From Text to Trust: Empowering AI-assisted Decision Making with Adaptive LLM-powered Analysis (Supplementary Material)

Zhuoyan Li
li4178@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Hangxiao Zhu
hangxiao@tamu.edu
Texas A&M University
College Station, Texas, USA

Zhuoran Lu
lu800@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Ziang Xiao
ziang.xiao@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Ming Yin
mingyin@purdue.edu
Purdue University
West Lafayette, Indiana, USA

ACM Reference Format:
Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, Ziang Xiao, and Ming Yin. 2025. From Text to Trust: Empowering AI-assisted Decision Making with Adaptive LLM-powered Analysis (Supplementary Material) . In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613904.3642625

A EMPIRICAL EXAMINATIONS OF IMPACTS OF LLM-POWERED ANALYSIS IN AI-ASSISTED DECISION MAKING (ADDITIONAL DETAILS)

The full demographic information and statistics of participants in the study of this phase (where we have three treatments, i.e., Control, Seq, and All) for the income prediction and recidivism prediction tasks are shown in Table A.1 and Table A.2, respectively.

B COMBINING HUMAN DECISIONS AND AI PREDICTIONS

In the main paper, we aim to measure the reliability of AI model recommendations. Following previous research [2, 6], we combined independent human decisions with AI model predictions to determine the targeted decision for each task instance. We evaluated the human-AI combination method [2] and several truth inference methods used in crowdsourcing for truth discovery. We detailed the process of evaluation below.

Simulating Human Independent Decision. To understand how humans independently make decisions on the task instance, we first conducted a pilot study on Prolific to collect independent human decision behavior data on income prediction tasks and recidivism prediction tasks. We recruited 40 participants for each task. Each recruited participant needed to complete 15 tasks. With the collection of human behavior data, we then fitted two-layer neural networks to simulate human independent decision behavior. We optimized these independent behavior models using Adam [3] with an initial learning rate of $1e-4$ and a batch size of each training iteration of 128. The number of training epochs is set as 10. The average accuracy of 5-fold validation for the fitted decision models is 0.81 for the income prediction task and 0.84 for the recidivism task, both of which we found to be satisfactory. We then utilized these fitted models to simulate independent human decisions $y^h_{\text{independent}}$ in the human-AI combination process to determine the potentially better decisions.

Comparing Combination Performance. We consider the human + AI combination method proposed in [2] and a few truth inference methods in crowdsourcing as baselines in the evaluation, including GLAD [7], CATD [4], LFC [5], EM [8], and MV [8]. These methods combine the human independent decisions $y^h_{\text{independent}}$ predicted by the fitted independent human behavior models and AI model recommendations y^m to produce combined decisions y_{combine}. The accuracy of y_{combine} on holdout task pools when using different methods to combine humans' (predicted) independent decisions and AI's decision recommendation is reported in Table B.1. In general, we found that the human + AI combination method proposed in [2] outperforms other baselines. By integrating human decisions with AI predictions, this method shows superior performance to either AI solo or human solo across the two types of decision making tasks. Consequently, we used the combined decisions y_{combine} from the human + AI combination method as the targeted decision in subsequent experiments to select the LLM rationale analysis.

C ALGORITHMIC SELECTION OF LLM-POWERED ANALYSIS IN AI-ASSISTED DECISION MAKING (ADDITIONAL DETAILS)

C.1 Algorithmic Framework Setting

The initial hidden state distribution $\mathcal{P}(\boldsymbol{h}_0|\boldsymbol{x}, y^m; \boldsymbol{\theta}_{\text{init}})$ is modeled as a Gaussian distribution conditioned on the task $\boldsymbol{x}$ and the AI model prediction y^m. Specifically, it is represented as $\mathcal{N}(\boldsymbol{h}^0; \boldsymbol{\mu}_{\boldsymbol{\theta}_{\text{init}}}(\boldsymbol{x}, y^m), \Sigma_{\boldsymbol{\theta}_{\text{init}}}(\boldsymbol{x}, y^m))$, where $\boldsymbol{\mu}_{\boldsymbol{\theta}_{\text{init}}}$ and $\Sigma_{\boldsymbol{\theta}_{\text{init}}}$ are parameterized by one-layer feedforward networks. For state updating, we also use a Gaussian distribution to characterize this process as $\mathcal{N}(\boldsymbol{h}^t; \boldsymbol{\mu}_{\boldsymbol{\theta}_{\text{update}}}(\boldsymbol{r}^t, a^t, \boldsymbol{h}^{t-1}), \Sigma_{\boldsymbol{\theta}_{\text{update}}}(\boldsymbol{r}^t, a^t, \boldsymbol{h}^{t-1}))$. Here, $\boldsymbol{r}^t$ is encoded as three parts: (1) the text embedding of the LLM-powered analysis encoded with a BERT model [1]; (2) the task

CHI '25, April 26–May 1, 2025, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/3613904.3642625

Demographics		Control (N = 41)	Seq (N = 45)	All (N = 48)
Gender	Male	26.9%	53.3%	37.5%
	Female	73.1%	44.4%	60.4%
	Others	0	2.3%	2.1%
Age	Below 35	60.9%	46.7%	54.1%
	35 -44	29.2%	33.3%	16.7%
	45 or above	9.9%	20%	29.2%
Race	White	51.2%	68.8%	58.4%
	Black	7.3%	24.4%	14.5%
	Hispanic	17.1%	2.2%	14.6%
	Others	24.4%	4.4%	12.5%
Education	High school or lower	21.9%	4.5%	4.2%
	Some college	26.8%	28.9%	43.75%
	Bachelor Degree	36.5%	46.6%	31.3%
	Graduate school or higher	14.6%	20%	20.8%
Average Trust with AI systems		2.75	3.26	3.16
Average knowledge level with AI systems		2.53	2.95	2.66

Table A.1: Details of the demographic backgrounds of participants for empirical examinations of the impact of LLM-powered analysis in the income prediction task. N represents the number of participants in that treatment.

Demographics		Control (N = 49)	Seq (N = 40)	All (N = 61)
Gender	Male	38.7%	37.5%	31.1%
	Female	57.1%	62.5%	68.9%
	Others	4.2%	0	0
Age	Below 35	47.1%	37.5%	45.9%
	35 -44	38.7%	27.5%	27.4%
	45 or above	14.2%	35%	26.7%
Race	White	73.4%	67.5%	60.7%
	Black	10.2%	12.5%	14.7%
	Hispanic	4.1%	7.5%	6.6%
	Others	12.2%	12.5%	18.0%
Education	High school or lower	16.3%	12.5%	22.3%
	Some college	24.4%	32.5%	23.5%
	Bachelor Degree	36.7%	40%	36.1%
	Graduate school or higher	22.5%	15%	18.1%
Average Trust with AI systems		3.06	2.82	2.88
Average knowledge level with AI systems		2.77	2.72	2.78

Table A.2: Details of the demographic backgrounds of participants for the empirical examinations of the impact of LLM-powered analysis in the recidivism prediction task. N represents the number of participants in that treatment.

	Human Solo	AI Solo	Human + AI	GLAD	CATD	LFC	EM	MV
Income Prediction	0.61	0.73	**0.76**	0.62	0.74	0.61	0.60	0.69
Recidivism Prediction	0.54	0.58	0.61	0.59	0.61	0.58	0.59	**0.62**

Table B.1: Comparing the accuracy of different combination methods in the income prediction and recidivism prediction tasks, respectively. The best result in each row is highlighted in bold.

feature that the LLM focuses on analyzing in this analysis (e.g., a person's occupation, a defendant's charge degree); and (3) the LLM's analysis on the polarity of this feature towards the final decision, which can be positive, neutral, or negative. For example, in the recidivism prediction task, a positive polarity indicates that the LLM believes that the feature will increase the probability of reoffending, a neutral polarity indicates that it has no effect, and a negative

polarity indicates that it will decrease the probability of reoffending. $\boldsymbol{\mu}_{\boldsymbol{\theta}_{\text{update}}}$ and $\Sigma_{\boldsymbol{\theta}_{\text{update}}}$ are parameterized by one-layer LSTM-like feedforward networks. The $\boldsymbol{h}^{t-1}$ is iteratively computed starting from the initial values: $\boldsymbol{h}^0 \sim \mathcal{N}(\boldsymbol{h}^0; \boldsymbol{\mu}_{\boldsymbol{\theta}_{\text{init}}}(\boldsymbol{x}, y^m), \Sigma_{\boldsymbol{\theta}_{\text{init}}}(\boldsymbol{x}, y^m))$. The dimension of the hidden state $\boldsymbol{h}^t$ is set as 128. For the final decision making, we employ a two-layer multilayer perceptron to model the decision based on the last hidden state.

C.2 Demographic Information of Participants

The full demographic backgrounds of participants in the study of this phase (where we have five treatments, i.e., Control, Seq, All, Alg, and Rank) for the income prediction and recidivism prediction tasks are shown in Table C.1 and Table C.2, respectively.

D LLM-POWERED ANALYSIS EXAMPLES

D.1 Income Prediction Task

Table D.1 provides additional examples of the GPT-4's analysis for the income prediction task.

D.2 Recidivism Prediction Task

Table D.2 provides additional examples of the GPT-4's analysis for the recidivism prediction task.

Demographics		Control (N = 66)	Human-Solo (N = 52)	Seq (N = 85)	All (N = 91)	Rank (N = 53)	Alg (N = 100)
Gender	Male	40.9%	36.5%	41.2%	48.4%	51.8%	52.0%
	Female	57.6%	59.6%	55.3%	51.6%	46.3%	47.0%
	Others	1.5%	3.9%	3.5%	0	1.9%	1.0%
Age	Below 35	49.9%	48.1%	46.9%	30.8%	51.4%	47.0%
	35 -44	31.8%	25.0%	27.1%	33.0%	24.1%	26.0%
	45 or above	18.3%	26.9%	26.0%	36.2%	24.5%	27.0%
Race	White	69.7%	57.6%	61.2%	57.1%	57.4%	67.0%
	Black	10.6%	25%	20%	22.0%	20.3%	15.0%
	Hispanic	9.1%	13.5%	2.4%	9.9%	11.2%	6.0%
	Others	10.6%	3.9%	16.4%	11.0%	11.1%	12.0%
Education	High school or lower	6.1%	3.8%	4.7%	8.8%	9.3%	20.0%
	Some college	30.3%	32.7%	29.4%	28.6%	24.1%	29.0%
	Bachelor Degree	45.5%	30.7%	47.1%	42.9%	51.8%	37.0%
	Graduate school or higher	18.1%	32.8	18.8%	19.7%	14.8%	14.0%
Average Trust with AI systems		2.86	3.55	3.29	3.32	3.29	3.09
Average knowledge level with AI systems		2.43	3.38	2.85	3.03	3.01	2.71

Table C.1: Details of the demographic backgrounds of participants for the evaluation of the effectiveness of the proposed algorithmic selection of LLM-powered analysis in the income prediction task. N represents the number of participants in that treatment.

Demographics		Control (N = 84)	Human-Solo (N = 49)	Seq (N = 59)	All (N = 68)	Rank (N = 49)	Alg (N = 88)
Gender	Male	41.6%	43.1%	47.3%	38.2%	15.2%	38.6%
	Female	57.1%	56.9%	50.9%	58.9%	82.6%	60.2%
	Others	1.3%	0%	1.8%	2.9%	2.2%	1.2%
Age	Below 35	41.6%	35.2%	43.8%	44.1%	47.7%	61.3%
	35 -44	34.5%	39.2%	22.80%	33.8%	19.5%	20.4%
	45 or above	23.9%	25.6%	33.4%	22.1%	32.8%	18.3%
Race	White	60.7%	56.8%	71.9%	60.3%	63.1%	62.5%
	Black	16.7%	29.4%	15.8%	14.7%	26.1%	14.7%
	Hispanic	8.3%	7.8%	0	4.4%	6.5%	9.1%
	Others	14.3%	6.0%	12.3%	20.6%	4.3%	13.7%
Education	High school or lower	17.8%	3.9%	10.5%	16.1%	17.4%	5.7%
	Some college	28.6%	29.4%	19.2%	29.4%	34.7%	34.1%
	Bachelor Degree	31.0%	43.1%	49.1%	33.8%	34.8%	46.6%
	Graduate school or higher	22.6%	23.5%	21.2%	20.7%	13.1%	13.6%
Average Trust with AI systems		3.15	3.35	3.00	3.14	2.73	3.03
Average knowledge level with AI systems		2.90	3.47	2.75	2.94	2.63	2.95

Table C.2: Details of the demographic backgrounds of participants for the evaluation of the effectiveness of the proposed algorithmic selection of LLM-powered analysis in the recidivism prediction task. N represents the number of participants in that treatment.

Income Prediction	
Gender	*Value:* Female *Concept:* Gender can influence earning potential due to societal and systemic inequalities. *In this case:* As a female, there might be gender pay gaps present, which **decreases** the likelihood of making over $50000 per year.
Age	*Value:* 56 *Concept:* Middle-aged individuals often have more work experience, which can lead to higher income. *In this case:* At 56, the person likely has significant experience, which **increases** the likelihood of making over $50000 per year.
Education Level	*Value:* 10 years *Concept:* Higher education is commonly linked to higher earning potential. *In this case:* With 10 years of education, this might be slightly below the threshold for high-earning positions, which **decreases** the likelihood of making over $50000 per year.
Marital Status	*Value:* Divorced *Concept:* Married individuals might have dual incomes or more stability, affecting income levels. *In this case:* Being divorced may lead to financial strain and does not provide the potential dual income advantage of married couples, which **decreases** the likelihood of making over $50000 per year.
Occupation	*Value:* Sales *Concept:* Certain professions have higher average incomes than others. *In this case:* Sales positions vary greatly in income, but without further specifics, it's challenging to determine, which **has no impact** on the likelihood of making over $50000 per year.
Work Type	*Value:* Self-employed *Concept:* The class of work can affect income level, with self-employed individuals having a high variance in earnings. *In this case:* Being self-employed in sales might offer high earning potential if the business is successful, which **has no impact** on the likelihood of making over $50000 per year due to the variability in success amongst self-employed individuals.
Working Hour Per Week	*Value:* 48 *Concept:* Working more hours can lead to higher income, especially in jobs with overtime pay or higher rates for longer hours. *In this case:* As a female, there might be gender pay gaps present, which **decreases** the likelihood of making over $50000 per year.

Table D.1: Analysis of the Income Prediction task.

Recidivism Prediction	
Gender	*Value:* Female *Concept:* Gender may impact recidivism predictions due to societal perceptions and treatment of different genders within the criminal justice system. *In this case:* Being female, which generally **decreases** the likelihood of recidivating compared to males, reflecting broader trends in criminal behavior and recidivism.
Age	*Value:* 23 *Concept:* Younger individuals are statistically more likely to reoffend, possibly due to lack of stability, economic opportunities, or maturity. *In this case:* At the age of 23, which **increases** the likelihood of recidivating, as younger age is often associated with higher recidivism rates.
Race	*Value:* Black (African American) *Concept:* Societal and systemic biases related to race can influence recidivism predictions, with some races potentially facing harsher predictions due to historical and ongoing discrimination. *In this case:* Being Black (African American), which might **increase** the likelihood of recidivating due to systemic biases that affect judiciary outcomes.
Prior Non-juvenile Crimes	*Value:* 0 *Concept:* The number of prior criminal charges is a strong predictor of recidivism, with more priors indicating a higher risk. *In this case:* With no non-juvenile criminal charges, which **decreases** the likelihood of recidivating, suggesting a lack of previous engagement with the criminal justice system.
Juvenile Misdemeanor Crimes	*Value:* 0 *Concept:* Juvenile misdemeanors indicate lesser criminal involvement than felonies but can still reflect patterns of behavior leading to recidivism. *In this case:* Having no juvenile misdemeanor charges, which **decreases** the likelihood of recidivating, representing a lower early involvement in criminal activities.
Juvenile Felony Crimes	*Value:* 0 *Concept:* Juvenile felony charges are considered indicators of early criminal behavior, which can predict future recidivism. *In this case:* Having no juvenile felony charges, which **decreases** the likelihood of recidivating since early criminal behavior is not present.
Charge Issue	*Value:* Driving While License Revoked *Concept:* The specific nature of the current charge can influence recidivism predictions, with certain offenses considered more likely to lead to reoffending. *In this case:* Charged with Driving While License Revoked, which **has no impact** on the likelihood of recidivating as it may not directly indicate a higher risk of violent or more serious criminal behavior.
Charge Degree	*Value:* Felony *Concept:* The severity of the charge can predict recidivism, with felonies often leading to harsher predictions than misdemeanors. *In this case:* Facing a felony charge, which **increases** the likelihood of recidivating because felonies are associated with more severe criminal behavior.

Table D.2: Analysis of the Recidivism Prediction task.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[2] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. 2021. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems* 34 (2021), 4421–4434.

[3] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[4] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. 2014. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment* 8, 4 (2014), 425–436.

[5] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of machine learning research* 11, 4 (2010).

[6] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.

[7] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems* 22 (2009).

[8] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. 2017. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment* 10, 5 (2017), 541–552.