zahrani. 2024. Unveiling the shadows: Beyond the hype of ai in education. *Heliyon*, 10(9).

Abi Aryan, Aakash Kumar Nain, Andrew McMahon, Lucas Augusto Meyer, and Harpreet Singh Sahota. 2023. The costly dilemma: generalization, evaluation and cost-optimal deployment of large language models. *arXiv preprint arXiv:2308.08061*.

Claus Boye Asmussen and Charles Møller. 2019. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1):1–18.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Gavin Brookes and Tony McEnery. 2019. The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, 21(1):3–21.

Justine Calma. 2023. Ai is an energy hog, but deepseek could change that. Accessed: 2025-04-11.

Alexia Cambon, Brent Hecht, Benjamin Edelman, Donald Ngwe, Sonia Jaffe, Amy Heger, Mihaela Vorvoreanu, Sida Peng, Jake Hofman, Alex Farach, Margarita Bermejo-Cano, Eric Knudsen, James Bono, Hardik Sanghavi, Sofia Spatharioti, David Rothschild, Daniel G. Goldstein, Eirini Kalliamvakou, Peter Cihon, and 3 others. 2023. Early llm-based tools for enterprise information workers likely provide meaningful boosts to productivity. Technical Report MSR-TR-2023-43, Microsoft.

David S Carrell, David J Cronkite, Bradley A Malin, John S Aberdeen, and Lynette Hirschman. 2016. Is the juice worth the squeeze? costs and benefits of multiple human annotators for clinical text de-identification. *Methods of information in medicine*, 55(04):356–364.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *CoRR*, abs/2305.05176.

Alexander Choi, Syeda Sabrina Akter, J.P. Singh, and Antonios Anastasopoulos. 2024. The LLM effect: Are humans truly using LLMs, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054, Miami, Florida, USA. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Stijn Daenekindt and Jeroen Huisman. 2020. Mapping the scattered field of research on higher education. a correlated topic model of 17,000 articles, 1991–2018. *Higher Education*, 80(3):571–587.

DeepLearning.AI. 2024. Falling llm token prices and what they mean for ai companies. Accessed: 2025-04-11.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Franck Dernoncourt, Ji Young Lee, Özlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *J. Am. Medical Informatics Assoc.*, 24(3):596–606.

AMY Drahota, Rosemary D Meza, Brigitte Brikho, Meghan Naaf, Jasper A Estabillo, Emily D Gomez, Sarah F Vejnoska, Sarah Dufek, Aubyn C Stahmer, and Gregory A Aarons. 2016. Community-academic partnerships: A systematic review of the state of the literature and recommendations for future research. *The Milbank Quarterly*, 94(1):163–214.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024a. Collabcoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Sally Gao, Milda Norkute, and Abhinav Agrawal. 2024b. Evaluating interactive topic models in applied settings. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexis Horst, Brian D. Schwartz, Jenifer A. Fisher, Nicole Michels, and Lon J. Van Winkle. 2019. Selecting and performing service-learning in a team-based learning format fosters dissonance, reflective capacity, self-examination, bias mitigation, and compassionate behavior in prospective medical students. *International Journal of Environmental Research and Public Health*, 16(20).

Erik Johannes Husom, Arda Goknil, Lwin Khin Shar, and Sagar Sen. 2024. The price of prompting: Profiling energy use in large language models inference. *arXiv preprint arXiv:2407.16893*.

Chandra Irugalbandara, Ashish Mahendra, Roland Daynauth, Tharuka Kasthuri Arachchige, Jayanaka Dantanarayana, Krisztian Flautner, Lingjia Tang, Yiping Kang, and Jason Mars. 2024. Scaling down to scale up: A cost-benefit analysis of replacing openai's llm with open source slms in production. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 280–291.

Sonia Jaffe, Neha Parikh Shah, Jenna Butler, Alex Farach, Alexia Cambon, Brent Hecht, Michael Schwarz, and Jaime Teevan. 2024. Generative ai in real-world workplaces. Technical Report MSR-TR-2024-29, Microsoft.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.

Yan Liu and He Wang. 2024. Who on earth is using generative ai? Policy Research Working Paper 10870, World Bank. License: CC BY 3.0 IGO.

Larry Michaelsen, Arletta Knight, and L. Fink. 2002. *Team-based learning: a transformative use of small groups*.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):2053951716679679.

Douglas C Montgomery. 2017. *Design and analysis of experiments*. John wiley & sons.

Rosanna Nagtegaal, Lars Tummers, Mirko Noordegraaf, and Victor Bekkers. 2020. Designing to debias: Measuring and reducing public managers' anchoring bias. *Public Administration Review*, 80(4):565–576.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Cassandra Overney, Belén Saldías, Dimitra Dimitrakopoulou, and Deb Roy. 2024. Sensemate: An accessible and beginner-friendly human-ai platform for qualitative data analysis. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 922–939, New York, NY, USA. Association for Computing Machinery.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Anne C. Gill Rachael A. Hernandez, Paul Haidet and Cayla R. Teal. 2013. Fostering students' reflection about bias in healthcare: Cognitive dissonance and the role of personal and normative standards. *Medical Teacher*, 35(4):e1082–e1089. PMID: 23102159.

Philip Resnik. 2024. Large language models are biased because they are large language models. *Preprint*, arXiv:2406.13138.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI.

Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. 2023. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9. IEEE.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Min Sun, Jing Liu, Junmeng Zhu, and Zachary LeClair. 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.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Stephen Walther. 2024. How much does it cost to call openai apis? Accessed: 2025-04-11.

Yinying Wang, Alex J Bowers, and David J Fikis. 2017. Automated text data mining analysis of five decades of educational leadership research literature: Probabilistic topic modeling of eaq articles from 1965 to 2014. *Educational administration quarterly*, 53(2):289–323.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The promises and pitfalls of using language models to measure instruction quality in education. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4375–4389, Mexico City, Mexico. Association for Computational Linguistics.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024. Fact-and-reflection (FaR) improves confidence calibration of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8702–8718, Bangkok, Thailand. Association for Computational Linguistics.

A `AI-Human-Deanchoring` Warning

To address the anchoring bias minimally, we introduced a new treatment `AI-Human-Deanchoring`. Similar to the `AI-Human` setting, the `AI-Human-Deanchoring` group also received the results generated by AI (i.e., `AI-Only` list) paired with the following prominently displayed instructions:

```
This document has annotations suggested by
the LLMs. It will be your task to decide
whether these annotations are correct or
not. Delete or modify annotations as you
see fit. However, we have found evidence of
anchoring bias when annotators receive LLM
suggestions. Anchoring bias is a cognitive
bias where an individual relies too heavily
on an initial piece of information (the
“anchor”) when making decisions. This means
that the initial suggestions provided by
the LLM might disproportionately influence
the final labels you create, potentially
reducing the diversity and originality of
the Final Topic List. It is important for
you to be aware of this bias and make
conscious efforts to critically evaluate
and adjust your topics and suggestions to
ensure the annotations are accurate and
unbiased. We ask you to be extra critical
while annotating these documents.
```

Our intention was to observe how experts react to the awareness of anchoring bias from LLM suggestions and whether they adjust their behavior accordingly. We also aimed to evaluate if merely knowing about the bias was effective enough to help annotators de-anchor.

B Stage 1: All topic lists

B.1 AI-Only topic list

Topic Name	Topic Definition
Education	This topic encompasses various aspects of the educational process, including instructional strategies, curriculum development, assessment methods, professional development for educators, student performance tracking, and educational objectives and goals alignment with standards.
Student Assessment and Achievement	This topic covers the processes and methodologies involved in evaluating student performance, including standardized testing, reading assessments, and other forms of academic evaluation. It also includes strategies for improving student achievement levels in core subjects like math, ELA, and science.
Professional Development	This topic involves the continuous education and skill development of teachers and educational staff, including the implementation of best teaching practices, collaboration among educators, and the use of technology and data to enhance teaching effectiveness.
Curriculum and Instruction	This topic focuses on the design, implementation, and evaluation of educational curricula and instructional materials. It includes the alignment of curriculum with educational standards, the development of instructional strategies to meet diverse learning needs, and the integration of technology into the learning environment.
School Improvement Planning	This topic covers the strategic planning processes schools undertake to improve academic performance and operational efficiency. It includes setting and aligning goals with educational standards, data-driven decision-making, and the implementation of interventions and supports to meet educational objectives.
Behavioral Interventions and Supports	This topic addresses strategies and programs designed to improve student behavior and create positive school environments. It includes the implementation of Positive Behavior Interventions and Supports (PBIS), discipline management strategies, and efforts to increase student engagement and accountability.
Parent and Community Engagement	This topic involves strategies and practices for involving parents and the community in the educational process. It includes parent-teacher communication, community partnerships to support student achievement, and stakeholder involvement in school decision-making processes.
Educational Technology	This topic covers the use of technology in educational settings, including the implementation of digital tools and resources to support teaching and learning, the use of assessment technologies, and the training of educators in effective technology integration.

Table 7: AI-Only topic list for stage 1. We generated the list using GPT-4o-mini model using chatGPT API.

B.2 Final topic list after stage 1

Topic Name	Topic Definition
Academic Assessments	This topic includes mandated annual state assessments like MAP and other district and school level assessments to evaluate academic progress.
Academic Goals	This topic covers the strategic planning processes schools undertake in aligning goals with educational standards and the implementation of interventions and supports to meet educational objectives in core subjects like math, ELA, and science.
Behavioral Goals	This topic addresses strategies and programs designed to improve student behavior and create positive school environments. It includes the implementation of Positive Behavior Interventions and Supports (PBIS), discipline management strategies, and efforts to increase student engagement and accountability.
Classroom Management	This topic covers how teachers develop and implement procedures to maximize instructional time/space/transitions/activities for efficiency in the classroom.
Classroom Environment/Culture	This topic covers how all members of the school community (administrators, teachers, and students) develop and implement pro-social behaviors inside and outside of academic instruction. This can include social-emotional learning (SEL) and fostering of pro-social attitudes and behaviors.
Curriculum	This topic covers what teachers do to plan, design, and develop materials to promote learning. This can include collaboration through professional learning communities (PLCs) as long as it is specifically around curriculum design.
Instruction	This topic covers what teachers do to deliver instruction during active academic time with students in the classroom. This includes instructional strategies and also collaboration in professional learning communities (PLCs) as long as it is specifically about how teachers engage with students in academics, instructional strategies, academic press, critical thinking, or formative assessment.
Professional Development	This topic involves the continuous education and skill development of teachers and educational staff, including evaluation of teachers, classroom observation, and collaboration around improving what teachers do to work with students.
Parent/Community Engagement	This topic involves strategies and practices for involving parents and the community (including school boards) in the educational process. It includes parent-teacher communication, community partnerships to support student achievement, and stakeholder involvement in school decision-making processes.
Technology Use/Integration	This topic covers the use and integration of technological tools, resources, and materials.
College and Career Readiness (CCR)	This topic covers college and career readiness (CCR) of students including Career & Technical Education credit hours and employment, military, and college placement.
Graduation	This topic involves the matriculation between grades and completed secondary state requirements. This is often expressed in the graduation rates of students.
Attendance	This topic involves the attendance rates and percents of students.

Table 8: Stage 1 Final Topic List curated by the participants.

B.3 All Group-Specific Topic List

Human-Only List	AI-Human List	AI-Human-Deanchoring List	Final Topic List	AI-Only List
State Assessment	School Assessment and Achievement	Student Assessment and Achievement	Academic Assessments	Student Assessment and Achievement
Localized Assessment				
	School Assessment and Achievement	Student Assessment and Achievement	Academic Goals	
	Data-Driven Decisionmaking			
Behavioral Goals/ Classroom Management	Behavioral Interventions and Support	Behavioral Interventions and Supports	Behavioral Goals	Behavioral Interventions and Supports
Student Support	Data-Driven Decisionmaking			
		Classroom Management	Classroom Management	
Student/ Teacher Relationships		Classroom Culture/ Environment	Classroom Environment/ Culture	
Localized Curriculum	Curriculum and Instruction	Curriculum	Curriculum	Curriculum and Instruction
	Collaboration			
Teaching Strategies	Curriculum and Instruction	Instruction	Instruction	
Teacher Evaluation Components	Collaboration			
Professional Development	Professional Development	Professional Development	Professional Development	Professional Development
Instructional Coach				
Stakeholder Engagement	Parent and Community Engagement	Parent and Community Engagement	Parent/Community Engagement	Parent and Community Engagement
Technology Integration		Educational Technology	Technology Use/Integration	Education Technology
College, Career, Readiness			College and Career Readiness (CCR)	
Graduation/ Matriculation Rate			Graduation	
Attendance			Attendance	
	~~District Alignment~~	~~Education~~		~~Education~~
		~~School Improvement Planning~~		~~School Improvement Planning~~

Table 9: Comparison of topic lists generated across conditions in Stage 1. Entries are grouped to show thematic overlap and consolidation across all lists. Struckthrough entries indicate topics that annotators collectively decided to discard during the final discussion phase.

C Stage-2 Detailed Results:

We provide computation details for the metrics reported in Table 3. For the analysis, each paragraph-level field was encoded as a 14-dimensional binary vector: 13 dimensions correspond to the presence or absence of each topic from the Final Topic List, and the final slot indicates a "None" label (no topic assigned). These vectors were used for computing precision, recall, and agreement metrics.

Annotators	precision	recall
A1	0.57	0.5
A2	0.71	0.67
A3	0.77	0.47
A4	0.70	0.44
A5	0.65	0.65
Avg	**0.68**	**0.55**

Table 10: For each annotator in Stage 2, the precision and recall percentages of the `AI-Only` annotations over these documents when measured against the annotations of experts acting under the Human-Only condition. Also, the averages of these LLM precision and recall percentages.

Average Precision and Recall To evaluate how closely LLM-generated annotations align with human judgment, we compute precision and recall by comparing the LLM-assigned topics to those assigned by human annotators under each treatment condition (`Human-Only`, `AI-Human`, and `AI-Human-Deanchoring`).

Using the `Human-Only` condition as ground truth, we found that the LLM achieved an average precision of 0.68 and a recall of 0.55. This means that while 68% of LLM predictions aligned with expert judgments, nearly half of the expert-identified topics were not captured by the model. Thus, the LLM shows reasonable accuracy, but limited coverage in replicating full expert insight.

	Human-Only	**AI-Human**	**AI-Human-Deanchoring**
Average Annotation Speed (words/min)	73.75	71.15	89.91
Average Annotator Agreement with AI (%)	54.64	73.44	71.41

Table 11: Comparison of average annotation speed (words per minute) and average Human-AI agreement across the three conditions.

Average Annotation Speed To understand how LLM support affects efficiency, we calculated annotation speed in words per minute (wpm). For each document field, we divided the number of words by the time each annotator took to complete it, then averaged these speeds by condition. As shown in Table 11, annotators in the `Human-Only` condition averaged 73.75 wpm. This dipped slightly in the `AI-Human` condition to 71.15 wpm, but surprisingly jumped to 89.91 wpm in the `AI-Human-Deanchoring` condition—even though those annotators were explicitly warned about bias. The results suggest that having AI suggestions, even with cautionary prompts, may encourage annotators to move faster—possibly by relying on the AI's suggestions rather than thinking through every decision from scratch.

Average Annotator Agreement with AI To assess how closely human annotators aligned with LLM-generated suggestions, we calculated the percentage of topic assignments that matched the `AI-Only` output. For each annotator–field pair, we compared the human-assigned topics to the AI's and computed the overlap. These agreement scores were then averaged within each condition (see Table 11).

Agreement varied by condition. In the `Human-Only` setting—where annotators had no AI support—the average agreement with the AI was 54.64%. This jumped to 73.44% in the `AI-Human` condition, suggesting that access to AI suggestions substantially influenced annotator decisions. In the `AI-Human-Deanchoring` condition, agreement remained similarly high at 71.41%, even though annotators were explicitly warned about potential bias. This suggests that simply cautioning annotators may not be enough to counter the influence of LLM outputs.

Inter-Annotator Agreement. To assess how consistently annotators applied the topic labels, we used Cohen's κ(Cohen, 1960), a standard measure for inter-rater agreement on categorical decisions. Because each document field was annotated by a pair of analysts within the same condition (see Table**??**), we were able to compute pairwise κ scores for each condition and then average them.

The results (Table 12) show that annotators aligned more closely when LLM suggestions were available. Agreement was highest in the `AI-Human` condition ($\kappa = 0.71$) and nearly as high in the `AI-Human-Deanchoring` setting ($\kappa = 0.69$). In

Agreement between	Human-Only	AI-Human	AI-Human-Deanchoring	Avg per Annotator
A1 and A4	0.48	0.72	0.79	**0.66**
A2 and A5	0.65	0.69	0.59	**0.64**
Avg per Condition	**0.57**	**0.71**	**0.69**	

Table 12: Agreement between annotator pairs across different treatment conditions. We report annotator agreement Cohen's κ for each pair per setting. The average agreement per annotator pair is higher for the settings with LLM suggestions, implying towards a potential anchoring effect.

contrast, agreement dropped in the `Human-Only` condition ($\kappa = 0.57$), where annotators worked independently. These findings suggest that LLM support—regardless of deanchoring prompts—tends to guide annotators toward similar decisions, potentially reflecting a convergence effect around AI-generated suggestions.