HARVARD

GRADUATE SCHOOL OF EDUCATION

ChatGPT vs. Machine Learning:
Assessing the Efficacy and Accuracy of Large Language Models for Automated Essay Scoring

Postdoctoral Fellow (Harvard University), Youngwon Kim

Assistant Professor (Bentley University), Reagan Mozer

Associate Professor(Harvard University), Luke W. Miratrix

Assistant Professor (Michigan State University), Shireen Al-Adeimi

Trustworthy AI Lab for Education Summer Online Symposium (2024)

6/11/2024

Objectives

- Background
- Research Questions
- Data Source
- Measures
- Methods
- Results
- Discussion
- Limitations & Future Studies

Background

Why Automated Essay Scoring (AES) Matters?

Background

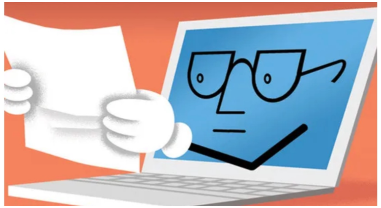
Why Automated Essay Scoring (AES) Matters?

- **Human grading:** Labor-intensive, time-consuming, and potentially susceptible to bias (Ramesh & Sanampudi, 2022)

Background

Why Automated Essay Scoring (AES) Matters?

- **Human grading:** Labor-intensive, time-consuming, and potentially susceptible to bias (Ramesh & Sanampudi, 2022)
- **AES:** Use of technology to evaluate and score written essays



Background

Why Automated Essay Scoring (AES) Matters?

- **Human grading:** Labor-intensive, time-consuming, and potentially susceptible to bias (Ramesh & Sanampudi, 2022)
- **AES:** Use of technology to evaluate and score written essays
- Entirely eliminating human grading efforts remains impractical in most real-world educational scenarios (Weegar & Idestam-Almquist, 2023)

Background

One of Current AES Approaches (Mozer & Miratrix, 2023)

- rcttext package: text analysis within randomized controlled trials

quanteda LIWC TAACO

Background

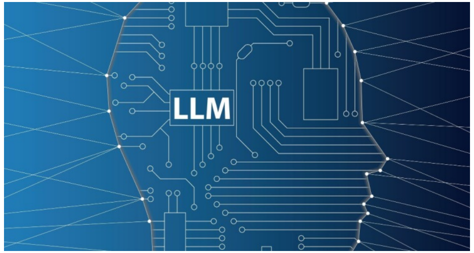
One of Current AES Approaches **(Mozer & Miratrix, 2023)**

- rcttext package: text analysis within randomized controlled trials
- Automated text feature extraction: Natural Language Processing tools (e.g., quanteda, LIWC and TACCO) analyze existing essays.
- Machine Learning Prediction: ML techniques predict scores based on the extracted text features.

Background

Emergence of Large Language Models (LLMs)

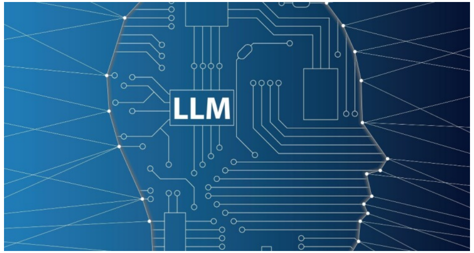
- **ChatGPT (since 2022):** Answers human questions with an AI that seems to have a perfect understanding of the language
- Potential benefits over existing ML methods



Background

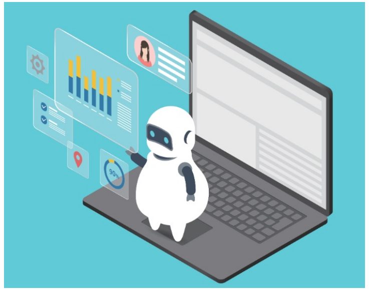
Emergence of Large Language Models (LLMs)

- **ChatGPT (since 2022):** Answers human questions with an AI that seems to have a perfect understanding of the language
- Potential benefits over existing ML methods
- Previous studies:
  - Successful: Scoring categorical outcomes, such as helpful/harmful (Touvron et al., 2023), preference (Lee et al., 2023), and polite/impolite (Ludwig et al., 2021)
  - Unsuccessful: Scoring continuous outcomes (Ludwig et al., 2021; Mayfield & Black, 2022)



Research Questions

1. How well does ML grading and LLM grading work for assessing student essays, as compared to a human-scored gold standard?
2. Which methods exhibit superior performance in grading for categorical and for continuous outcomes?



Data Source

Catalyzing Comprehension Discussion and Debate (CCDD) study **(Snow et al., 2009)**

- WG study: Persuasive essays written by grade 4-8 students about whether iPads should be added to their school
- A team of 7 research assistants, experienced in English language teaching, scored and classified the essays (Gold Standard)
- Out of the initial 3,542 essays, 2,687 were selected
- These essays were sourced from 23 schools (13 control, 10 treatment)

Measures

Quality of Writing Assessment (Continuous)

- Holistic Writing Rubric (NAEP, 2017): 1) Development of Ideas, 2) Organization, 3) Language Facility and Convention
- 7-point scale with higher scores indicating greater quality (1-7)

Measures

Quality of Writing Assessment (Continuous)

- Holistic Writing Rubric (NAEP, 2017): 1) Development of Ideas, 2) Organization, 3) Language Facility and Convention
- 7-point scale with higher scores indicating greater quality (1-7)

Essays Opinion Classification (Categorical)

- 5 distinct opinions on iPad usage in schools:
  - Affirmative (Allow iPads in school)
  - Negative (Do not allow iPads in school)
  - Other: Balanced (Allow iPads in school with restrictions)
    Ambivalent (Not clear on stance)
    No Argument (Not argumentative stance and off-topic)

Methods

Machine Learning (ML) Modeling

- Tree-based Models (trained on 5% to 90% of data)
  - Random Forest (RF), Regularized Random Forest (RRF)
  - Stochastic Gradient Boosting (GBM), Extreme Gradient Boosting (XGBOOST)

Methods

Machine Learning (ML) Modeling

- Tree-based Models (trained on 5% to 90% of data)
  - Random Forest (RF), Regularized Random Forest (RRF)
  - Stochastic Gradient Boosting (GBM), Extreme Gradient Boosting (XGBOOST)

Large Language Models (LLMs)

- API: ChatGPT 3.5-Turbo-0615 (older) & ChatGPT 3.5-Turbo-0125 (updated)
- Types of prompts: Base, Few-shot, Few-shot + Chain-of-Thought (CoT)
- No Fine-tuning vs Fine-tuning (90 essays for essay classification, 91 essays for writing quality)

Methods

3 Different Types of Prompts

Zero-Shot	Few-shot	Few-shot + CoT
Evaluate and score the overall quality of the essay on iPad usage.	Evaluate and score the overall quality of the essay on iPad usage. + Labeled Examples	Evaluate and score the overall quality of the essay on iPad usage. + Labeled Examples + When evaluating and scoring the given text, consider three criteria and the examples above.

Methods

Prompt Examples - Base

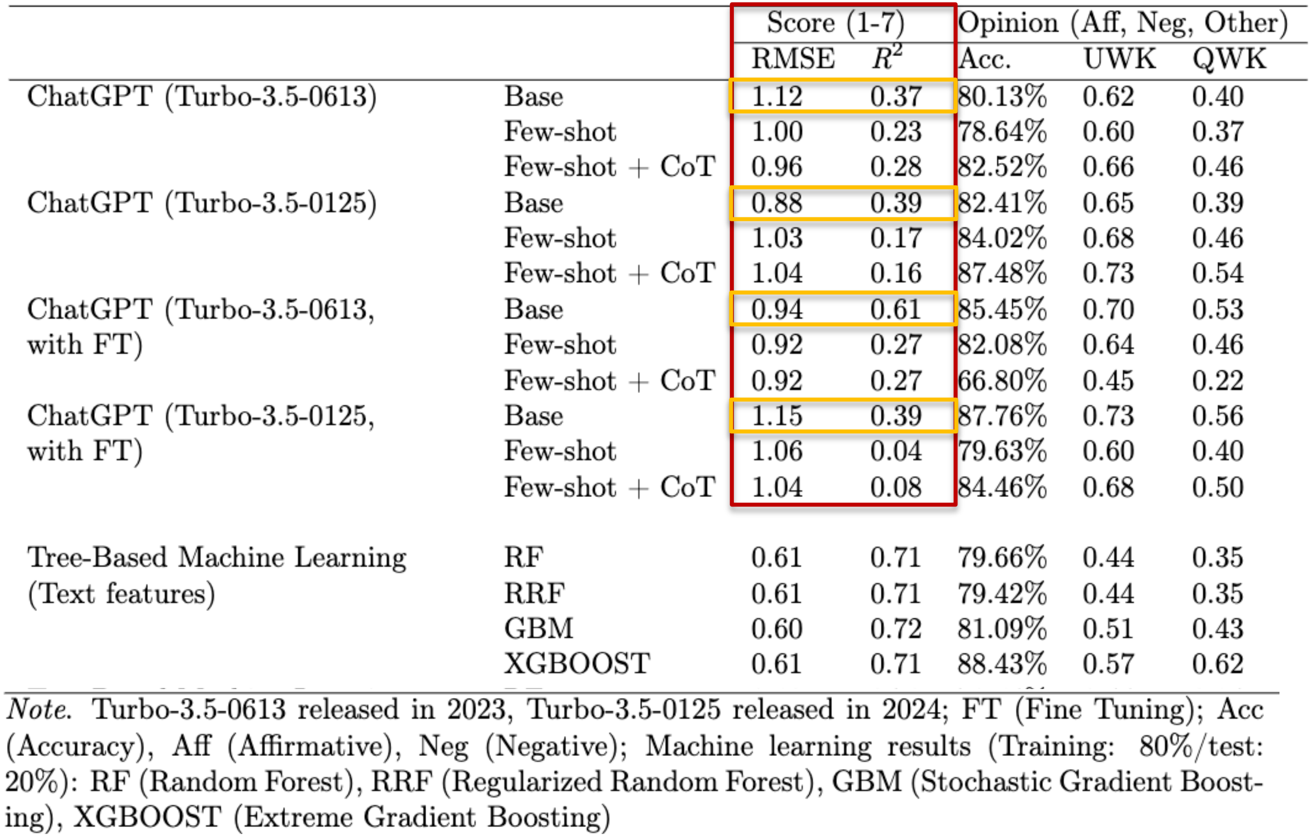
Quality of Writing	Essay Classification
You are an expert essay grader for students in grades 4-7. The evaluation should consider three criteria: 1) Development of Ideas, measuring the depth, complexity, and richness of details and examples; 2) Organization, focusing on the logical structure, coherence, and overall focus of ideas; 3) Language facility and convention, evaluating clarity, effectiveness in sentence structure, word choice, voice, tone, grammar, usage, and mechanics. In the given text, evaluate and score the overall quality of the essay on iPad usage in schools. Use a 7-point scale, where a higher score indicates greater quality. Present your response as only the numeric score.	You are an expert essay grader for students in grades 4-7. In the given text, evaluate and categorize the stance on iPad usage in schools into one of the following: 1) Allow iPads in school (AFF), 2) Do not allow iPads in school (NEG), 3) or if the essay does not fit into either of these categories (OTHER). Present your response as either AFF, NEG, or OTHER.

Methods

Evaluation Metrics

- Quality of writing
  - RMSE
  - R^2: 0 to 1
- Essay classification
  - Accuracy
  - Unweighted Kappa (UWK):
    - Ranges: -1 to 1
    - 0–0.20 slight | 0.21–0.40 fair | 0.41–0.6 moderate | 0.61–0.80 substantial | 0.81–1 almost perfect (Landis & Koch, 1977)
  - Quadratic Weighted Kappa(QWK):
    - 0.70 acceptable agreement (Williamson et al, 2012)

Results (LLMs vs MLs)



		Score (1-7)		Opinion (Aff, Neg, Other)		
		RMSE	R^2	Acc.	UWK	QWK
ChatGPT (Turbo-3.5-0613)	Base	1.12	0.37	80.13%	0.62	0.40
	Few-shot	1.00	0.23	78.64%	0.60	0.37
	Few-shot + CoT	0.96	0.28	82.52%	0.66	0.46
ChatGPT (Turbo-3.5-0125)	Base	0.88	0.39	82.41%	0.65	0.39
	Few-shot	1.03	0.17	84.02%	0.68	0.46
	Few-shot + CoT	1.04	0.16	87.48%	0.73	0.54
ChatGPT (Turbo-3.5-0613, with FT)	Base	0.94	0.61	85.45%	0.70	0.53
	Few-shot	0.92	0.27	82.08%	0.64	0.46
	Few-shot + CoT	0.92	0.27	66.80%	0.45	0.22
ChatGPT (Turbo-3.5-0125, with FT)	Base	1.15	0.39	87.76%	0.73	0.56
	Few-shot	1.06	0.04	79.63%	0.60	0.40
	Few-shot + CoT	1.04	0.08	84.46%	0.68	0.50
Tree-Based Machine Learning (Text features)	RF	0.61	0.71	79.66%	0.44	0.35
	RRF	0.61	0.71	79.42%	0.44	0.35
	GBM	0.60	0.72	81.09%	0.51	0.43
	XGBOOST	0.61	0.71	88.43%	0.57	0.62

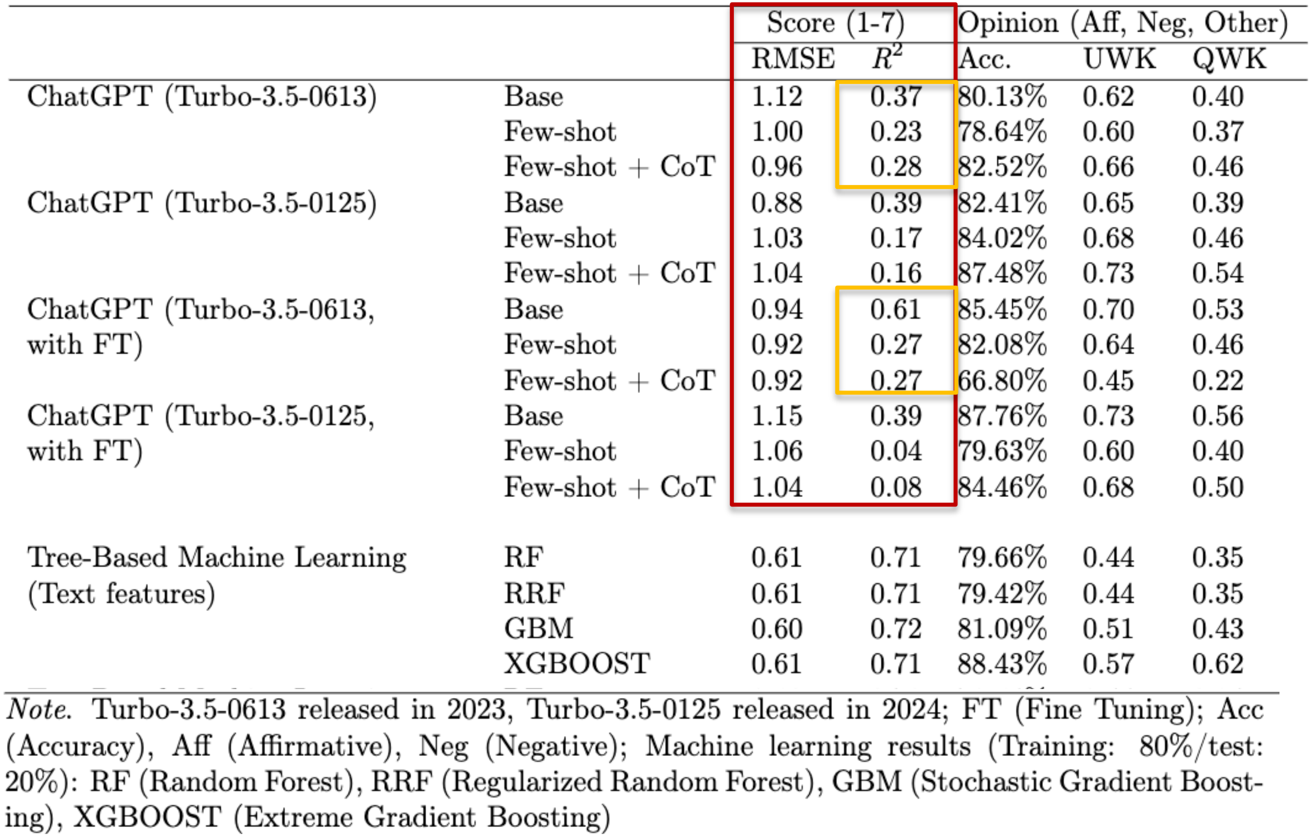
Note. Turbo-3.5-0613 released in 2023, Turbo-3.5-0125 released in 2024; FT (Fine Tuning); Acc (Accuracy), Aff (Affirmative), Neg (Negative); Machine learning results (Training: 80%/test: 20%): RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

Quality of Essay

<Prompting Approaches>

- Base prompts > few-shot approaches
- Few-shot learning approaches (mixed results)

Results (LLMs vs MLs)



		Score (1-7)		Opinion (Aff, Neg, Other)		
		RMSE	R^2	Acc.	UWK	QWK
ChatGPT (Turbo-3.5-0613)	Base	1.12	0.37	80.13%	0.62	0.40
	Few-shot	1.00	0.23	78.64%	0.60	0.37
	Few-shot + CoT	0.96	0.28	82.52%	0.66	0.46
ChatGPT (Turbo-3.5-0125)	Base	0.88	0.39	82.41%	0.65	0.39
	Few-shot	1.03	0.17	84.02%	0.68	0.46
	Few-shot + CoT	1.04	0.16	87.48%	0.73	0.54
ChatGPT (Turbo-3.5-0613, with FT)	Base	0.94	0.61	85.45%	0.70	0.53
	Few-shot	0.92	0.27	82.08%	0.64	0.46
	Few-shot + CoT	0.92	0.27	66.80%	0.45	0.22
ChatGPT (Turbo-3.5-0125, with FT)	Base	1.15	0.39	87.76%	0.73	0.56
	Few-shot	1.06	0.04	79.63%	0.60	0.40
	Few-shot + CoT	1.04	0.08	84.46%	0.68	0.50
Tree-Based Machine Learning (Text features)	RF	0.61	0.71	79.66%	0.44	0.35
	RRF	0.61	0.71	79.42%	0.44	0.35
	GBM	0.60	0.72	81.09%	0.51	0.43
	XGBOOST	0.61	0.71	88.43%	0.57	0.62

Note. Turbo-3.5-0613 released in 2023, Turbo-3.5-0125 released in 2024; FT (Fine Tuning); Acc (Accuracy), Aff (Affirmative), Neg (Negative); Machine learning results (Training: 80%/test: 20%): RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

Quality of Essay

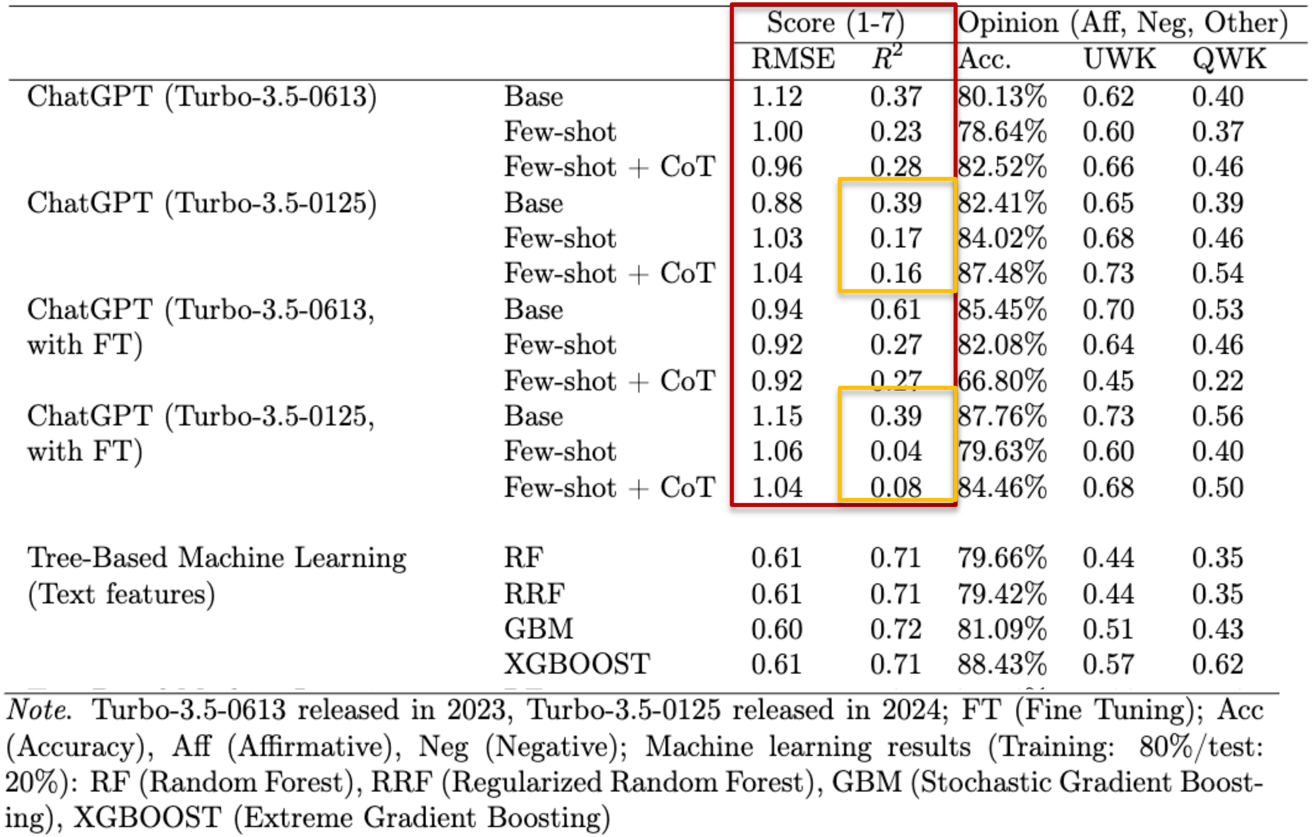
<Prompting Approaches>

- Base prompts > few-shot approaches
- Few-shot learning approaches (mixed results)

<Model Versions>

- Older model often perform better than new model

Results (LLMs vs MLs)



		Score (1-7)		Opinion (Aff, Neg, Other)		
		RMSE	R^2	Acc.	UWK	QWK
ChatGPT (Turbo-3.5-0613)	Base	1.12	0.37	80.13%	0.62	0.40
	Few-shot	1.00	0.23	78.64%	0.60	0.37
	Few-shot + CoT	0.96	0.28	82.52%	0.66	0.46
ChatGPT (Turbo-3.5-0125)	Base	0.88	0.39	82.41%	0.65	0.39
	Few-shot	1.03	0.17	84.02%	0.68	0.46
	Few-shot + CoT	1.04	0.16	87.48%	0.73	0.54
ChatGPT (Turbo-3.5-0613, with FT)	Base	0.94	0.61	85.45%	0.70	0.53
	Few-shot	0.92	0.27	82.08%	0.64	0.46
	Few-shot + CoT	0.92	0.27	66.80%	0.45	0.22
ChatGPT (Turbo-3.5-0125, with FT)	Base	1.15	0.39	87.76%	0.73	0.56
	Few-shot	1.06	0.04	79.63%	0.60	0.40
	Few-shot + CoT	1.04	0.08	84.46%	0.68	0.50
Tree-Based Machine Learning (Text features)	RF	0.61	0.71	79.66%	0.44	0.35
	RRF	0.61	0.71	79.42%	0.44	0.35
	GBM	0.60	0.72	81.09%	0.51	0.43
	XGBOOST	0.61	0.71	88.43%	0.57	0.62

Note. Turbo-3.5-0613 released in 2023, Turbo-3.5-0125 released in 2024; FT (Fine Tuning); Acc (Accuracy), Aff (Affirmative), Neg (Negative); Machine learning results (Training: 80%/test: 20%): RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

Quality of Essay

<Prompting Approaches>

- Base prompts > few-shot approaches
- Few-shot learning approaches (mixed results)

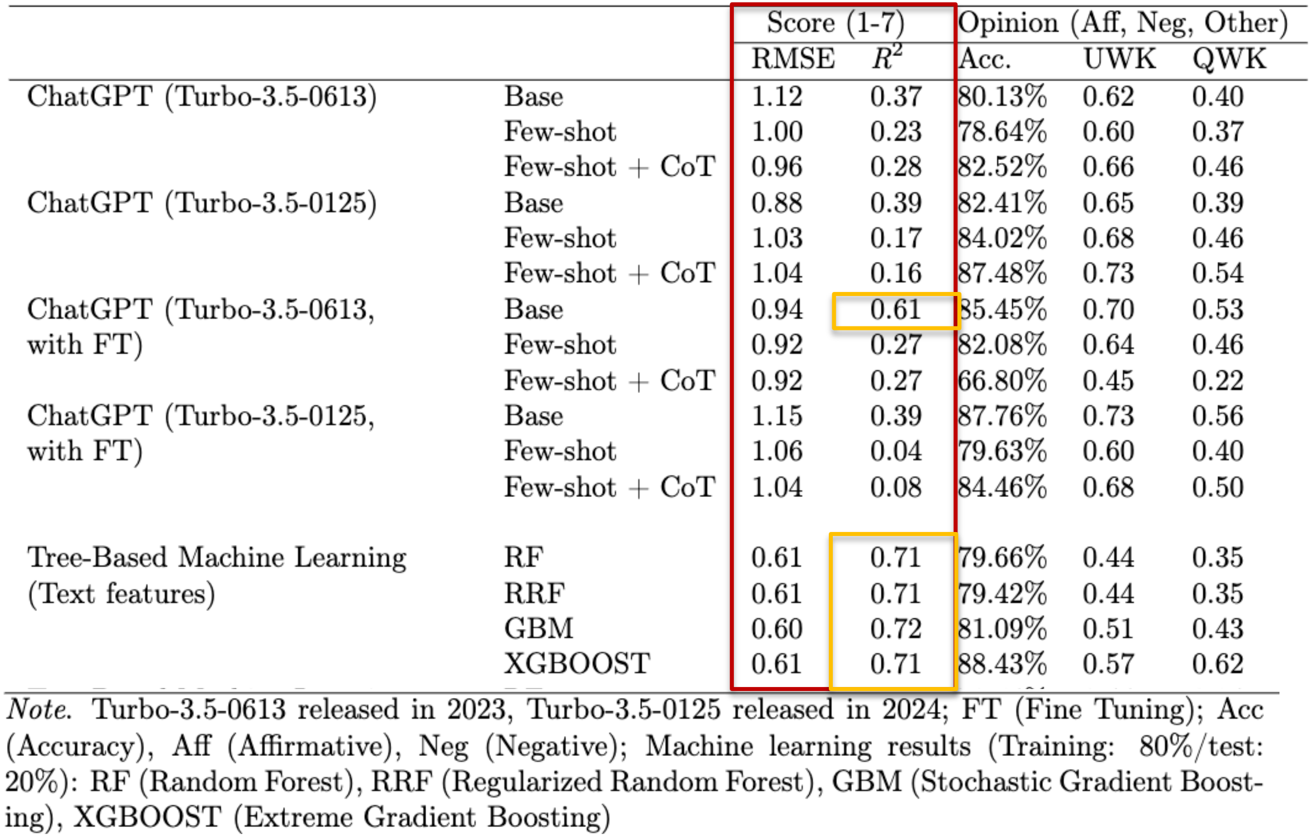
<Model Versions>

- Older model often perform better than new model

<Fine Tuning Impact>

- Older models often perform better than new model

Results (LLMs vs MLs)



		Score (1-7)		Opinion (Aff, Neg, Other)		
		RMSE	R^2	Acc.	UWK	QWK
ChatGPT (Turbo-3.5-0613)	Base	1.12	0.37	80.13%	0.62	0.40
	Few-shot	1.00	0.23	78.64%	0.60	0.37
	Few-shot + CoT	0.96	0.28	82.52%	0.66	0.46
ChatGPT (Turbo-3.5-0125)	Base	0.88	0.39	82.41%	0.65	0.39
	Few-shot	1.03	0.17	84.02%	0.68	0.46
	Few-shot + CoT	1.04	0.16	87.48%	0.73	0.54
ChatGPT (Turbo-3.5-0613, with FT)	Base	0.94	0.61	85.45%	0.70	0.53
	Few-shot	0.92	0.27	82.08%	0.64	0.46
	Few-shot + CoT	0.92	0.27	66.80%	0.45	0.22
ChatGPT (Turbo-3.5-0125, with FT)	Base	1.15	0.39	87.76%	0.73	0.56
	Few-shot	1.06	0.04	79.63%	0.60	0.40
	Few-shot + CoT	1.04	0.08	84.46%	0.68	0.50
Tree-Based Machine Learning (Text features)	RF	0.61	0.71	79.66%	0.44	0.35
	RRF	0.61	0.71	79.42%	0.44	0.35
	GBM	0.60	0.72	81.09%	0.51	0.43
	XGBOOST	0.61	0.71	88.43%	0.57	0.62

Note. Turbo-3.5-0613 released in 2023, Turbo-3.5-0125 released in 2024; FT (Fine Tuning); Acc (Accuracy), Aff (Affirmative), Neg (Negative); Machine learning results (Training: 80%/test: 20%): RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

Quality of Essay

<Prompting Approaches>

- Base prompts > few-shot approaches
- Few-shot learning approaches (mixed results)

<Model Versions>

- Older model often perform better than new model

< Fine Tuning Impact >

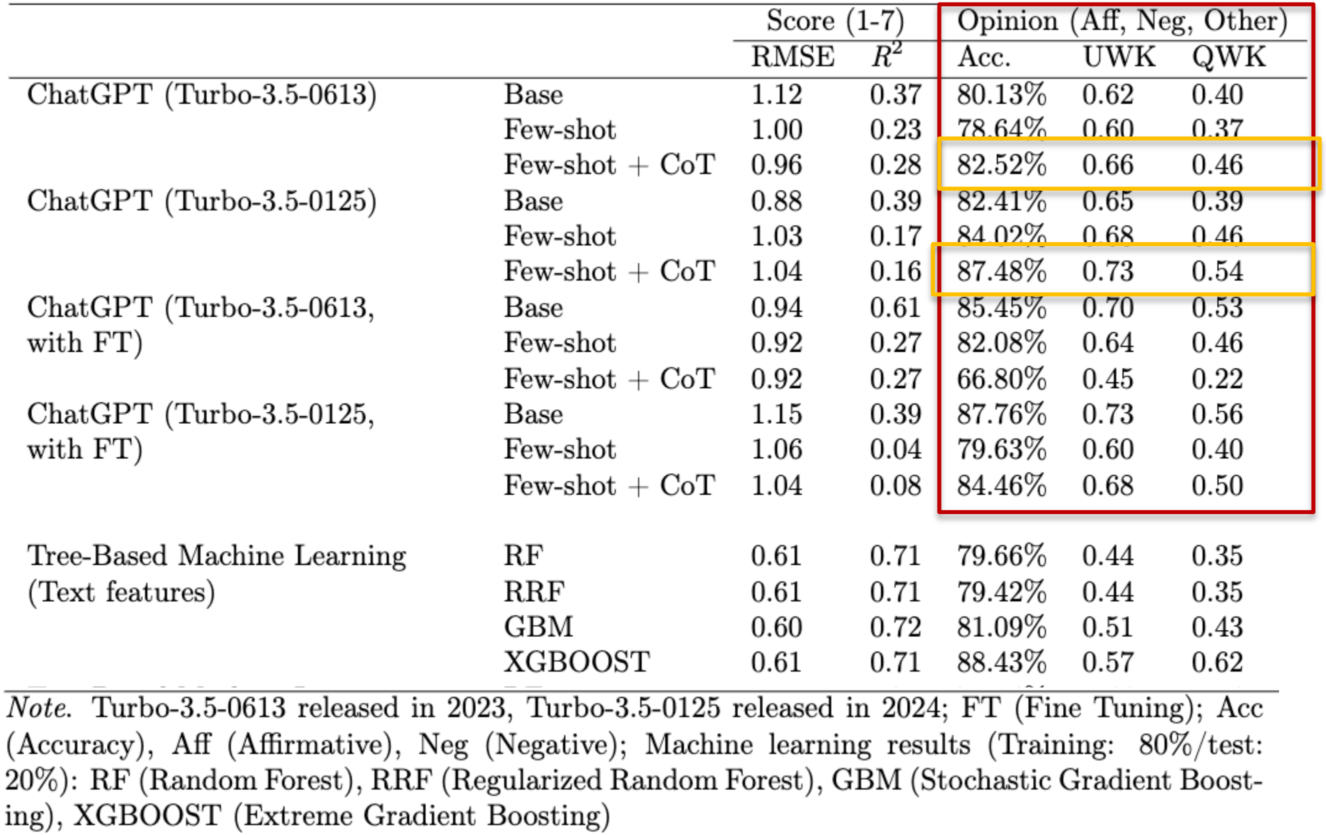
- Older model often perform better than new model

< ChatGPT vs ML>

- Tree-based ML methods are always better than GPTs

Results (LLMs vs MLs)

UWK - 0.41–0.60 moderate agreement
0.61–0.80 substantial agreement
QWK - 0.70 acceptable agreement



		Score (1-7)		Opinion (Aff, Neg, Other)		
		RMSE	R^2	Acc.	UWK	QWK
ChatGPT (Turbo-3.5-0613)	Base	1.12	0.37	80.13%	0.62	0.40
	Few-shot	1.00	0.23	78.64%	0.60	0.37
	Few-shot + CoT	0.96	0.28	82.52%	0.66	0.46
ChatGPT (Turbo-3.5-0125)	Base	0.88	0.39	82.41%	0.65	0.39
	Few-shot	1.03	0.17	84.02%	0.68	0.46
	Few-shot + CoT	1.04	0.16	87.48%	0.73	0.54
ChatGPT (Turbo-3.5-0613, with FT)	Base	0.94	0.61	85.45%	0.70	0.53
	Few-shot	0.92	0.27	82.08%	0.64	0.46
	Few-shot + CoT	0.92	0.27	66.80%	0.45	0.22
ChatGPT (Turbo-3.5-0125, with FT)	Base	1.15	0.39	87.76%	0.73	0.56
	Few-shot	1.06	0.04	79.63%	0.60	0.40
	Few-shot + CoT	1.04	0.08	84.46%	0.68	0.50
Tree-Based Machine Learning (Text features)	RF	0.61	0.71	79.66%	0.44	0.35
	RRF	0.61	0.71	79.42%	0.44	0.35
	GBM	0.60	0.72	81.09%	0.51	0.43
	XGBOOST	0.61	0.71	88.43%	0.57	0.62

Note. Turbo-3.5-0613 released in 2023, Turbo-3.5-0125 released in 2024; FT (Fine Tuning); Acc (Accuracy), Aff (Affirmative), Neg (Negative); Machine learning results (Training: 80%/test: 20%): RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

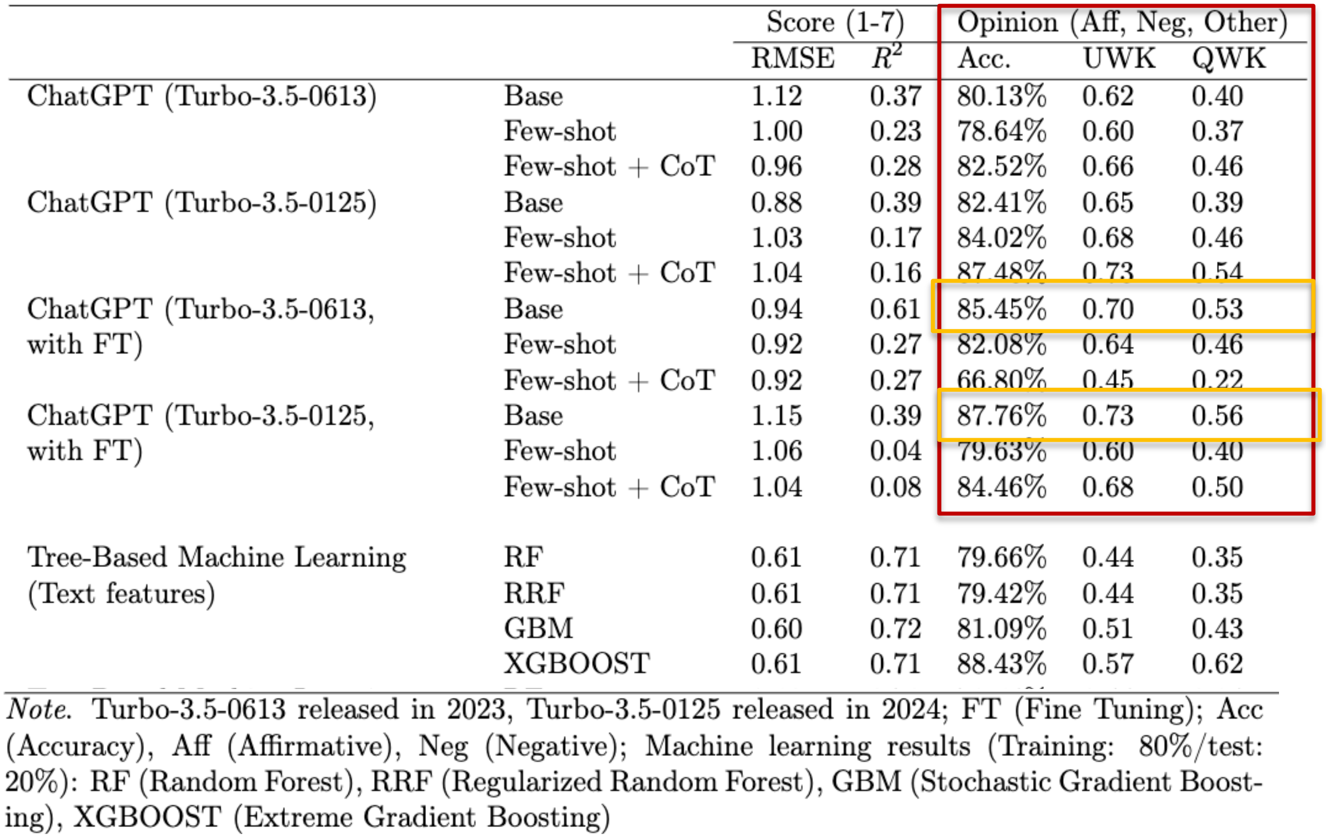
Essay Classification

<Prompting Approaches + Fine Tuning Impact >

- No FT: Few-shot+CoT

Results (LLMs vs MLs)

UWK - 0.41–0.60 moderate agreement
0.61–0.80 substantial agreement
QWK - 0.70 acceptable agreement



		Score (1-7)		Opinion (Aff, Neg, Other)		
		RMSE	R^2	Acc.	UWK	QWK
ChatGPT (Turbo-3.5-0613)	Base	1.12	0.37	80.13%	0.62	0.40
	Few-shot	1.00	0.23	78.64%	0.60	0.37
	Few-shot + CoT	0.96	0.28	82.52%	0.66	0.46
ChatGPT (Turbo-3.5-0125)	Base	0.88	0.39	82.41%	0.65	0.39
	Few-shot	1.03	0.17	84.02%	0.68	0.46
	Few-shot + CoT	1.04	0.16	87.48%	0.73	0.54
ChatGPT (Turbo-3.5-0613, with FT)	Base	0.94	0.61	85.45%	0.70	0.53
	Few-shot	0.92	0.27	82.08%	0.64	0.46
	Few-shot + CoT	0.92	0.27	66.80%	0.45	0.22
ChatGPT (Turbo-3.5-0125, with FT)	Base	1.15	0.39	87.76%	0.73	0.56
	Few-shot	1.06	0.04	79.63%	0.60	0.40
	Few-shot + CoT	1.04	0.08	84.46%	0.68	0.50
Tree-Based Machine Learning (Text features)	RF	0.61	0.71	79.66%	0.44	0.35
	RRF	0.61	0.71	79.42%	0.44	0.35
	GBM	0.60	0.72	81.09%	0.51	0.43
	XGBOOST	0.61	0.71	88.43%	0.57	0.62

Note. Turbo-3.5-0613 released in 2023, Turbo-3.5-0125 released in 2024; FT (Fine Tuning); Acc (Accuracy), Aff (Affirmative), Neg (Negative); Machine learning results (Training: 80%/test: 20%): RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

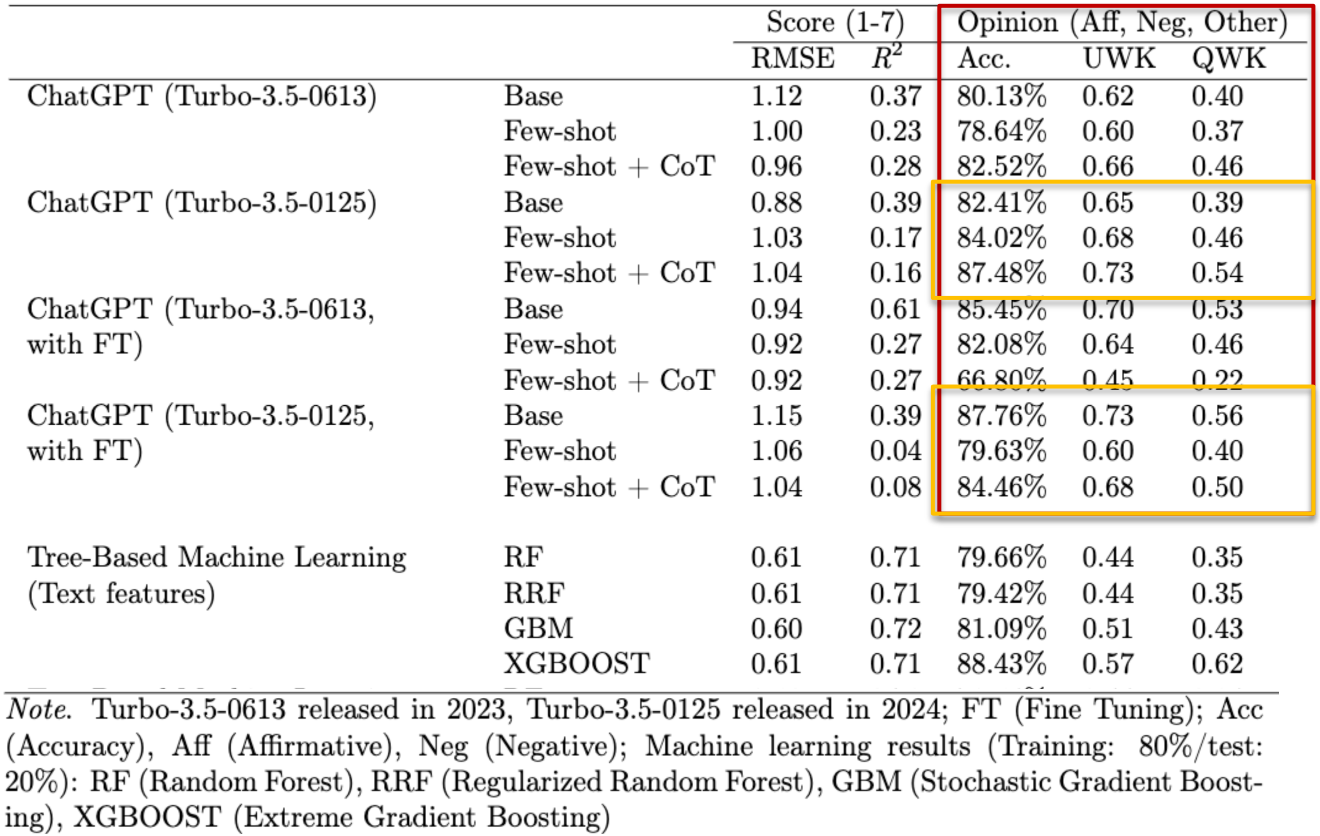
Essay Classification

<Prompting Approaches + Fine Tuning Impact >

- No FT: Few-shot+CoT
- FT: Base prompts

Results (LLMs vs MLs)

UWK - 0.41–0.60 moderate agreement
0.61–0.80 substantial agreement
QWK - 0.70 acceptable agreement



		Score (1-7)		Opinion (Aff, Neg, Other)		
		RMSE	R^2	Acc.	UWK	QWK
ChatGPT (Turbo-3.5-0613)	Base	1.12	0.37	80.13%	0.62	0.40
	Few-shot	1.00	0.23	78.64%	0.60	0.37
	Few-shot + CoT	0.96	0.28	82.52%	0.66	0.46
ChatGPT (Turbo-3.5-0125)	Base	0.88	0.39	82.41%	0.65	0.39
	Few-shot	1.03	0.17	84.02%	0.68	0.46
	Few-shot + CoT	1.04	0.16	87.48%	0.73	0.54
ChatGPT (Turbo-3.5-0613, with FT)	Base	0.94	0.61	85.45%	0.70	0.53
	Few-shot	0.92	0.27	82.08%	0.64	0.46
	Few-shot + CoT	0.92	0.27	66.80%	0.45	0.22
ChatGPT (Turbo-3.5-0125, with FT)	Base	1.15	0.39	87.76%	0.73	0.56
	Few-shot	1.06	0.04	79.63%	0.60	0.40
	Few-shot + CoT	1.04	0.08	84.46%	0.68	0.50
Tree-Based Machine Learning (Text features)	RF	0.61	0.71	79.66%	0.44	0.35
	RRF	0.61	0.71	79.42%	0.44	0.35
	GBM	0.60	0.72	81.09%	0.51	0.43
	XGBOOST	0.61	0.71	88.43%	0.57	0.62

Note. Turbo-3.5-0613 released in 2023, Turbo-3.5-0125 released in 2024; FT (Fine Tuning); Acc (Accuracy), Aff (Affirmative), Neg (Negative); Machine learning results (Training: 80%/test: 20%): RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

Essay Classification

<Prompting Approaches + Fine Tuning Impact >

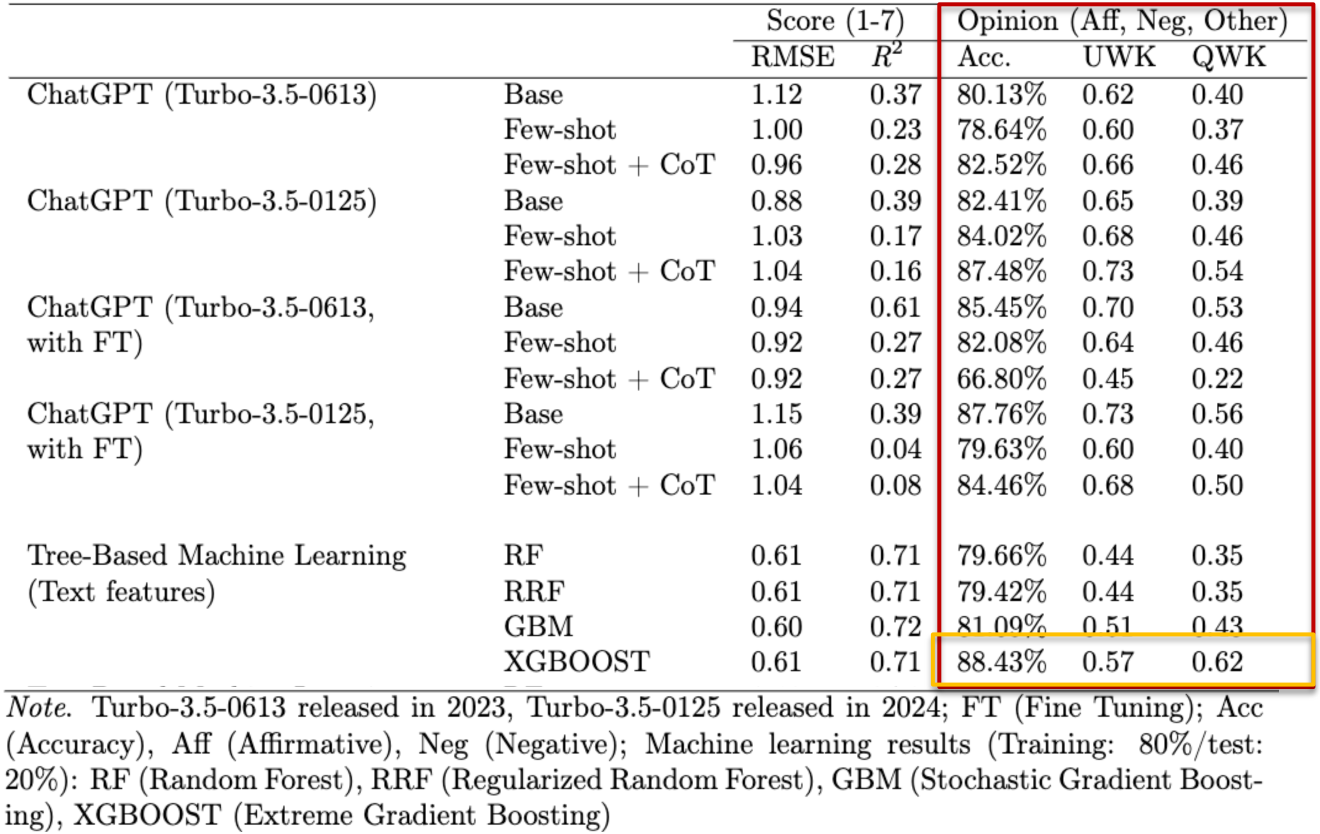
- No FT: Few-shot+CoT
- FT: Base prompts

<Model Versions>

- Newer ChatGPT > older ChatGPT

Results (LLMs vs MLs)

UWK - 0.41–0.60 moderate agreement
0.61–0.80 substantial agreement
QWK - 0.70 acceptable agreement



		Score (1-7)		Opinion (Aff, Neg, Other)		
		RMSE	R^2	Acc.	UWK	QWK
ChatGPT (Turbo-3.5-0613)	Base	1.12	0.37	80.13%	0.62	0.40
	Few-shot	1.00	0.23	78.64%	0.60	0.37
	Few-shot + CoT	0.96	0.28	82.52%	0.66	0.46
ChatGPT (Turbo-3.5-0125)	Base	0.88	0.39	82.41%	0.65	0.39
	Few-shot	1.03	0.17	84.02%	0.68	0.46
	Few-shot + CoT	1.04	0.16	87.48%	0.73	0.54
ChatGPT (Turbo-3.5-0613, with FT)	Base	0.94	0.61	85.45%	0.70	0.53
	Few-shot	0.92	0.27	82.08%	0.64	0.46
	Few-shot + CoT	0.92	0.27	66.80%	0.45	0.22
ChatGPT (Turbo-3.5-0125, with FT)	Base	1.15	0.39	87.76%	0.73	0.56
	Few-shot	1.06	0.04	79.63%	0.60	0.40
	Few-shot + CoT	1.04	0.08	84.46%	0.68	0.50
Tree-Based Machine Learning (Text features)	RF	0.61	0.71	79.66%	0.44	0.35
	RRF	0.61	0.71	79.42%	0.44	0.35
	GBM	0.60	0.72	81.09%	0.51	0.43
	XGBOOST	0.61	0.71	88.43%	0.57	0.62

Note. Turbo-3.5-0613 released in 2023, Turbo-3.5-0125 released in 2024; FT (Fine Tuning); Acc (Accuracy), Aff (Affirmative), Neg (Negative); Machine learning results (Training: 80%/test: 20%): RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

Essay Classification

<Prompting Approaches + Fine Tuning Impact >

- No FT: Few-shot+CoT
- FT: Base prompts

<Model Versions>

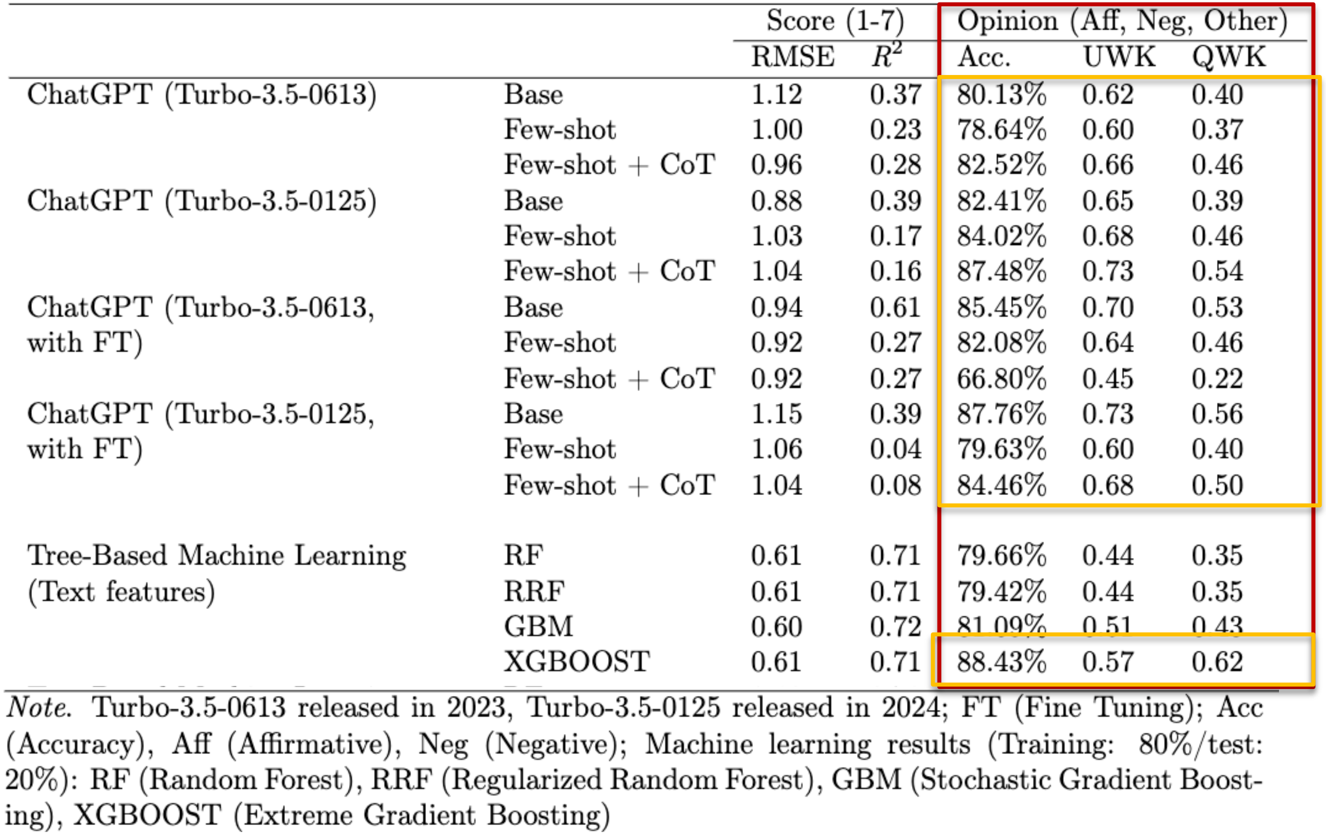
- Newer ChatGPT > older ChatGPT

< ChatGPT vs ML>

- XGBoost was the strongest tree-based model.

Results (LLMs vs MLs)

UWK - 0.41–0.60 moderate agreement
0.61–0.80 substantial agreement
QWK - 0.70 acceptable agreement



		Score (1-7)		Opinion (Aff, Neg, Other)		
		RMSE	R^2	Acc.	UWK	QWK
ChatGPT (Turbo-3.5-0613)	Base	1.12	0.37	80.13%	0.62	0.40
	Few-shot	1.00	0.23	78.64%	0.60	0.37
	Few-shot + CoT	0.96	0.28	82.52%	0.66	0.46
ChatGPT (Turbo-3.5-0125)	Base	0.88	0.39	82.41%	0.65	0.39
	Few-shot	1.03	0.17	84.02%	0.68	0.46
	Few-shot + CoT	1.04	0.16	87.48%	0.73	0.54
ChatGPT (Turbo-3.5-0613, with FT)	Base	0.94	0.61	85.45%	0.70	0.53
	Few-shot	0.92	0.27	82.08%	0.64	0.46
	Few-shot + CoT	0.92	0.27	66.80%	0.45	0.22
ChatGPT (Turbo-3.5-0125, with FT)	Base	1.15	0.39	87.76%	0.73	0.56
	Few-shot	1.06	0.04	79.63%	0.60	0.40
	Few-shot + CoT	1.04	0.08	84.46%	0.68	0.50
Tree-Based Machine Learning (Text features)	RF	0.61	0.71	79.66%	0.44	0.35
	RRF	0.61	0.71	79.42%	0.44	0.35
	GBM	0.60	0.72	81.09%	0.51	0.43
	XGBOOST	0.61	0.71	88.43%	0.57	0.62

Note. Turbo-3.5-0613 released in 2023, Turbo-3.5-0125 released in 2024; FT (Fine Tuning); Acc (Accuracy), Aff (Affirmative), Neg (Negative); Machine learning results (Training: 80%/test: 20%): RF (Random Forest), RRF (Regularized Random Forest), GBM (Stochastic Gradient Boosting), XGBOOST (Extreme Gradient Boosting)

Essay Classification

<Prompting Approaches + Fine Tuning Impact >

- No FT: Few-shot+CoT
- FT: Base prompts

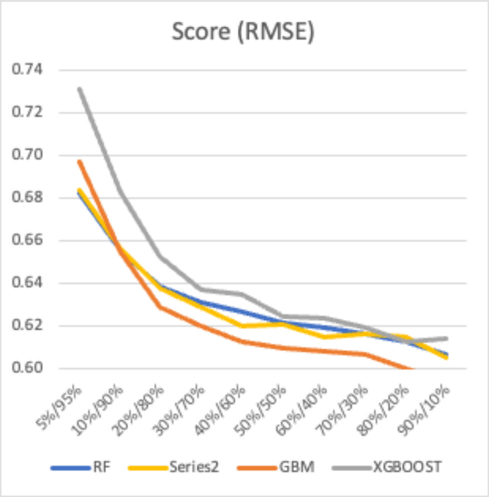
<Model Versions>

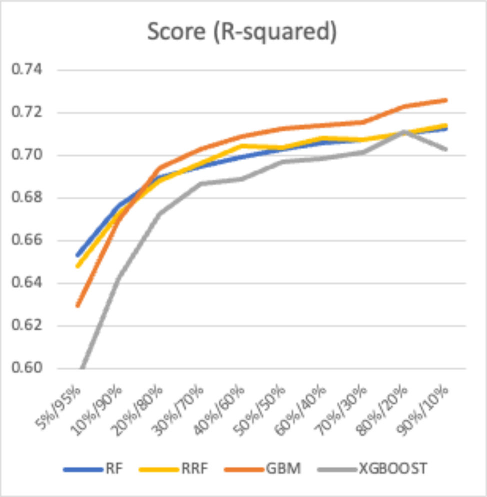
- Newer ChatGPT > older ChatGPT

< ChatGPT vs ML>

- XGBoost was the strongest tree-based model.
- Except XGBOOST, GPTs are generally outperform ML methods

Results (MLs – Quality of Essay)

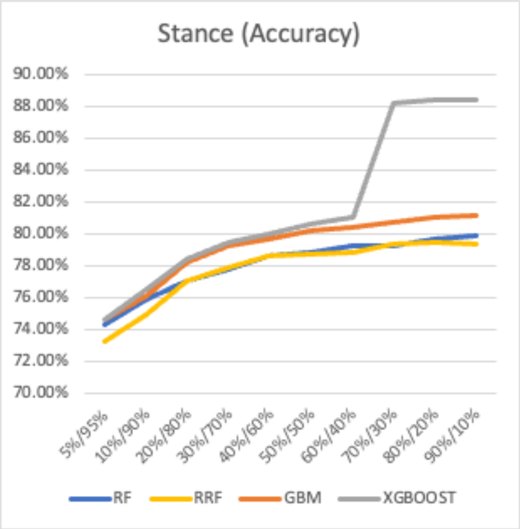


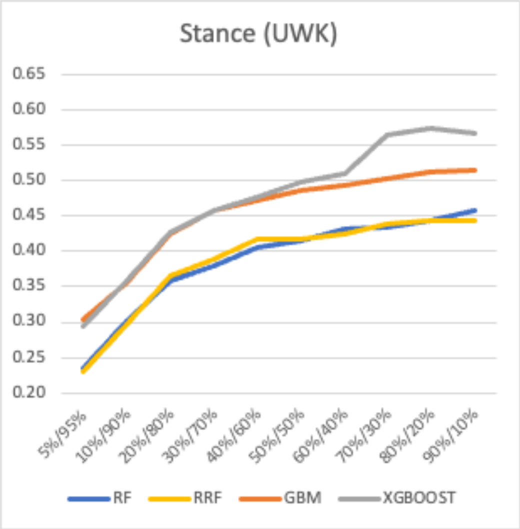


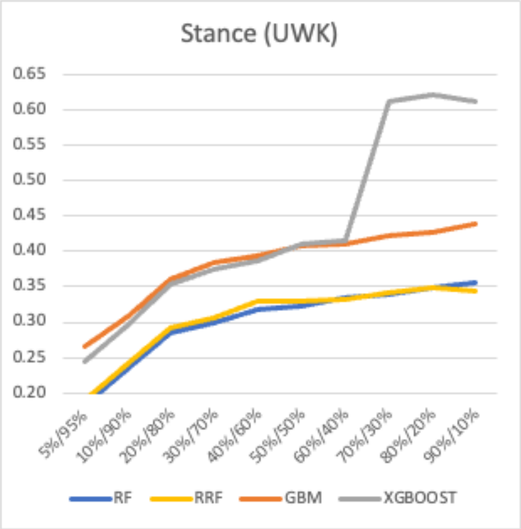
- RF and RRF: Similar performance / better performances before 20%/80% than GBM and XGBOOST
- GBM: Overall best performance
- XGBOOST: Worst performance

Results (MLs – Essay Classification)

UWK - 0.41–0.60 moderate agreement
0.61–0.80 substantial agreement
QWK - 0.70 acceptable agreement







- GBM and XGBOOST: Similar performance until 60%/40%
- XGBOOST: Best performance after 70%/30%
- RF and RRF: Similar performance

Discussion

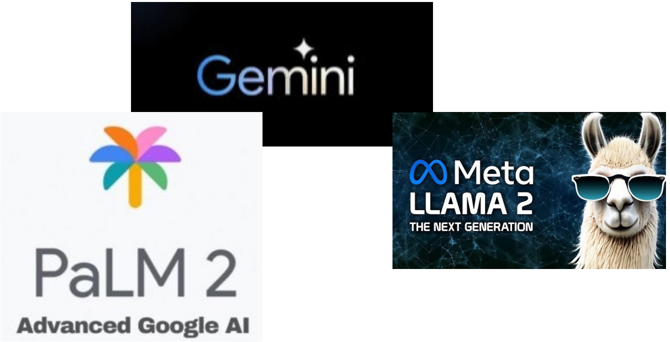
- LLMs show some promise in accurate essay grading, leveraging their language understanding and reasoning
- While excelling in categorizing essays, LLMs encounter challenges in scoring continuous outcomes, even with fine-tuning
- Mixed results in prompting approaches
- Fine tuning and updates matters

Limitations & Future Studies

- Data and model dependence
- Expanding model comparisons

Implications

- Prompting techniques and fine-tuning
- Web interface vs. ChatGPT API
- Fully replacing human grading is still a distant goal

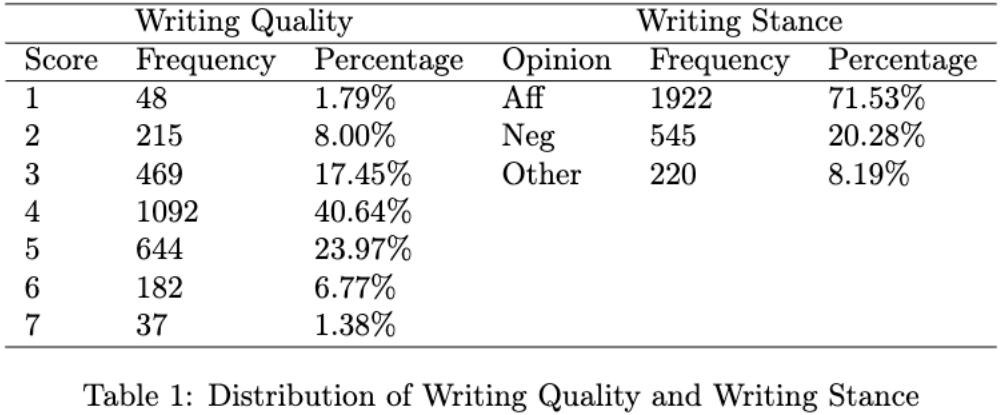


Questions

Thank you for your attention!

Special thanks to my supervisors Drs. Miratrix, Mozer, and Al-adimi.

For further inquiries, please reach out to youngwon_kim@gse.harvard.edu.



	Writing Quality		Writing Stance		
Score	Frequency	Percentage	Opinion	Frequency	Percentage
1	48	1.79%	Aff	1922	71.53%
2	215	8.00%	Neg	545	20.28%
3	469	17.45%	Other	220	8.19%
4	1092	40.64%			
5	644	23.97%			
6	182	6.77%			
7	37	1.38%			

Table 1: Distribution of Writing Quality and Writing Stance