

Beyond Volume: How Cognitive Fingerprints Predict Residual Gain in AI Tutoring

Aleena K Raj

Learning Analytics & Psychometrics
Kaplan North America
aleena.kraj@kaplan.edu

Nina Deng

Learning Analytics & Psychometrics
Kaplan North America
nina.deng@kaplan.com

Abstract

This study evaluates an AI tutoring tool in a pre-licensure exam preparation product. Using residual gain modeling, elastic-net feature selection, and Gaussian Mixture clustering on LLM-derived cognitive fingerprints, the study identifies meaningful learner profiles and shows that the cognitive depth of student–AI interaction, not volume, drives measurable learning gain.

1 Introduction

Generative AI-based tutoring tools are increasingly integrated into online learning platforms. While conversational interfaces can enhance engagement and personalization, their efficacy remains under-evaluated. Early research on intelligent tutoring systems often fails to isolate students' baseline capabilities from treatment effects (Shadish et al., 2002; Pelánek et al., 2016). Recent studies report highly heterogeneous results (Kestin et al., 2024; Mollick and Mollick, 2024), and some reveal that unguided generative AI use leads to cognitive offloading and degraded critical thinking (Bastani et al., 2025).

One challenge is that many evaluations report aggregate treatment effects, implicitly assuming that all students benefit uniformly from AI tutoring. When the intervention's effect depends fundamentally on *how* students use the tool—not merely *whether* they use it—aggregate metrics will systematically underestimate (or entirely miss) the tool's value for heterogeneous learners.

Motivated by this gap, the present study aims to address the following research objectives in the context of an online exam preparation product:

- Understand the efficacy of an AI tutor chatbot by residual score gain after controlling baseline ability.
- Identify the characteristics of student-AI interaction that are associated with residual gain.

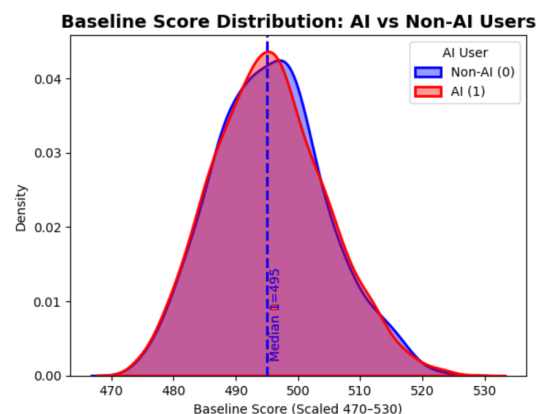


Figure 1: Pre-assessment score distributions for AI users ($n = 667$) and non-AI users ($n = 3,467$). The near-complete overlap indicates minimal selection bias on observed baseline ability.

- Identify learner engagement profiles which exhibit different residual gain.

2 Data

Chat history, course activity, and assessment data were collected from a cohort enrolled in an online pre-licensure exam preparation course between August 2024 and April 2025. The final sample comprises 667 AI tutor users (who interacted with the chatbot at least once) and 3,467 non-AI users (who did not interact with chatbot). Both groups took pre- and post-assessments to calculate scaled score gains. Pre-assessment score distributions were nearly identical across both groups (Figure 1), indicating minimal selection bias between the two groups.

3 Methods

The analysis proceeds through a four-step pipeline. Each step's output serves as input to the next; the sequential architecture is deliberate and has implications for model specification (see §3.1).

3.1 Residual Gain Modeling

Residual gain for AI user i is defined as:

$$r_i = g_i^{(\text{obs})} - \hat{g}(b_i; \hat{\theta}_{\text{non-AI}}) \quad (1)$$

where $g_i^{(\text{obs})}$ is the observed score gain, b_i is the baseline (pre-assessment) score, and $\hat{g}$ is a quadratic ordinary-least-squares model with parameters $\hat{\theta}_{\text{non-AI}}$ estimated exclusively from the non-AI cohort ($n = 3,467$). Functional form was selected by comparing linear and quadratic OLS specifications via a nested F-test. Distributional fairness of the residual was assessed by an ANOVA on residuals across baseline-score tertiles. Reference-model diagnostics are reported in §4.1.

The reference model deliberately conditions on baseline score. The other learning behavior variables (e.g., total questions answered, study intensity) are excluded because they serve as predictors in the downstream elastic-net decomposition (§3.3). Including them in both stages would constitute double adjustment: the residual is *designed* to contain all behavioral signal so that the elastic-net can allocate it. This follows standard practice in value-added modeling, where prior achievement is the primary conditioning variable and post-treatment covariates are excluded to avoid “bad controls” (Angrist and Pischke, 2009; Sanders and Horn, 1998; Raudenbush, 2004).

3.2 Cognitive Fingerprinting

To evaluate interaction quality, each student–AI conversation was classified using Gemini-1.5-Flash (temperature = 0.1) via a structured few-shot prompt, an established approach for text annotation (Gilardi et al., 2023; Ziems et al., 2024). Classifications spanned two dimensions:

- **Content section** (S): five exam topics (A–E).
- **Cognitive depth** (D): five levels adapted from Bloom’s Taxonomy (Bloom, 1956)—Factual Recall, Conceptual Understanding, Application/Problem-Solving, Analysis/Reasoning, and Metacognitive/Strategic.

The Cartesian product $S \times D$ yields a 25-cell grid. Each student’s proportion vector across this grid constitutes their **cognitive fingerprint**: a unique signature of how they engage with the AI across content areas and reasoning levels.

Classifier validation. Face validity was established through manual inspection of student prompt language: interactions classified as Analysis/Reasoning contained reasoning-oriented language (e.g., “why does X fail under Y”), while Factual Recall items contained definition-seeking queries (e.g., “what is X”). Convergent validity was assessed by examining within-student cross-section consistency: students who engaged at higher cognitive levels in one content section tended to do so across sections, a pattern inconsistent with random classification. Furthermore, a pilot inter-rater reliability design is planned in the future work (§6).

3.3 Feature Selection and Composite Indices

Twenty-five candidate predictors spanning four families—total course activity, cognitive depth (proportions across the five Bloom-adapted levels defined in §3.2), content topic, and student-AI interaction volume—were analyzed using elastic-net regression (Zou and Hastie, 2005) with residual gain as the outcome. Baseline score was included as a control. The elastic-net penalty combines an ℓ_1 component (which forces weak predictors to exactly zero, aligned with the feature-selection objective) and an ℓ_2 component (which stabilizes correlated features); the overall strength α and the ℓ_1/ℓ_2 mixing ratio were selected via 5-fold cross-validation, with fitted values reported in §4.2. The full predictor list, organized by family, appears in Appendix B.

Predictors with non-zero weights were retained and partitioned into two orthogonal composite indices: the **Effort Index** (course-activity features) and the **AI-Leverage Index** (AI-interaction features). These indices serve as inputs to the persona analysis in the clustering step.

3.4 Behavioral Clustering

A Gaussian Mixture Model (Reynolds, 2009) was fit on the Effort Index $\times$ AI-Leverage Index space. Baseline score was excluded so that personas reflect *behavior*, not ability. The optimal number of clusters was determined by comparing BIC, AIC, and Silhouette score across $k \in \{3, 4, 5\}$. Cluster stability was assessed via the Adjusted Rand Index (ARI) across 10 random initializations, and per-student assignment certainty was quantified using posterior class probabilities. Silhouette is interpreted as a secondary diagnostic given that it penalizes the soft, overlapping cluster boundaries

characteristic of GMMs and was originally developed for hard-partition methods such as k -means. Cluster diagnostics are reported in §4.3.

4 Results

4.1 Aggregate Effect

Reference-model diagnostics. The quadratic reference model ($R^2 = 0.053$) significantly outperformed a linear specification ($R^2 = 0.041$; nested $F = 44.26$, $p < 0.001$). The fitted equation is $\hat{g}(x) = -0.0066x^2 + 6.41x - 1544.38$; the negative leading coefficient captures the ceiling effect for high-baseline learners. An ANOVA on residuals across baseline tertiles was non-significant ($F = 2.98$, $p = 0.052$), indicating that residual gain does not systematically favor any ability stratum.

Aggregate residual gain. The mean residual gain across all 667 AI users is $+0.50$ points (directionally positive, not statistically significant). The median is exactly 0.00, with $SD = 6.22$ and a range from -10.54 to $+28.59$. This wide range and near-zero average suggest substantial heterogeneity and motivate the persona-level decomposition that follows. It indicate that *aggregate evaluation* is an insufficient paradigm when treatment effects are heterogeneous and likely depend on student-AI interaction quality.

4.2 Feature Selection

Model fit. Cross-validation selected $\alpha = 0.1851$ and ℓ_1 -ratio = 0.95; the ℓ_2 component thus contributes only 5% of the total penalty, consistent with the feature-selection objective.

Selected predictors. Sixteen of 25 predictors maintained non-zero weights. Six features positively predicted residual gain (Figure 2): Total Questions Answered ($+1.65$), Study Efficiency ($+0.89$), Study Intensity ($+0.65$), Full-Length Test Effort ($+0.28$), Analysis/Reasoning% ($+0.23$), and Average Tokens per Interaction ($+0.02$).

The course-activity family yielded the largest weights, including Total Questions Answered, Study Efficiency, and Study Intensity. This confirms that increased practice alongside AI tutoring drives greater residual gains. Within the cognitive-depth dimension, notably, Analysis/Reasoning had a meaningful positive prediction, while Factual/Recall negatively predicted gain (-0.32).

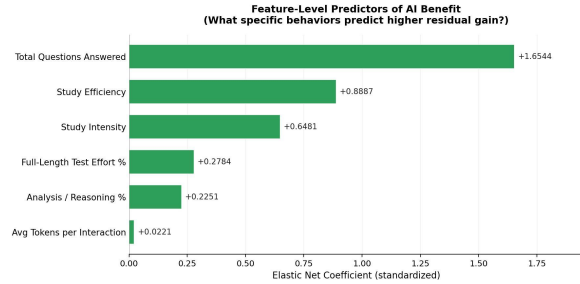


Figure 2: Standardized elastic-net coefficients for features positively predicting residual gain. All AI-volume metrics received zero weight (not shown).

Crucially, all AI interaction volume metrics received exactly zero weight. The number of chat sessions, student-AI interactions, and tokens contributed nothing to residual gain prediction once course effort and cognitive depth were in the model. This result is robust to reasonable perturbation of the elastic-net hyperparameters: volume features occupy a flat region of the coefficient surface with no predictive signal.

4.3 Learner Personas

Cluster diagnostics. All three model-selection criteria converged on $k = 4$: BIC dropped 540 points from $k = 3$ to $k = 4$ before rising at $k = 5$, AIC followed an analogous trajectory, and Silhouette peaked at $k = 4$ (Table 2 in Appendix). Under the final model, cluster assignments were highly stable across 10 random initializations ($ARI = 0.993 \pm 0.008$), and 96.7% of students were assigned to their persona with $> 80\%$ posterior probability (only 0.9% below 60%). The Silhouette score of 0.155 is low by conventional thresholds but expected given its sensitivity to the soft, overlapping cluster boundaries characteristic of GMMs; the personas nevertheless exhibit strong interpretive distinctness which is illustrated below.

Persona profiles. Four personas emerge from the GMM clustering (Table 1):

The cognitive fingerprints of these personas are qualitatively distinct (Figure 4). Maximizers exhibit a pronounced spike on Analysis/Reasoning; Grinders collapse almost entirely into Factual Recall despite comparable course effort; Satisficers show a balanced mid-depth profile; Sideliners are undifferentiated across all dimensions. This interpretive distinctness complements the statistical stability reported above and addresses concerns about geometric cluster separation.

Persona	n	%	Cognitive Signature	Residual Gain
Maximizer	25	3.7	Analysis + Application	+2.82
Satisficer	255	38.2	Conceptual + Application	+1.00
Grinder	43	6.4	Factual Recall	+0.77
Sidelineer	344	51.6	Shallow, undifferentiated	-0.07

Table 1: Behavioral definition, cognitive signature, and mean residual gain of the four learner personas.

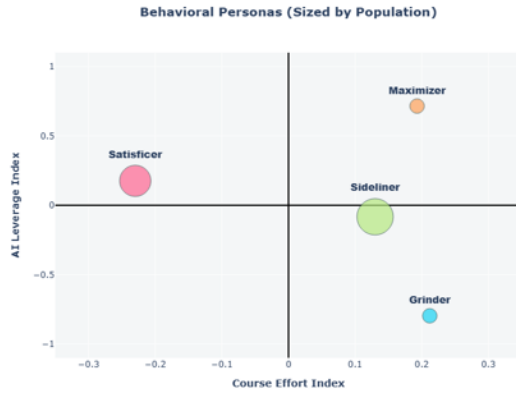


Figure 3: Learning personas in the Course Effort $\times$ AI Leverage plane (bubble size proportional to n -count).

The Maximizer–Grinder comparison. Figure 3 present the four learning personas across the two dimensions: Course Effect (x-axis) and AI-Leverage (y-axis). It is worth noting that both Maximizers and Grinders share high course effort but differ sharply in AI-Leverage. The residual gain gap of +2.05 points (+2.82 vs. +0.77) isolates the association between AI interaction cognitive depth and learning gain, net of effort. This within-treatment comparison holds the primary confounder (course effort) approximately constant while varying the dimension of interest (cognitive depth), providing the strongest quasi-experimental evidence available in this observational design.

Figure 4 further highlights the distinct cognitive-depth fingerprints of the four personas. Maximizers were uniquely characterized by substantial engagement in Analysis/Reasoning. Conversely, Grinders collapsed almost entirely into Factual/Recall despite high course effort, while Satisficers and Sidelineers clustered centrally with low-depth engagement.

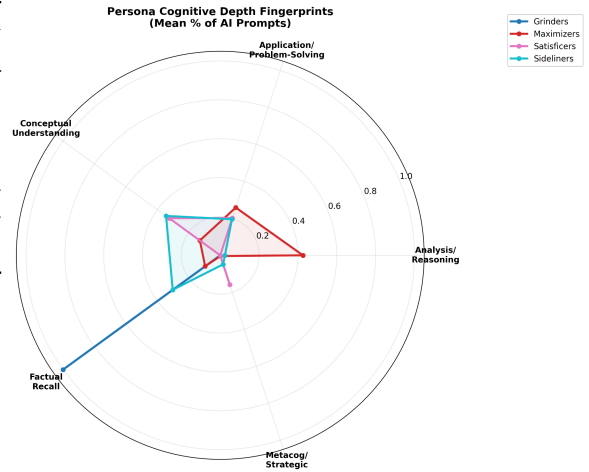


Figure 4: Mean cognitive depth distribution of AI interactions by persona. Maximizers spike on Analysis/Reasoning; Grinders collapse to Factual Recall.

5 Conclusion

This study demonstrates that the efficacy of AI tutoring is driven by the cognitive depth of student-AI interaction rather than sheer usage volume. Learners engaging in higher-order thinking (especially Analysis and Reasoning) achieved greater residual gains, whereas those focused primarily on Factual Recall derive negligible benefit.

The finding aligns with Bastani et al. (2025), who showed that unguided ChatGPT improved practice performance but harmed actual learning, while a purpose-built tutor with guardrails, especially providing scaffolding hints instead of giving away answers, avoided this harm. Our Maximizer profile represents the behavioral signature of deep, guided-like AI engagement; the Sidelineer profile represents unguided, shallow use.

The non-significant aggregate effect (+0.50) is largely explained by population composition: 51.6% of AI users are Sidelineers (gain $\approx$ 0), while only 3.7% are Maximizers (gain = +2.82). Aggregate efficacy metrics will systematically miss persona-level variation, and product decisions based on aggregate effects alone risk missing the important engagement factors driving learning gains.

By shifting the evaluation of AI tools from quantitative usage to qualitative cognitive fingerprints, this research provides vital evidence that student-AI interfaces should be optimized with Socratic scaffolding and deep-thinking prompts to foster deep cognitive engagement. Further research should focus on validating these findings across

broader academic domains and diverse student populations.

6 Future Work

To further strengthen the study, the following extensions are planned as immediate next step to validate the LLM-as-a-judge classification framework of cognitive depth.

- **Inter-rater reliability** A stratified sample of 50 interactions (8 per cognitive depth level, stratified by Gemini’s assigned label, plus 10 randomly drawn from the full pool) will be independently coded by human subject matter experts blind to the model’s classifications. The human experts will use the identical rubric definitions and few-shot examples supplied to the Gemini classifier. Cohen’s κ , per-class agreement, and a confusion matrix will be computed against the human experts. Both inter-rater agreement and human-model agreement will be evaluated.

Limitations

This study is quasi-experimental: students were not randomly assigned to AI usage or to specific engagement patterns, so unobserved confounders (e.g., metacognitive awareness, intrinsic motivation) cannot be ruled out despite baseline-ability adjustment and the effort-matched Maximizer–Grinder contrast. Furthermore, statistical power for the Maximizer persona is also limited ($n = 25$); the +2.82 residual-gain estimate is stable across cluster reassignments (96.7% posterior certainty) but derives from a small subgroup.

A future randomized trial study assigning students to scaffolded versus unscaffolded and unguided AI interaction, and replicated in a larger cohort, would strengthen the conclusion from current quasi-experimental study to a causal estimate of depth-of-engagement effects.

Finally, findings were derived from a single exam preparation product; cross-domain validation is needed to assess whether the same learning persona structure and AI interaction depth-over-volume result will generalize in other domains and subjects.

References

- Joshua D. Angrist and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.
- Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122.
- Benjamin S. Bloom. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals*. David McKay Company.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Greg Kestin, Kelly Miller, Anna Kiales, Timothy Milbourne, and Giada Ponti. 2024. AI tutoring outperforms active learning. *Harvard Initiative for Learning and Teaching Working Paper*.
- Ethan Mollick and Lilach Mollick. 2024. Instructors as innovators: A future-focused approach to new AI learning opportunities, with prompts. *Wharton School Working Paper*.
- Radek Pelánek, Jiří Řihák, and Jan Papoušek. 2016. [Impact of data collection on interpretation and evaluation of student models](#). In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 40–49. Association for Computing Machinery.
- Stephen W. Raudenbush. 2004. What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1):121–129.
- Douglas A. Reynolds. 2009. Gaussian mixture models. In Stan Z. Li and Anil Jain, editors, *Encyclopedia of Biometrics*, pages 659–663. Springer.
- William L. Sanders and Sandra P. Horn. 1998. Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3):247–256.
- William R. Shadish, Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.
- Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

k	BIC ($\downarrow$)	AIC ($\downarrow$)	Silhouette ($\uparrow$)
3	1042.1	965.5	0.097
4	502.0	398.5	0.155
5	537.7	407.1	0.099

Table 2: Model-selection criteria across candidate cluster counts. All three criteria identify $k = 4$ as optimal.

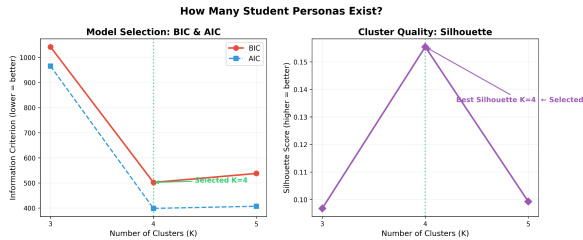


Figure 5: Cluster model selection. BIC and AIC (left) both minimize at $k = 4$; Silhouette (right) peaks at $k = 4$. All three criteria converge on four behavioral personas.

A Model Selection for Clustering

In Table 2, the BIC dropped 540 points from $k = 3$ to $k = 4$ before rising at $k = 5$; AIC followed an analogous trajectory. In Figure 5 Silhouette peaked at $k = 4$. The convergence of three independent criteria substantially reduces the risk that the four-cluster solution is an artifact of any single metric.

B Candidate Predictors

Table 3 lists all candidate predictors submitted to the elastic-net model, organized by family.

Family	Predictor	Definition
Course Activity	Total Questions Answered	Raw count of all practice questions answered
	Study Intensity	Total study hours / (study-window days + 1)
	Study Efficiency	Total questions answered $\div$ total study hours
	Full-Length Test Effort %	Share of total questions from full-length practice tests
	Quiz Effort %	Proportion of effort directed toward quizzes
	Question Effort %	Proportion of effort directed toward question solving
	Full-length Attempt %	Share of practice sessions that were full-length tests
	Quiz Attempt %	Proportion of attempts focused on Quiz practice
	QBank Attempt %	Proportion of attempts focused on QBank practice
Cognitive Depth	Questions per Quiz	Average questions per Quiz session
	Questions per QBank	Average questions per QBank session
	Analysis/Reasoning %	Proportion of AI interactions classified as Analysis/Reasoning
	Application %	Proportion classified as Application/Problem-Solving
	Conceptual Understanding %	Proportion classified as Conceptual Understanding
	Factual Recall %	Proportion classified as Factual Recall
Topic	Metacognitive/Strategic %	Proportion classified as Metacognitive/Strategic
	Human-to-AI Token Ratio	Mean ratio of student to AI tokens per interaction
	Contextual / Non-Content %	Proportion of AI interactions classified as Contextual/Non-Content (non-academic or logistical)
AI Volume	Section Focus % (A–E)	Proportion of AI interactions per content from A to E
AI Volume	AI Usage Volume	Standardized composite of session/interaction/token counts
	Avg Tokens per Interaction	Mean total tokens per interaction

Table 3: Full list of candidate predictors for elastic-net feature selection.