UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

Exploring Generative Artificial Intelligence for Data Extraction in Single Case Design Meta-analysis

Yaosheng Lou, Mariola Moeyaert, Piet Wesling, Benjamin Solomon, Xin Li

Brief Introduction

- Single-case experimental designs (SCEDs) is rigorous experimental design involving repeated measures over time for evaluating intervention effectiveness.
- Meta-analysis plays an essential role in aggregating research evidence to enhance the generalizability of findings.
- **The substantial time investment** is one of the major challenges for researchers, particularly during the data extraction process.
- In SCED meta-analyses, researchers must extract both textual and graphical data.

Two Components of Data Extraction

Textual Data Extraction:

Data on variables need to be extracted from the full-text of articles.

Method

Participants

Three elementary school students with severe disabilities participated in the study. Participant criteria included: (a) ability to verbally answer questions or point to words/pictures to answer questions, (b) having an IQ of 55 or below and meeting the federal criteria for intellectual disability (c) ability to stay in an instructional area for 10 minutes; and (d) teacher recommendation. Two boys and one girl, ages 10 to 12 years old, met these criteria. The special education teacher helped procure parental consent.

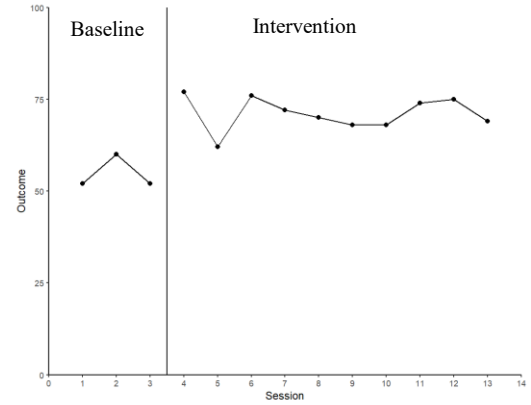
All of the students were classified as having intellectual disability. All participants were in Grade 5. Juan is a Hispanic male with a traumatic brain injury. Dalton is a Caucasian male and Jackie a Caucasian female; both have Down syndrome. Juan used spoken language to communicate while Dalton and Jackie used spoken language and pictures. All students could read some sight words, but only Juan had basic word decoding skills. Juan and Dalton were able to answer some literal comprehension questions with picture support at the end of a short passage. Jackie could answer some simple literal comprehension questions after a sentence was read. All students had IEP (Individualized Education Program) goals related to comprehension and reading words. Please see Table 1 for additional demographic information.

The interventionist who implemented the study and collected the primary data was a second-year doctoral student in special education with eight years teaching experience in public schools with students with severe disabilities. Inter-observer reliability and procedural fidelity data were collected by a graduate assistant.

(An example from Courtade et al., 2017)

Graphical Data Extraction:

Data points of SCED graphs need to be extracted to calculate effect sizes.



Study	Author	Year	Study Design	Intervention Program	Outcome Domain	Case	Age	Grade	Gender	Disability Status
1	Alison et a	2017	4	Shared story reading	Literacy	1	8	2	M	Autism Spectrum Disorder (ASD)
1	Alison et a	2017	4	Shared story reading	Literacy	2	8	2	M	Autism Spectrum Disorder (ASD)
1	Alison et a	2017	4	Shared story reading	Literacy	3	10	4	M	Autism Spectrum Disorder (ASD)
2	Browder e	2011	4	Scripted task analytic le	Literacy; Engagement	1	8	NA	F	severe intellectual disability and cerebral palsy
2	Browder e	2011	4	Scripted task analytic le	Literacy; Engagement	2	9	NA	F	severe intellectual disability and legally blind
2	Browder e	2011	4	Scripted task analytic le	Literacy; Engagement	3	6	NA	M	severe intellectual disability and cerebral palsy
3	Browder e	2015	4	electronic story-mappi	Literacy	1	8	2	M	Autism Spectrum Disorder
3	Browder e	2015	4	electronic story-mappi	Literacy	2	9	3	M	Autism Spectrum Disorder
3	Browder e	2015	4	electronic story-mappi	Literacy	3	10	4	F	Autism Spectrum Disorder
4	Cheek et a	2019	2	Online module plus eC	Literacy; Engagement	1	9	NA	M	SID, autism, cerebral palsy (CP), and hearing
4	Cheek et a	2019	2	Online module plus eC	Literacy; Engagement	2	7	NA	F	SID,autism, and speech delays
4	Cheek et a	2019	2	Online module plus eC	Literacy; Engagement	3	9	NA	M	SID, autism, andattention deficit hyperactivity disorder
5	Collins et a	2019	4	Computer-Aided Listen	Literacy, communicati	1	21	12	M	autism spectrum disorder (ASD) and intellectual disability
5	Collins et a	2019	4	Computer-Aided Listen	Literacy, communicati	2	16	10	M	autism spectrum disorder (ASD) and intellectual disability
5	Collins et a	2019	4	Computer-Aided Listen	Literacy, communicati	3	21	12	M	autism spectrum disorder (ASD) and intellectual disability
6	Courtade e	2017	4	Read-Alouds of Grade-l	Literacy	1	10	5	M	MID; traumatic brain injury
6	Courtade e	2017	4	Read-Alouds of Grade-l	Literacy	2	12	5	M	MID; Down syndrome
6	Courtade e	2017	4	Read-Alouds of Grade-l	Literacy	3	10	5	F	MID; Down syndrome

Graph ID	Session	Outcome	Phase
1_1	1	52	0
1_1	2	61	0
1_1	3	52	0
1_1	4	78	1
1_1	5	63	1
1_1	6	77	1
1_1	7	73	1
1_1	8	71	1
1_1	9	69	1
1_1	10	69	1
1_1	11	74	1
1_1	12	75	1
1_1	13	70	1
1_2	1	71	0
1_2	2	64	0
1_2	3	71	0
1_2	4	79	1

Previous Literature

Textual Data Extraction

- To date, many studies have examined the potential of GenAI in textual data extraction. Studies have found that the performance of GenAI can vary depending on several factors:
  - Prompt (Hamed et al., 2023)
  - Complexity of variables (Motzfeldt Jensen et al., 2025)
  - Study domains (Schmidt et al. 2024):
    - human clinical studies > animal studies > human social science studies

- No study has reported the accuracy of textual data extraction in the context of SCEDs.

Graphical Data Extraction

- To best of our knowledge, no study has evaluated the potential of GenAI for extracting data from longitudinal graphs (e.g., SCED graphs).

Research Questions

Textual Data Extraction:

- How does GenAI perform in extracting textual data from SCED studies?
- What are the differences in extraction accuracy between manual extraction and different GenAI models?

Graphical Data Extraction:

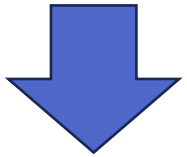
- How does GenAI perform in extracting graphical data from SCED graphs?
- Does GenAI's accuracy in graphical data extraction vary depending on graph characteristics (e.g., the number of data points in baseline/intervention, trend, variability, effect size, and graph type)?

Textual Data Extraction

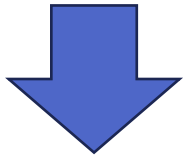
Selection of Materials

Studies for textual data extraction were drawn from SCED meta-analyses identified through a systematic search in Web of Science:

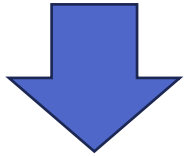
(a) a meta-analysis including SCED studies;
(b) published in 2024;
(c) peer-reviewed with full-text availability
(d) related to the field of education, such as “skill development”, “reading intervention”
(e) available in English.



This search resulted in six eligible meta-analyses.



These six meta-analyses included 106 primary studies. After excluding non-SCED studies and replications, **93** unique SCED studies remained for testing in this study.



Download 93 SCED studies in PDF format for data extraction

Prompt for Textual Data Extraction

You are a data extraction and information retrieval specialist. Your task is to extract information strictly from the uploaded PDFs (I will upload them later, one by one). If a required field is not explicitly present in the text, enter NA. Do not infer, guess, or use general knowledge. Model settings: Set temperature = 0.

Output file:

1. Create an Excel with one sheet data.
2. Row 1 headers (A–L): Study | Author | Year | Study Design | Intervention Program | Outcome Domain | Case | Age | Grade | Gender | Disability Status | Physical Setting
3. Start data at row 2. Each participant = one row. Duplicate study-level variables across all participants in that study. Study-level variables include Study, Author, Year, Study Design, Intervention Program, Outcome Domain.

Filling Rules:

- **Study:** Assign plain numbers 1, 2, 3… in the order the PDFs are uploaded/processed.
- **Author**: Extract the names of the authors. Use APA in-text format. For example, for 1–2 authors: "Smith & Lee"; for 3 or more: "Smith et al."
- **Year**: Use the publication year. If multiple, prefer the one labeled "published."
- **Study Design** (Enter numeric code only from list below):
  1= Reversal design
  2= Multiple baseline design
  3= Alternating treatments design
  4= Multiple probes design
  5= Changing criterion design
  6= Combined SCED designs (hybrids of two or more designs)
  7= AB/ABA
  8= Other
- **Intervention Program:** Use the exact program name reported.
- **Outcome Domain:** Choose from the list. Separate multiple with semicolons (e.g., "Literacy; Engagement").
  Math (e.g., calculation, counting, numeracy)
  Literacy (e.g., reading, writing, listening, comprehension)
  Communication/social skills (e.g., conversation, requesting, play interaction,)
  Problem behavior (e.g., off-task, disruption, aggression, noncompliance)
  Engagement (e.g., on-task, task completion, compliance, activity engagement)
  Others (specify in parentheses, e.g., Others (acceptance of foods))
- **Case:** Within each study, assign 1, 2, 3… in the order participants are described.
- **Age**: Extract each participant's age in years only. If not reported, NA.
- **Grade**: Use standardized codes (PreK, K, 1, 2, …, 12, Postsecondary, higher ed). If not reported, NA.
- **Gender**: Extract each participant's gender. If not reported, NA.
- **Disability Status:** Record verbatim diagnostic category/eligibility (e.g., ASD, ADHD, ID). If multiple labels per participant, separate with semicolons.
- **Physical Setting** (If multiple locations fall within the same category, code using that single category number. Use code 5 only when locations span across different categories. If not reported, enter NA.)
  1= Classroom/school/University
  2= Home
  3= Clinic/lab
  4= Multiple/combination
  5= Others

Methods for Textual Data Extraction

Manual Extraction

- Two coders with experience in conducting SCED meta-analyses extract the data independently
- The extracted datasets were compared and merged into one dataset.

GenAI Extractions

- ChatGPT 5
- Gemini 2.5 Pro
- Claude Opus 4.1

Gold Standard Dataset

After extracting data using four methods, two human coders reviewed all discrepancies and established the gold standard dataset through discussion.

Results for Textual Data Extraction

Each extracted dataset was then compared with the gold standard.

Classify	Data Extraction Performance
True positive	All relevant data was found and correct
False positive	Relevant data was found, but some was missing
	Relevant data was found, but some was incorrect
	Data that was incorrect was found
False negative	Data exist but were not detected (i.e., reported as “NA”)

Metric	Equation	Meaning
Precision	$\text{Precision} = \frac{TP}{TP + FP}$	How many of the extracted results were correct.
Recall	$\text{Recall} = \frac{TP}{TP + FN}$	How many of the correct results were actually found.
F1 Score	$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$	Harmonic mean of Precision and Recall.

Results for Textual Data Extraction

A total of 2,068 values were extracted across 93 articles

	Precision	Recall	F1 Score
Manual Extraction	98.35%	99.56%	98.95%
ChatGPT	84.79%	99.83%	91.70%
Gemini	97.09%	99.70%	98.38%
Claude	96.23%	98.84%	97.52%

- **Manual Extraction** remains the method with highest precision (98.35%) and good recall (99.56%).
- **ChatGPT** demonstrated the highest recall (99.83%), meaning it seldom misses relevant items, but relatively low precision (84.79%).
- **Gemini** performs very close to the manual extraction, with high precision (97.09%) and recall (99.70%). There is only a 0.57% difference in F1 score relative to manual extraction.
- **Claude** also demonstrated great performance (Precision 96.23%, Recall 98.84%), with an F1-score difference of 0.86% relative to Gemini and 1.43% relative to manual extraction.

Graphical Data Extraction

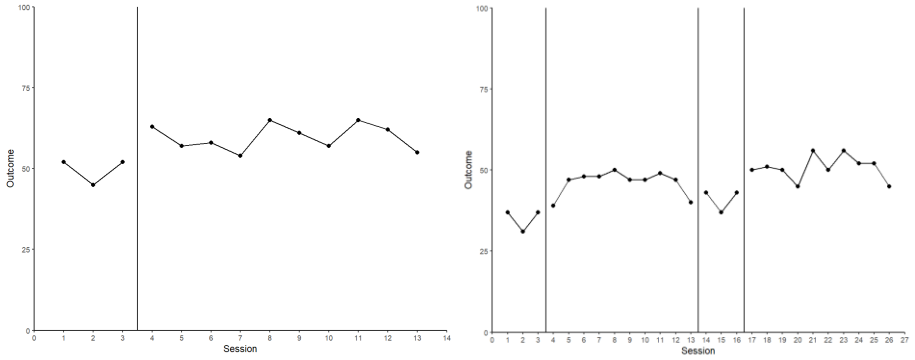
Methods for Graphical Data Extraction

640 graphs were simulated

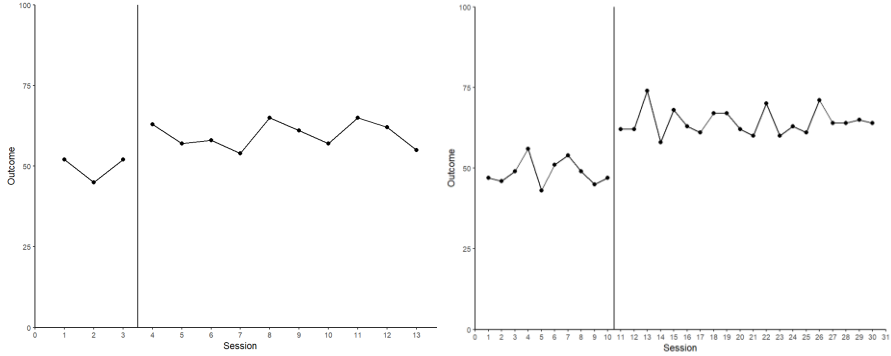
- 64 conditions (2 x 2 x 2 x 2 x 2 x 2) in total: representing the most basic yet realistic combination
- 10 graphs per condition were simulated

Graph Characteristics	Value
Number of data points in baseline phase	3, 10
Number of data points in intervention phase	10, 20
Trend (only intervention phase)	No trend; Trend (30 degrees)
Variability	SD = 4; SD = 10;
Effect size	No effect: SMD = 0; With effect: SMD = 3;
Graph Type	AB graph; ABAB graph

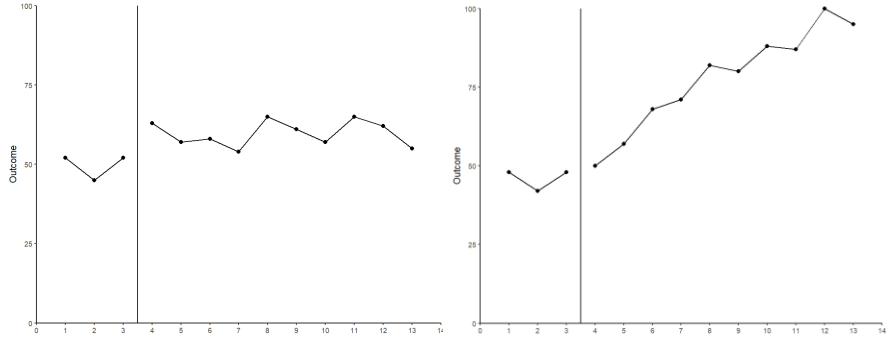
Graph Type



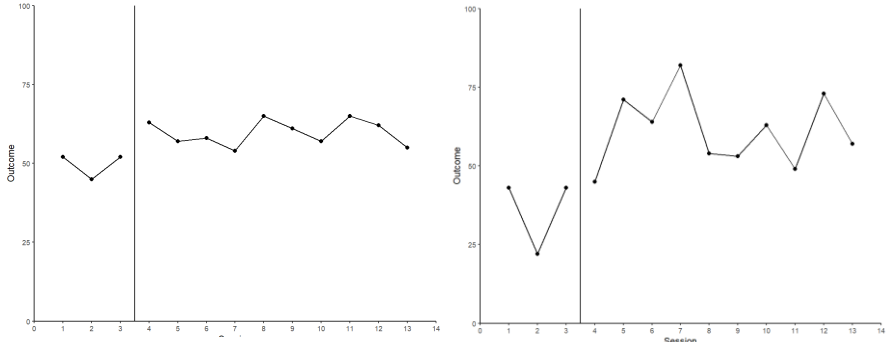
data points



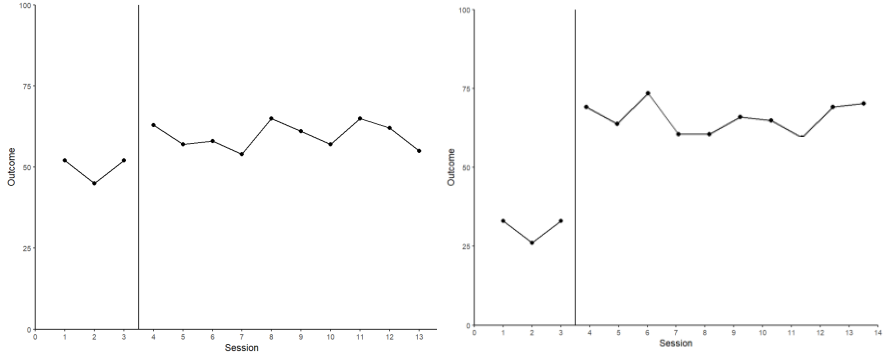
Trend



Variability



Effect size



Prompt for Graphical Data Extraction

You are an expert in quantitative data extraction from single-case experimental design (SCED) graphs.

Task:

Extract all visible data points from the uploaded line graph image.

Each point represents one session (x-axis) and one outcome (y-axis).

Determine each point's phase based on visual cues (e.g., shaded areas, vertical dashed lines, or phase labels):

- Phase 0 = Baseline phase (A)

- Phase 1 = Intervention phase (B)

Follow these rules carefully:

1. Estimate each point's Session (x-axis) and Outcome (y-axis) values to two decimal places using the axis scales visible in the image.

2. Identify phases based on shading, vertical dividers, or labels.

3. Output the results in a Markdown table with the following headers: Graph ID | Session | Outcome | Phase

5. Graph ID = the filename (without the extension).

6. Do not describe or explain the graphs. Only output the final combined Markdown table.

Methods for Graphical Data Extraction

Only GenAI extractions

- ChatGPT 5
- Gemini 2.5 Pro
- Claude Opus 4.1

Lin's Concordance Correlation Coefficient (CCC) was used to evaluate the performance of GenAI models in each graph characteristics

- Inter-coder reliability for continuous data (McBride, 2005)

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

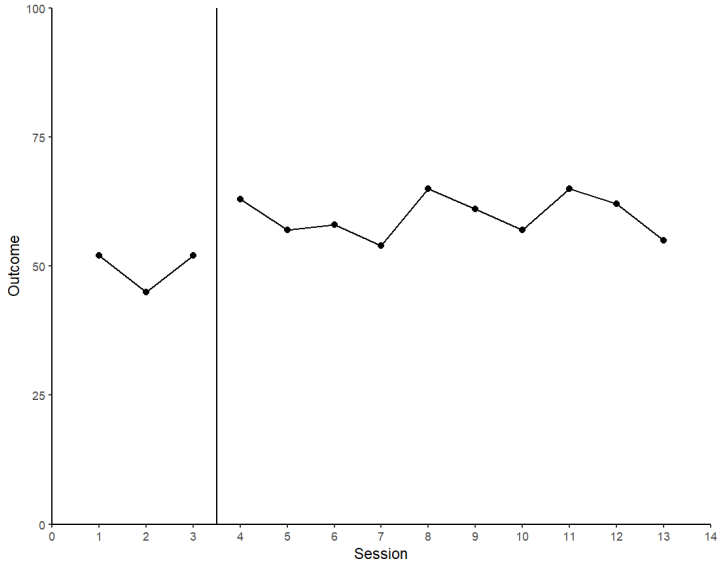
σ_x, σ_y: standard deviations of x and y
μ_x, μ_y: means of x and y

- ρ_c = 1: perfect agreement
- ρ_c = 0: no agreement

Methods for Graphical Data Extraction

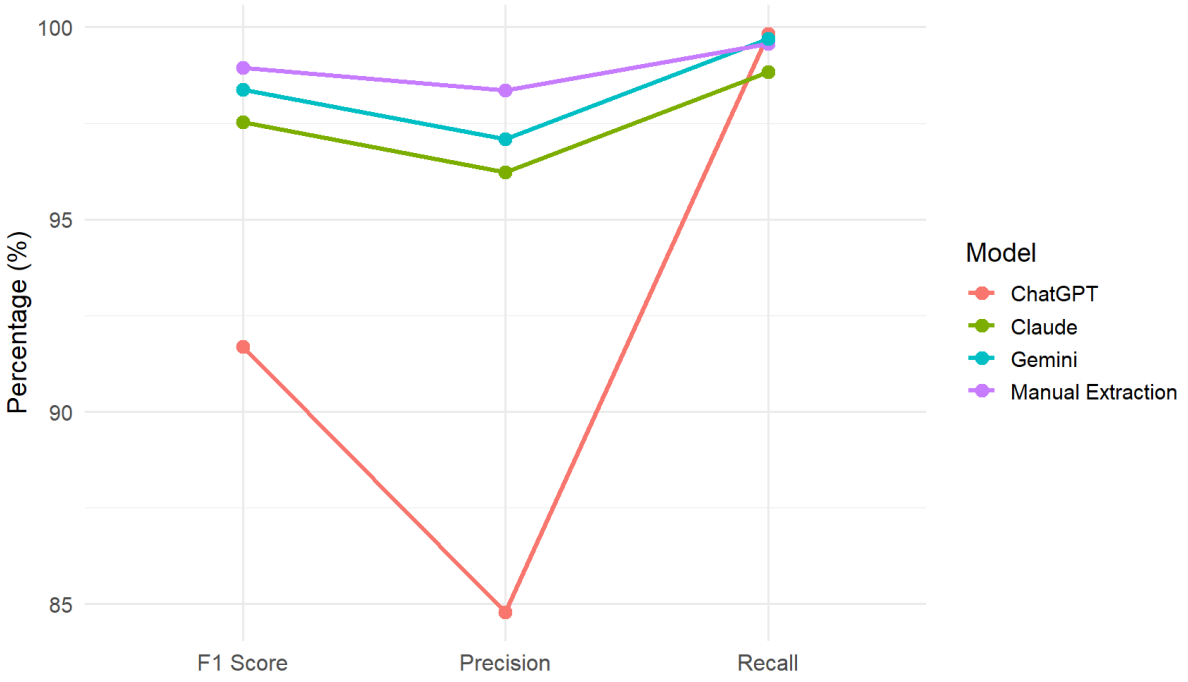
	Overall	# of baseline		# of intervention		Trend		Variability		Effect size		Graph Type	
		3	10	10	20	0°	30°	SD=4	SD=10	SMD=0	SMD=3	AB	ABAB
ChatGPT	0.890	0.867	0.903	0.926	0.868	0.883	0.894	0.945	0.857	0.901	0.878	0.954	0.849
Gemini	0.974	0.984	0.973	0.991	0.964	0.971	0.976	0.985	0.967	0.978	0.970	0.997	0.960
Claude	0.964	0.975	0.956	0.978	0.955	0.952	0.973	0.986	0.951	0.965	0.962	0.999	0.941

- Lin' CCC > 0.95 was considered acceptable (McBride, 2005)
- Gemini achieved the highest overall accuracy, whereas ChatGPT demonstrated the lowest
- Graphs with a smaller number of data points (e.g., three data points in the baseline phase, ten in the intervention phase) and lower variability tended to have higher accuracy
- Effect size and trend did not influence accuracy
- AB graphs consistently showed greater extraction accuracy than ABAB graphs

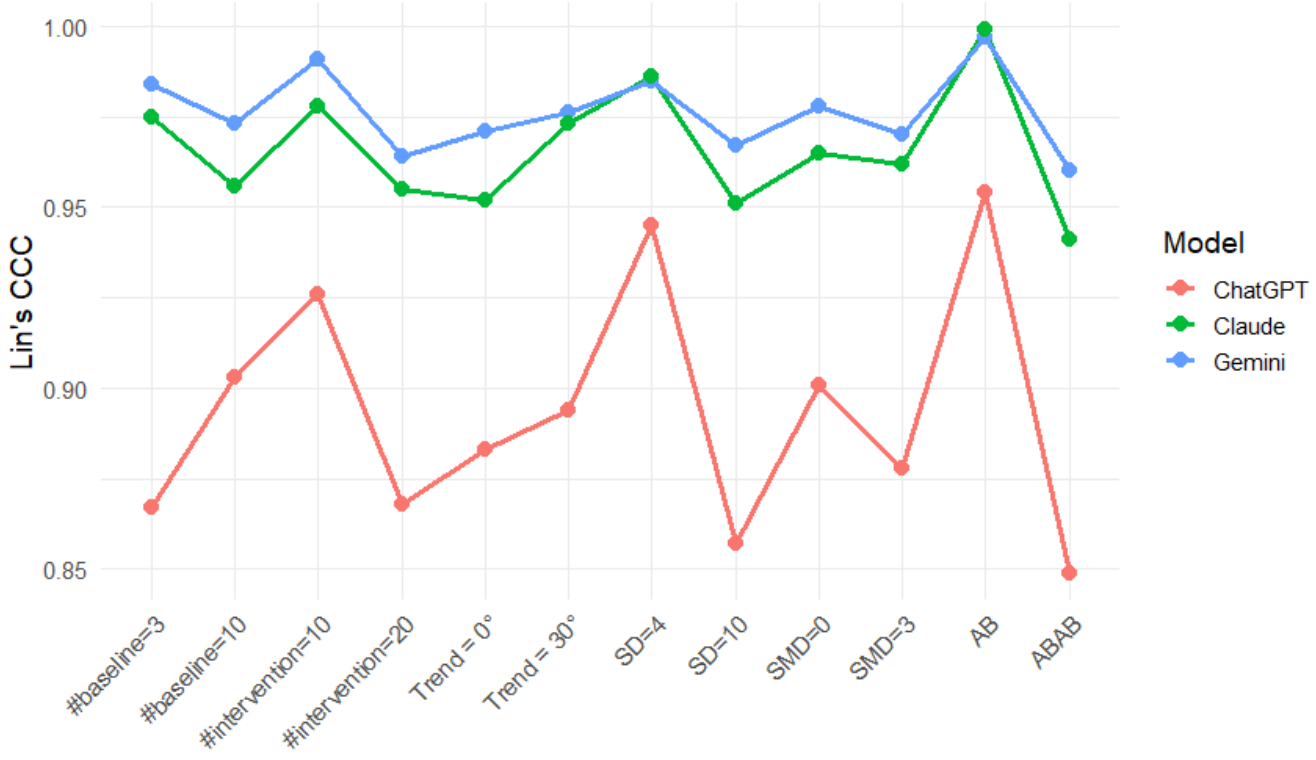


Visualization

Textual Data Extraction



Graphical Data Extraction



UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

Questions?

Yaosheng Lou

Email: ylou@albany.edu