An Information-Theoretic Evaluation Framework for Benchmark and Model Diagnosis in Knowledge Tracing

Houru Jiang[1,*], **Zixi Wang**[1,*], **Tengteng Cheng**[2], **Xueyi Li**[2]
Mingliang Hou[2,†], Jiaqi Zheng[2], Renqiang Luo[3], Teng Guo[2], Zitao Liu[2]
[1]Jinan University, Guangzhou, China
[2]Guangdong Institute of Smart Education, Jinan University, Guangzhou, China
[3]Jilin University, Changchun, China
teemohold@outlook.com
[*]Equal contribution. [†]Corresponding author.

Abstract

Knowledge tracing (KT) models are predominantly evaluated using aggregate metrics such as area under the curve (AUC) and accuracy. However, these global scores obscure where the remaining errors originate and fail to indicate whether a benchmark is approaching saturation. While estimating a global theoretical performance limit is challenging in realistic KT settings, it is possible to quantify local predictability. To address this, we propose an information-theoretic evaluation framework for KT benchmark diagnosis. We use Context Tree Weighting (CTW) on item-response histories and current-item queries as an operational causal uncertainty coordinate, while distinguishing it from the unobserved Local Irreducible Uncertainty (LIU) under the full KT information set. By projecting predictions onto this shared uncertainty coordinate, we evaluate model performance gains across distinct entropy bands rather than only at the global level. Comprehensive evaluations on NIPS Task 3/4 and Algebra 2005 reveal that model improvements are highly non-uniform. Modern KT models show substantial gains in high-entropy regions, and additional item-aware references, log-loss, and equal-frequency analyses support this localization. The framework also flags regions where apparent gains require checks for noise-sensitive behavior. By surfacing these local modeling failures alongside genuine gains, this approach provides a diagnostic tool for studying both residual predictive structure and the limitations of current KT benchmarks and models.

1 Introduction

Knowledge tracing (KT) predicts a student's future responses from their past interaction history, serving as a foundational task in artificial intelligence (AI) for education. As modern KT architectures grow increasingly sophisticated, performance on standard benchmarks continues to climb. However, the evaluation paradigm remains dominated by aggregate metrics such as area under the curve (AUC) and accuracy. These global scores are useful for leaderboard rankings, but they fail to answer two diagnostic questions that matter for progress: *where does meaningful benchmark headroom remain?* and *what kind of errors is a model reducing?* A marginal $+0.01$ boost in global AUC cannot distinguish between a model that refines already-easy cases, one that extracts genuine structure from difficult interactions, and one whose apparent gain is concentrated in unstable regions where outcomes are close to random under the observable protocol.

Preprint.

To address this gap, we propose an *information-theoretic* evaluation framework for KT benchmark and model diagnosis. The framework separates two roles that aggregate evaluation conflates. First, **uncertainty characterization** establishes a shared local coordinate for prediction difficulty. Because the true Local Irreducible Uncertainty (LIU) is not directly observable in finite KT data, we use Context Tree Weighting (CTW) as an operational causal anchor: it assigns each prediction point a fold-consistent entropy estimate using past item-response history and the current queried item. Second, **predictive realization** measures how much task-facing KT models and anchor-side neural probes improve on that coordinate. This decoupling lets us ask not only whether a benchmark is moving on average, but also whether gains occur in low-uncertainty regions, high-uncertainty regions with exploitable structure, or high-uncertainty regions that call for item or data inspection.

Our contributions are as follows: (i) We introduce an information-theoretic evaluation paradigm that shifts KT assessment from global leaderboard scores to localized benchmark and model diagnosis; (ii) We formalize the separation between local irreducible uncertainty and realized predictive performance, with CTW serving as an operational causal anchor rather than as the ground-truth LIU; (iii) We provide a fold-consistent empirical protocol based on CTW-anchored entropy bands, gain-over-anchor curves, and item-level remaining-uncertainty/information-gain diagnostics; and (iv) Through broad aligned experiments on NIPS Task 3/4 and Algebra 2005, we show that the largest reliable gains occur in high-entropy regions while also identifying high-uncertainty, low-information-gain cases that warrant further reliability checks, with item-aware anchor and proper-score analyses testing the robustness of the localization.

2 Related Work

KT has progressed from Bayesian KT [Corbett and Anderson, 1994] to neural sequence models such as DKT [Piech et al., 2015], memory-based models such as DKVMN [Zhang et al., 2017], attention-based variants such as SAKT and AKT [Pandey and Karypis, 2019, Ghosh et al., 2020], and strong modern baselines such as Deep-IRT and simpleKT [Yeung, 2019, Liu et al., 2023b]. Toolkits and benchmarks including ASSISTments, EdNet, the NeurIPS 2020 Education Challenge, and pyKT have improved reproducibility [Heffernan and Heffernan, 2014, Choi et al., 2020b, Wang et al., 2021b, Liu et al., 2022]. However, these comparisons still primarily report aggregate metrics such as AUC and accuracy.

Our work is closer to probabilistic forecast evaluation. Information-theoretic uncertainty [Cover and Thomas, 2006], universal coding tools such as CTW [Willems et al., 1995], and proper scoring rules such as Brier score and log-loss [Brier, 1950, Gneiting and Raftery, 2007] provide tools for separating uncertainty from realized predictive quality. We use CTW as a common causal anchor and evaluate model gains along entropy bands rather than only at the benchmark level. This makes the contribution diagnostic rather than architectural: the objective is to determine where benchmark progress is reliable, where model capacity changes predictions, and where apparent gains are more consistent with unstable or noise-like item behavior.

3 Framework

3.1 Problem setup and local uncertainty

In standard KT, we observe a student's interaction history up to time t, denoted as $X_t = \{(q_1, r_1), \ldots, (q_t, r_t)\}$, where q_k represents the practiced item and $r_k \in \{0, 1\}$ represents the correctness of the response. Given the next item q_{t+1}, the task is to predict the probability of a correct response:

$$\hat{p}_{t+1} = \hat{P}(R_{t+1} = 1 \mid X_t, q_{t+1}). \tag{1}$$

To formalize the information available at this prediction step, we define the observable filtration as $\mathcal{F}_t = \sigma(X_t, q_{t+1})$. Based on this, we define the *local irreducible uncertainty* (LIU) of the prediction point as the conditional entropy of the true response R_{t+1}:

$$\mathrm{LIU}(X_t, q_{t+1}) = H(P(R_{t+1} \mid X_t, q_{t+1}))\,. \tag{2}$$

This term represents the theoretical entropy remaining under the admissible KT information boundary, i.e., the inherent unpredictability even if we had a perfect model for the student's historical data. In

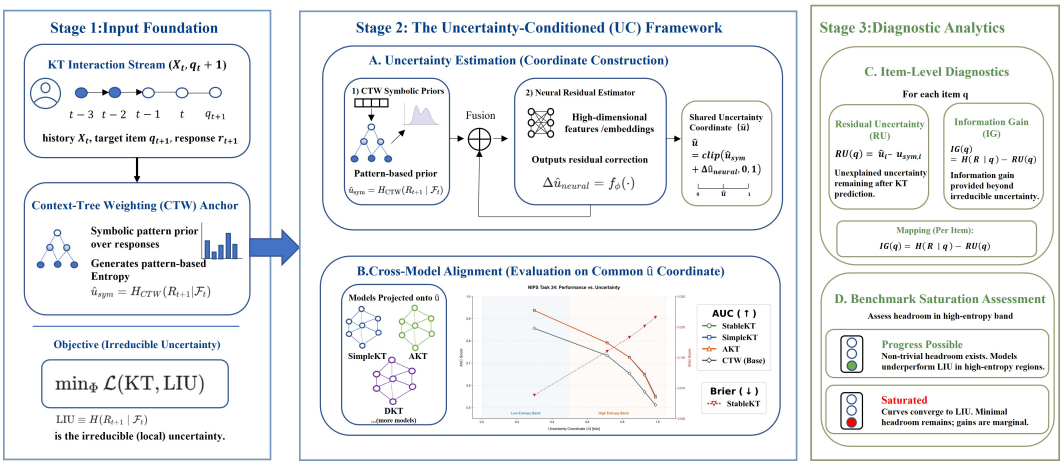


Figure 1: Overview of the proposed information-theoretic evaluation framework. A causal anchor assigns each prediction point to a shared uncertainty coordinate; model performance is then read along this coordinate to reveal where benchmark headroom remains and where model behavior warrants further diagnosis.

finite benchmark data, LIU is not directly observed; it must be studied through operational causal estimates that respect the information available at prediction time.

Rather than aggregating all predictions into a single global metric, our framework evaluates models locally. For each prediction instance i across the dataset, we compute an operational CTW uncertainty coordinate $\hat{u}_{\mathrm{ctw},i}$ as a causal proxy for LIU_i, partition the samples into distinct uncertainty buckets $\mathcal{B}_k$, and evaluate the model's performance $\mathrm{Perf}_m(k)$ within each specific bucket.

3.2 Entropy-risk decomposition and causal anchor

Let $P_t = P(R_{t+1} \mid \mathcal{F}_t)$ be the true conditional response distribution. For any $\mathcal{F}_t$-measurable probabilistic predictor $\hat{P}_\theta$, the expected log-loss (cross-entropy) naturally decomposes into two components:

Fact 1 (Local entropy-risk decomposition)**.**

$$\mathbb{E}\big[-\log \hat{P}_\theta(R_{t+1} \mid \mathcal{F}_t)\big] = \underbrace{H(R_{t+1} \mid \mathcal{F}_t)}_{\textit{mean LIU}} + \underbrace{\mathbb{E}\big[\mathrm{KL}(P_t \,\|\, \hat{P}_\theta)\big]}_{\textit{epistemic gap}}. \tag{3}$$

This decomposition illustrates that standard benchmark-level error inextricably mixes LIU with model mismatch (the epistemic gap). A global score can decrease either because a model reduces epistemic error in a meaningful region or because gains and failures cancel across regions with very different local uncertainty.

All entropies and KL divergences in the theoretical identities use base-2 logarithms, so they are measured in bits. The outer expectation in Fact 1 averages over prediction contexts; the local entropy is the entropy of P_t at a given context.

To study uncertainty operationally, we use Context Tree Weighting (CTW) as an autoregressive causal anchor. In the real-data implementation, its context contains past item-response tokens, encoded as `I:item|R:resp`, and a query token `Q:item` for the current item. At prediction step $t+1$, these tokens are constructed from X_t and q_{t+1}; the current response is observed only after prediction, and future responses are excluded. Thus item identity is a predictor input, while an explicit Q-matrix or skill-tag representation is not required. The corresponding uncertainty coordinate is $\hat{u}_{\mathrm{ctw},t} = H(p_{\mathrm{ctw}})$. This is a model-dependent estimate under the causal information boundary $\mathcal{F}_t = \sigma(X_t, q_{t+1})$, not a response-only coordinate or a certified Bayes limit. Gains over CTW

can reflect limitations of its finite-depth context model even when predictors share this information boundary. We also test localization under an independent ItemRasch reference that models item difficulty and online ability differently (Appendix E).

Furthermore, we introduce a fused CTW+NN predictor. This model augments the pure CTW anchor by applying a neural refinement of the form $p_{\mathrm{ctw}} \mapsto \sigma(\mathrm{logit}(p_{\mathrm{ctw}}) + f_\phi(X_t, q_{t+1}))$. In our evaluation protocol, the pure CTW model defines the shared uncertainty axis (i.e., the buckets), while the CTW+NN model acts as an anchor-side neural probe. It shows whether exploitable structural signals still exist beyond the capabilities of the symbolic source model.

3.3 Operational bands, remaining uncertainty, and information gain

Given an aligned evaluation set, our empirical protocol proceeds as follows: First, we compute a fold-consistent CTW prediction $p_{\mathrm{ctw},i}$ for each evaluation point i. Second, we convert this probability into an estimated entropy score $\hat{u}_{\mathrm{ctw},i} = H(p_{\mathrm{ctw},i})$. Third, we partition the samples into distinct uncertainty buckets (using fixed-width bands in the original analysis and equal-frequency quintiles in the additional robustness analysis). Finally, we report both the absolute performance and the performance gain over the CTW anchor within each bucket. This resulting band view is deliberately model-agnostic: any KT baseline can be evaluated on this exact same CTW-defined coordinate.

The same autoregressive anchor also enables fine-grained, item-level diagnostics. For any given item q, let H_q^{marg} denote the entropy of its marginal empirical correctness, and let H_q^{ctw} denote its average CTW conditional entropy across all predictions for that item. We define the *remaining uncertainty* as $\mathrm{RU}(q) = H_q^{\mathrm{ctw}}$ and the *information gain* as $\mathrm{IG}(q) = H_q^{\mathrm{marg}} - H_q^{\mathrm{ctw}}$. A high RU identifies items whose outcomes remain unstable under the current observable protocol (e.g., potentially due to guessing, ambiguous wording, missing observables, or overly coarse step granularity); a high IG highlights items with lower anchor entropy than item-marginal entropy. This plug-in difference is a diagnostic rather than a guaranteed estimate of mutual information; anchor misspecification can affect either quantity.

3.4 Benchmark and model diagnosis criteria

The framework answers three diagnostic questions. First, **benchmark diagnosis** asks whether high-uncertainty regions with substantial support still admit reliable gains over the causal anchor; if strong models remain separated there, aggregate saturation claims are premature. Second, **model diagnosis** asks where a model improves and whether those improvements are supported by proper scoring behavior, not only by threshold accuracy. Third, **noise-sensitive behavior** asks whether apparent improvements occur in high-RU/low-IG regions where the item-conditioned anchor leaves substantial uncertainty. Such movements should not be interpreted as progress unless they are stable across folds and accompanied by proper-score improvements.

4 Experimental Protocol

We evaluate five-fold aligned exports on NIPS Task 3/4 and Algebra 2005. ASSIST2015 is reported as a global auxiliary check, while the local diagnostic figures focus on NIPS Task 3/4 and Algebra 2005 because these datasets support the band-level and item-level diagnostics used in the main figures. Each prediction point is assigned a CTW entropy value and placed into one of five entropy bands $[0.0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1.0]$. We report AUC, accuracy, and Brier score globally and within bands. Accuracy gives the decision-level view, while Brier score is used as a proper-score check on whether probability quality improves in the same region.

The model pool includes the CTW anchor, the CTW+NN neural probe, and a broad KT sweep spanning DKT, DKT+, DKT-forget, KQN, SAKT, SAINT, AKT, simpleKT, sparseKT, FoLiBiKT, stableKT, extraKT, Deep-IRT, ATKT, ATKTFix, DIMKT, ATDKT, and HawkesKT [Yeung and Yeung, 2018, Nagatani et al., 2019, Lee and Yeung, 2019, Choi et al., 2020a, Guo et al., 2021, Huang et al., 2023, Im et al., 2023, Li et al., 2024a,b, Shen et al., 2022, Liu et al., 2023a, Wang et al., 2021a]. The main bucket coordinate is always computed from CTW alone, so task-facing KT baselines do not define the axis on which they are judged. Model diagnosis is then based on the alignment between gain-over-anchor curves, proper-score changes, fold consistency, and item-level RU/IG signals. The additional robustness analysis uses a single SimpleKT residual probe, denoted CTW+NN-SKT, on

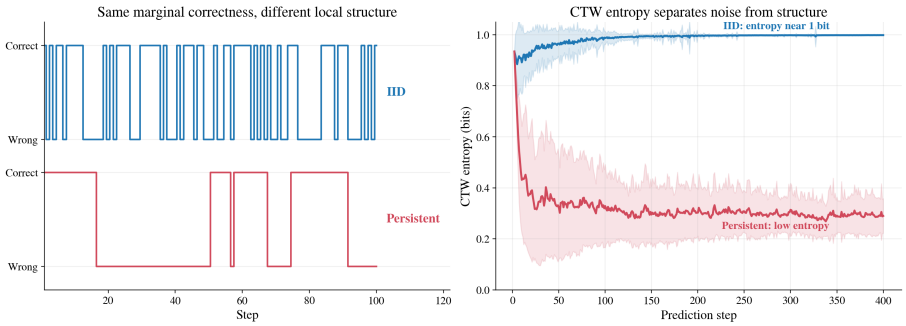


Figure 2: Synthetic sanity check. Two regimes have nearly identical marginal correctness near 0.5, but CTW separates noise-like responses from run-structured responses with strong causal predictability.

a separate common evaluation cohort; it is distinct from the original reported CTW+NN export. Appendix E reports its log-loss, calibration, item-aware reference, and quintile results separately.

5 Results

5.1 Sanity check: same marginal correctness, different causal uncertainty

Before real benchmarks, we verify the property needed from the anchor. We construct two binary response regimes with nearly identical marginal correctness rates. The first is an i.i.d. Bernoulli(0.5) process. The second is a symmetric persistent two-state Markov process: the initial state is uniform over $\{0, 1\}$, and each subsequent response copies the previous response with probability 0.95 and flips with probability 0.05. This process has stationary marginal correctness 0.5 but forms long correct/incorrect runs. Across 64 sequences of length 400, the observed correctness rates are 0.492 and 0.495, yet average CTW entropy is 0.986 bits for the i.i.d. regime and 0.318 bits for the persistent regime. CTW predictive accuracy is 0.501 versus 0.945.

5.2 Global aligned performance is insufficient for diagnosis

Global aligned performance is used as context rather than as the main diagnostic target. Across the aligned folds, CTW+NN improves the CTW anchor on all three datasets while remaining an anchor-side neural probe rather than the definition of the uncertainty axis: $+0.0823$ AUC and $+0.0690$ accuracy on NIPS Task 3/4, $+0.0835$ AUC and $+0.0160$ accuracy on Algebra 2005, and $+0.0455$ AUC and $+0.0142$ accuracy on ASSIST2015. The task-facing KT sweep is reported in Table 1. These numbers show that stronger predictors improve on average, but they do not reveal whether the improvement comes from already predictable regions, genuinely difficult but structured regions, or unstable regions where apparent gains require caution. The central claim is local: global scores provide a sanity check, while the entropy-conditioned bands below test where benchmark headroom and model limitations actually lie.

5.3 Benchmark diagnosis through entropy bands

The main empirical pattern is local rather than global. On NIPS Task 3/4, the accuracy improvement of CTW+NN over CTW grows from $+0.0026$ in the lowest entropy band to $+0.0886$ in the highest. On Algebra 2005, gains are small in lower bands but reach $+0.0283$ accuracy and 0.0681 Brier reduction in the highest band. Moreover, the highest band is not a tiny tail: it contains roughly 74.8% of NIPS Task 3/4 samples and 55.6% of Algebra 2005 samples in the shared aligned subsets. These high-entropy regions are therefore central to the benchmark rather than peripheral outliers.

5.4 Model diagnosis through gain profiles

The band profile should be read as a diagnostic object rather than as a replacement for global model ranking. The absolute-accuracy row in Figure 3 answers a first-order question: does the CTW coordinate actually stratify prediction difficulty? In both main datasets, the answer is yes. Accuracy

Table 1: Task-facing KT baselines across the three benchmark datasets. Each metric cell is formatted as a two-line entry with the mean emphasized and the standard deviation reduced for readability. Entries are mean ± standard deviation over five folds; lower Brier is better. Best entries are bolded and second-best entries are underlined within each dataset/metric block. The unified table keeps only the KT models available on all three datasets. All NIPS Task 3/4 rows use the shared aligned subset of 1,354,735 points. For Algebra 2005, all rows use 464,725 points except ATKT (271,122) and ATKTFix (278,835). For ASSIST2015 under the convention $concept = question$, all rows use 672,960 points.

	NIPS Task 3/4			Algebra 2005			ASSIST2015		
Model	AUC	ACC	Brier	AUC	ACC	Brier	AUC	ACC	Brier
simpleKT	**0.8014** ± 0.0003	0.7309 ± 0.0004	0.1798 ± 0.0005	0.8147 ± 0.0225	0.8065 ± 0.0072	0.1389 ± 0.0060	0.7231 ± 0.0012	0.7502 ± 0.0006	0.1713 ± 0.0003
stableKT	0.8011 ± 0.0005	**0.7310** ± 0.0003	**0.1796** ± 0.0003	0.8200 ± 0.0181	0.8070 ± 0.0063	0.1365 ± 0.0052	0.7256 ± 0.0010	0.7506 ± 0.0002	0.1707 ± 0.0002
AKT	0.7999 ± 0.0005	0.7292 ± 0.0008	0.1805 ± 0.0004	0.8277 ± 0.0015	0.8110 ± 0.0015	0.1334 ± 0.0007	0.7262 ± 0.0008	0.7519 ± 0.0008	0.1708 ± 0.0002
FoLiBiKT	0.7993 ± 0.0008	0.7291 ± 0.0010	0.1809 ± 0.0002	0.8026 ± 0.0141	0.8017 ± 0.0049	0.1413 ± 0.0043	0.7261 ± 0.0013	0.7520 ± 0.0004	0.1707 ± 0.0004
extraKT	0.7982 ± 0.0020	0.7285 ± 0.0016	0.1810 ± 0.0011	0.8215 ± 0.0029	0.8064 ± 0.0018	0.1365 ± 0.0011	0.7257 ± 0.0004	0.7506 ± 0.0002	0.1712 ± 0.0001
sparseKT-soft	0.7979 ± 0.0002	0.7276 ± 0.0003	0.1813 ± 0.0002	0.8155 ± 0.0009	0.8031 ± 0.0011	0.1395 ± 0.0006	0.7206 ± 0.0013	0.7494 ± 0.0004	0.1718 ± 0.0003
sparseKT-topk	0.7974 ± 0.0010	0.7271 ± 0.0008	0.1814 ± 0.0004	0.8137 ± 0.0012	0.8020 ± 0.0009	0.1407 ± 0.0002	0.7161 ± 0.0066	0.7490 ± 0.0024	0.1728 ± 0.0016
ATKT	0.7776 ± 0.0062	0.7086 ± 0.0045	0.1904 ± 0.0026	**0.8373** ± 0.0229	**0.8179** ± 0.0189	**0.1280** ± 0.0128	**0.8069** ± 0.0100	**0.7787** ± 0.0062	**0.1494** ± 0.0035
DKT+	0.7660 ± 0.0006	0.7010 ± 0.0007	0.1947 ± 0.0003	0.8098 ± 0.0016	0.8078 ± 0.0006	0.1372 ± 0.0005	0.7283 ± 0.0007	0.7510 ± 0.0004	0.1702 ± 0.0002
DKT	0.7659 ± 0.0002	0.7005 ± 0.0004	0.1948 ± 0.0001	0.8091 ± 0.0007	0.8072 ± 0.0005	0.1374 ± 0.0004	0.7267 ± 0.0005	0.7506 ± 0.0006	0.1706 ± 0.0001
KQN	0.7626 ± 0.0011	0.6982 ± 0.0006	0.1961 ± 0.0005	0.7883 ± 0.0062	0.7985 ± 0.0020	0.1437 ± 0.0017	0.7250 ± 0.0007	0.7500 ± 0.0003	0.1709 ± 0.0001
SAINT	0.7621 ± 0.0303	0.6964 ± 0.0258	0.1975 ± 0.0105	0.7660 ± 0.0133	0.7780 ± 0.0031	0.1607 ± 0.0039	0.7007 ± 0.0020	0.7441 ± 0.0015	0.1765 ± 0.0009
Deep-IRT	0.7599 ± 0.0029	0.6951 ± 0.0025	0.1971 ± 0.0013	0.7949 ± 0.0020	0.8006 ± 0.0007	0.1421 ± 0.0005	0.7185 ± 0.0003	0.7501 ± 0.0004	0.1719 ± 0.0001
ATKTFix	0.7599 ± 0.0032	0.6956 ± 0.0024	0.1973 ± 0.0012	0.7841 ± 0.0051	0.7941 ± 0.0035	0.1453 ± 0.0017	0.7213 ± 0.0035	0.7485 ± 0.0013	0.1719 ± 0.0008
SAKT	0.7481 ± 0.0004	0.6844 ± 0.0010	0.2031 ± 0.0011	0.7867 ± 0.0037	0.7964 ± 0.0016	0.1477 ± 0.0022	0.7112 ± 0.0004	0.7472 ± 0.0001	0.1756 ± 0.0007

Table 2: CTW+NN minus CTW improvement by CTW entropy band. Positive ΔACC and negative ΔBrier indicate gains over the anchor.

Band	NIPS Task 3/4		Algebra 2005	
	ΔACC	ΔBrier	ΔACC	ΔBrier
$[0.0, 0.2)$	$+0.0026$	-0.0075	-0.0002	-0.0006
$[0.2, 0.4)$	$+0.0037$	-0.0144	$+0.0002$	-0.0019
$[0.4, 0.6)$	$+0.0055$	-0.0161	-0.0001	-0.0041
$[0.6, 0.8)$	$+0.0147$	-0.0229	$+0.0021$	-0.0087
$[0.8, 1.0]$	$+\mathbf{0.0886}$	$-\mathbf{0.0378}$	$+\mathbf{0.0283}$	$-\mathbf{0.0681}$

falls as CTW entropy increases, and this trend is visible not only for the CTW anchor but also for the KT baselines. This matters because the coordinate is computed without access to future labels or to the task-facing KT model being evaluated. It therefore provides a common, causal axis on which different predictors can be compared.

The delta row answers a second-order question: where is there still exploitable headroom relative to the anchor? If a benchmark were nearly exhausted under the current observable protocol, then high-entropy regions should show little remaining separation among strong models. That is not the observed pattern. On NIPS Task 3/4, the CTW+NN neural probe barely moves the lowest-entropy band but gains almost nine accuracy points in the highest band. The main KT baselines follow the same qualitative profile. This shows that the difficult region is not merely random noise: it is difficult for the symbolic anchor, but still contains structure that stronger predictors can exploit.

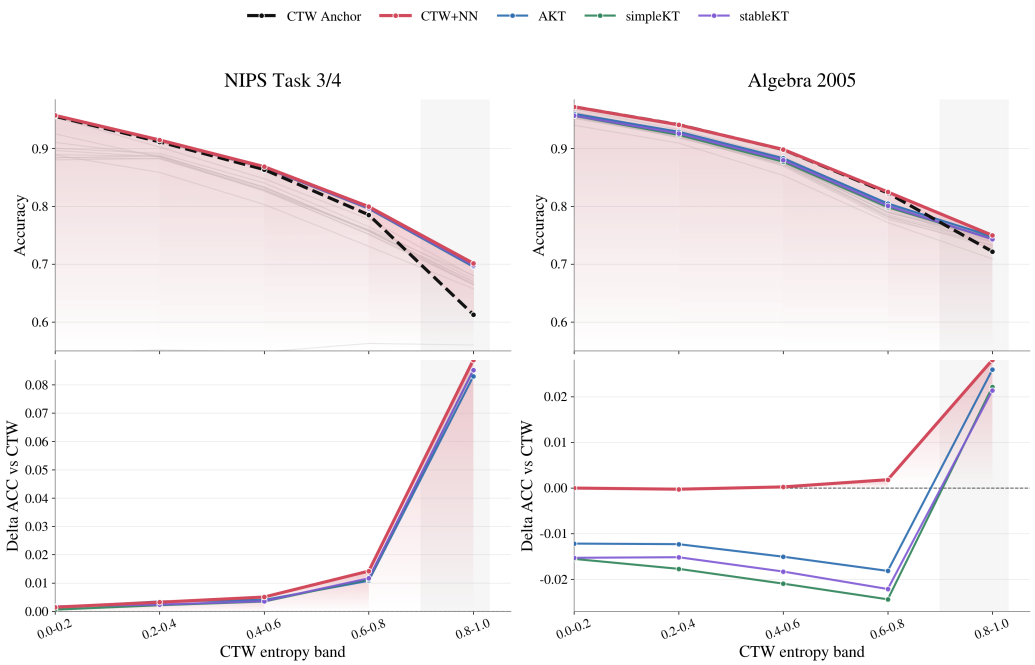


Figure 3: Entropy-conditioned accuracy bands under the CTW anchor. Top: absolute accuracy by CTW entropy band. Bottom: accuracy gain relative to CTW. Higher CTW-entropy bands have lower accuracy, but the largest gains over CTW also occur in these highest-entropy regions, indicating difficult but not exhausted benchmark regions.

Algebra 2005 shows a weaker low-band pattern but the same high-band conclusion. In the first four entropy bands, CTW+NN is close to CTW in accuracy, while several KT baselines can sit below the anchor depending on coverage and model family. In the highest band, however, CTW+NN separates clearly, and the Brier reduction is largest. This is exactly the regime where a local diagnostic is more informative than a global score: aggregate performance averages together easy regions that are already largely solved and high-uncertainty regions where model choice still matters.

This interpretation also clarifies why CTW+NN is included. It is not used to define the benchmark axis, and it is not presented as a new KT leaderboard method. Its role is to act as an anchor-side neural probe: if a simple neural refinement of the causal anchor improves most in high-entropy regions, then those regions should not be dismissed as irreducible noise. They are hard, but not exhausted. By contrast, flat delta curves in high-entropy bands across strong models would indicate a more saturated benchmark under the same observable protocol.

The same profile is also a model-diagnosis tool, but the roles must remain separate: CTW+NN is an anchor-side probe for remaining signal, not a task-facing KT baseline. For task-facing models, a gain is most credible when threshold accuracy and Brier score move in the same direction within the same region. Isolated accuracy changes in high-entropy regions should be inspected rather than immediately counted as progress, because high entropy can arise from both exploitable hidden structure and effectively random variation under the observed features. This is why the framework pairs band-level gains with item-level RU/IG diagnostics: the band profile identifies where a model changes behavior, while RU/IG asks whether that region contains contextual signal or unstable item behavior.

5.5 Sensitivity and noise-sensitive item diagnostics

Two additional checks support the same interpretation. First, repartitioning samples by a refined uncertainty coordinate $\hat{u}_{refined} = H(\hat{p}_{fused})$ preserves the non-uniform headroom profile, while smoothing the NIPS gain curve because the fused predictor already absorbs part of the structure left unresolved by CTW. On Algebra 2005, higher refined-uncertainty buckets still exhibit larger

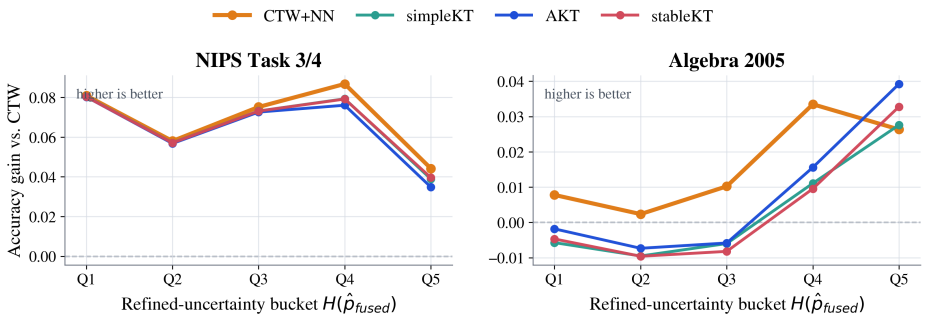


Figure 4: Sensitivity analysis using the refined uncertainty coordinate $\hat{u}_{refined} = H(\hat{p}_{fused})$. The compact view focuses on accuracy gains relative to the CTW anchor, the primary headroom signal used in the band analysis. The refined coordinate preserves a non-uniform gain profile while avoiding the visual clutter of duplicating absolute-metric curves.

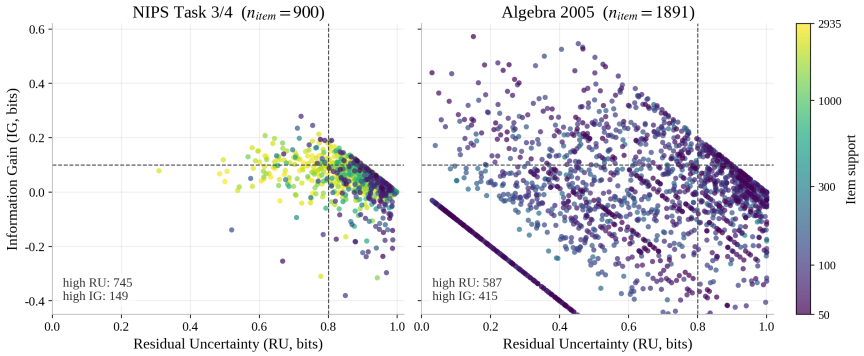


Figure 5: Item-level diagnostic map under the CTW anchor. RU captures remaining item uncertainty and IG captures contextual signal beyond item marginal correctness. High-RU/low-IG regions are the main warning cases for noise-sensitive model behavior, while high-IG regions indicate contextual structure that models can exploit.

accuracy gains over the CTW anchor; on NIPS Task 3/4, the peak moves from the last bucket to the upper-middle region, consistent with the refined coordinate already encoding part of the additional signal.

Second, item-level RU and IG separate item regions whose outcomes remain unstable from regions where history carries strong predictive signal. This is the main mechanism by which the framework surfaces noise-sensitive model behavior. High-RU/low-IG items motivate inspection for ambiguity, guessing, missing observables, or problematic granularity; if a model's apparent improvement is concentrated in such items without proper-score or fold-stability support, the improvement is a warning signal for unstable prediction behavior; it does not by itself establish that the model has learned random noise. High-IG items, in contrast, point to regions where better temporal or prerequisite modeling should pay off. Figure 5 shows this support-filtered diagnostic map.

The RU/IG view makes the distinction concrete. High-RU/low-IG regions have high anchor entropy relative to item-marginal entropy, so apparent model gains on them require proper-score and fold-stability checks before being treated as progress. Figure 6 compares all task-facing baselines with a direct two-score diagnostic: the x-coordinate asks whether accuracy improves over CTW, the y-coordinate asks whether the proper score improves over CTW, and the green quadrant marks gains supported by both criteria. The two-score view converts the RU/IG map from an item-level warning signal into a model-level reliability check. The strongest NIPS models cluster in the green quadrant for high-IG regions, whereas Algebra and high-RU/low-IG panels contain more leftward or low-y points. Thus uncertainty bands identify *where* headroom remains, while RU/IG distinguishes exploitable structure from unstable item behavior.

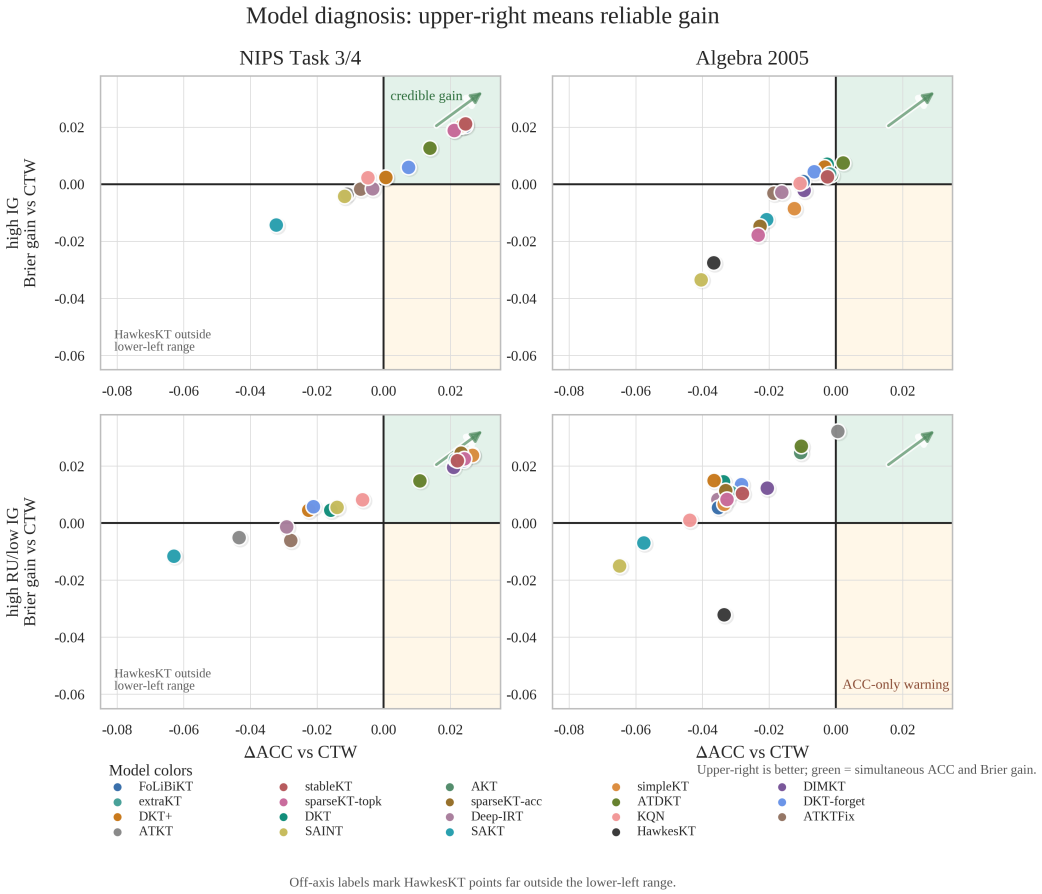


Figure 6: Model diagnosis scatter plot for task-facing KT baselines. Columns separate datasets and rows separate RU/IG regions. Each point is one KT model, with color identifying the model. The horizontal axis reports ΔACC versus CTW and the vertical axis reports Brier gain ($-\Delta$Brier) versus CTW, so farther to the upper-right is better. Points in the green upper-right quadrant improve both threshold accuracy and probability quality; points moving right but not up indicate accuracy-only gains that require caution.

Robustness to the reference and evaluation resolution. An independent ItemRasch reference uses an empirical-Bayes item estimate and a causal online student state, providing an alternative to CTW's item-token context model. In its highest-entropy band, CTW+NN-SKT improves accuracy by $+0.0420$ on NIPS and $+0.0788$ on Algebra, with favorable Brier and log-loss changes and five-fold confidence intervals. On the CTW coordinate, highest-band log-loss changes are -0.0902 and -0.1578 nats, respectively. Equal-frequency quintiles resolve gradients inside the large fixed-width band, and more than 93% of high-CTW-entropy points occur after interaction 20. These checks support localization across alternative item-conditioned reference models and beyond cold start; full results and protocol distinctions appear in Appendix E.

6 Discussion and Limitations

The results support information-theoretic reporting as a benchmark- and model-diagnosis tool. A benchmark close to saturation should show tight clustering of strong models especially in high-uncertainty regions. We do not observe that pattern for NIPS Task 3/4 or Algebra 2005: the hardest CTW-defined regions still contain the largest gains. This suggests that aggregate KT scores hide substantial local structure and that residual structure remains under the tested references and evaluation protocols; it does not quantify distance to the Bayes limit.

CTW is used here as a causal anchor and common analysis coordinate, not as a complete model of student learning. This localizes where headroom remains while avoiding the confound that would arise if every model defined its own uncertainty axis. The refined-coordinate and Brier checks reduce the risk that the conclusion is an artifact of CTW-only bucketization or threshold accuracy. At the same time, the framework avoids treating every high-entropy gain as progress: high-RU/low-IG items expose regions where model changes should be interpreted as noise-sensitive unless supported by proper-score improvement and fold stability.

The analysis has several scope boundaries. The operational coordinates are model-dependent estimates, and their entropy is not a certified lower bound on achievable error. Five-fold intervals describe variation across the retained folds; overlapping training sets limit interpretations based on independent-fold assumptions. The additional CTW+NN-SKT analysis uses a distinct common cohort and must not be numerically pooled with the original reported branch. First, alternative uncertainty coordinates may reveal complementary structure. Second, some raw KT exports have dataset-specific coverage differences, so the main claims emphasize shared aligned subsets. Third, Figure 6 reports model diagnosis at the RU/IG-region level, while the paper reports item-level RU/IG summaries for NIPS Task 3/4 and Algebra 2005.

7 Conclusion

We introduced an information-theoretic evaluation framework for KT benchmark and model diagnosis. By separating uncertainty characterization from predictive realization and using CTW entropy as a shared causal coordinate, the framework shows where model progress occurs. Across broad aligned sweeps, the largest gains concentrate in high-uncertainty regions, demonstrating that these benchmarks contain substantial exploitable structure beyond what aggregate scores reveal.

The practical message is that KT benchmarks should not be judged only by aggregate score movement. A small global improvement can be meaningful if it occurs in a difficult, high-uncertainty region with substantial support and proper-score improvement; conversely, gains concentrated in high-RU/low-IG regions may reflect noise-sensitive behavior. A minimal adoption path is straightforward: report global metrics, add entropy-conditioned gain curves on a fixed causal coordinate, and inspect RU/IG regions when the curve shows remaining headroom.

Acknowledgments and Disclosure of Funding

This work received no funding. The authors declare no competing interests.

References

Glenn W. Brier. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, 78(1):1–3, January 1950. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Byungsoo Kim, Yeongmin Cha, Dongmin Shin, Chan Bae, and Jaewe Heo. Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, L@S '20, pages 341–344, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450379519. doi: 10.1145/3386527.3405945. URL https://doi.org/10.1145/3386527.3405945.

Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. EdNet: A Large-Scale Hierarchical Dataset in Education. In Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán, editors, *Artificial Intelligence in Education*, pages 69–73, Cham, 2020b. Springer International Publishing. ISBN 9783030522407. doi: 10.1007/978-3-030-52240-7_13.

Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, December 1994. ISSN 1573-1391. doi: 10.1007/BF01099821. URL https://doi.org/10.1007/BF01099821.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, September 2006. ISBN 9780471241959 9780471748823. doi: 10.1002/047174882X. URL https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X.

Aritra Ghosh, Neil Heffernan, and Andrew S. Lan. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 2330–2339, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403282. URL https://doi.org/10.1145/3394486.3403282.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Xiaopeng Guo, Zhijie Huang, Jie Gao, Mingyu Shang, Maojing Shu, and Jun Sun. Enhancing knowledge tracing via adversarial training. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 367–375, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475554. URL https://doi.org/10.1145/3474085.3475554.

Neil T. Heffernan and Cristina Lindquist Heffernan. The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, December 2014. ISSN 1560-4306. doi: 10.1007/s40593-014-0024-x. URL https://doi.org/10.1007/s40593-014-0024-x.

Shuyan Huang, Zitao Liu, Xiangyu Zhao, Weiqi Luo, and Jian Weng. Towards robust knowledge tracing models via k-sparse attention. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 2441–2445, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3592073. URL https://doi.org/10.1145/3539618.3592073.

Yoonjin Im, Eunseong Choi, Heejin Kook, and Jongwuk Lee. Forgetting-aware linear bias for attentive knowledge tracing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 3958–3962. ACM, October 2023. doi: 10.1145/3583780.3615191. URL http://dx.doi.org/10.1145/3583780.3615191.

Jinseok Lee and Dit-Yan Yeung. Knowledge query network for knowledge tracing: How knowledge interacts with skills. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK19, pages 491–500, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362566. doi: 10.1145/3303772.3303786. URL https://doi.org/10.1145/3303772.3303786.

Xueyi Li, Youheng Bai, Teng Guo, Zitao Liu, Yaying Huang, Xiangyu Zhao, Feng Xia, Weiqi Luo, and Jian Weng. Enhancing length generalization for attention based knowledge tracing models with linear biases. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 5918–5926. International Joint Conferences on Artificial Intelligence Organization, 8 2024a. doi: 10.24963/ijcai.2024/654. URL https://doi.org/10.24963/ijcai.2024/654. Main Track.

Xueyi Li, Youheng Bai, Teng Guo, Ying Zheng, Mingliang Hou, Bojun Zhan, Yaying Huang, Zitao Liu, Boyu Gao, and Weiqi Luo. Extending Context Window of Attention Based Knowledge Tracing Models via Length Extrapolation. In Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto Bugarín-Diz, José M. Alonso-Moral, Senén Barro, and Fredrik Heintz, editors, *Frontiers in Artificial Intelligence and Applications*. IOS Press, October 2024b. ISBN 9781643685489. doi: 10.3233/FAIA240651. URL https://ebooks.iospress.nl/doi/10.3233/FAIA240651.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Jiliang Tang, and Weiqi Luo. pyKT: a python library to benchmark deep learning based knowledge tracing models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. doi: 10.52202/068431-1347. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/75ca2b23d9794f02a92449af65a57556-Abstract-Datasets_and_Benchmarks.html.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Boyu Gao, Weiqi Luo, and Jian Weng. Enhancing deep knowledge tracing with auxiliary tasks. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pages 4178–4187, New York, NY, USA, 2023a. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583866. URL https://doi.org/10.1145/3543507.3583866.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. simplekt: A simple but tough-to-beat baseline for knowledge tracing, 2023b. URL https://arxiv.org/abs/2302.06881.

Koki Nagatani, Qian Zhang, Masahiro Sato, Yan-Ying Chen, Francine Chen, and Tomoko Ohkuma. Augmenting knowledge tracing by considering forgetting behavior. In *The World Wide Web Conference*, WWW '19, pages 3101–3107, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313565. URL https://doi.org/10.1145/3308558.3313565.

Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. *CoRR*, abs/1907.06837, 2019. URL http://arxiv.org/abs/1907.06837.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf.

Shuanghong Shen, Zhenya Huang, Qi Liu, Yu Su, Shijin Wang, and Enhong Chen. Assessing student's dynamic knowledge state by exploring the question difficulty effect. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 427–437, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531939. URL https://doi.org/10.1145/3477495.3531939.

Chenyang Wang, Weizhi Ma, Min Zhang, Chuancheng Lv, Fengyuan Wan, Huijie Lin, Taoran Tang, Yiqun Liu, and Shaoping Ma. Temporal cross-effects in knowledge tracing. In *Proceedings of the*

14th ACM International Conference on Web Search and Data Mining, WSDM '21, pages 517–525, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441802. URL https://doi.org/10.1145/3437963.3441802.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E. Turner, Richard G. Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. Results and insights from diagnostic questions: The neurips 2020 education challenge, 2021b. URL https://arxiv.org/abs/2104.04034.

F. M.J. Willems, Y. M. Shtarkov, and T. J. Tjalkens. The context-tree weighting method: basic properties. *IEEE Trans. Inf. Theor.*, 41(3):653–664, May 1995. ISSN 0018-9448. doi: 10.1109/18.382012. URL https://doi.org/10.1109/18.382012.

Chun-Kit Yeung. Deep-irt: Make deep learning based knowledge tracing explainable using item response theory, 2019. URL https://arxiv.org/abs/1904.11738.

Chun-Kit Yeung and Dit-Yan Yeung. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, L@S '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358866. doi: 10.1145/3231644.3231647. URL https://doi.org/10.1145/3231644.3231647.

Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 765–774, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052580. URL https://doi.org/10.1145/3038912.3052580.

A Extended Theoretical Framework and Proofs

A.1 Binary entropy properties and the LIU proxy

For a binary correctness probability $p \in [0, 1]$, the entropy in bits is

$$H_2(p) = -p \log_2 p - (1-p) \log_2 (1-p). \tag{A.1}$$

Its first two derivatives on $(0, 1)$ are

$$\frac{dH_2}{dp}(p) = \log_2 \frac{1-p}{p}, \qquad \frac{d^2 H_2}{dp^2}(p) = -\frac{1}{(\ln 2)\, p(1-p)} < 0. \tag{A.2}$$

Equation (A.2) shows three properties that matter for the appendix. First, $H_2(p)$ is symmetric around $p = 0.5$. Second, it is strictly increasing on $[0, 0.5]$ and strictly decreasing on $[0.5, 1]$. Third, it is strictly concave. Entropy measures the uncertainty of a predictive distribution; it is not a loss for evaluating whether a confident prediction is correct. As a result, $H_2(p)$ is small when the next response is causally predictable and maximal at one bit when the best available causal probability is near chance.

Fact A.1 (Local entropy-risk decomposition)**.** *For any $\mathcal{F}_t$-measurable probabilistic predictor $\hat{P}_\theta$,*

$$\mathbb{E}\big[-\log \hat{P}_\theta(R_{t+1} \mid \mathcal{F}_t)\big] = H(R_{t+1} \mid \mathcal{F}_t) + \mathbb{E}\big[\mathrm{KL}(P_t \| \hat{P}_\theta)\big]. \tag{A.3}$$

Proof. Let $P_t(r) = P(R_{t+1} = r \mid \mathcal{F}_t)$ for $r \in \{0, 1\}$. Then

$$\mathbb{E}\big[-\log \hat{P}_\theta(R_{t+1} \mid \mathcal{F}_t)\big] = \mathbb{E}\Big[\sum_{r \in \{0,1\}} P_t(r)\big(-\log \hat{P}_\theta(r \mid \mathcal{F}_t)\big)\Big]. \tag{A.4}$$

Adding and subtracting $\log P_t(r)$ inside the sum yields

$$\mathbb{E}\Big[\sum_r P_t(r)(-\log P_t(r))\Big] + \mathbb{E}\Big[\sum_r P_t(r) \log \frac{P_t(r)}{\hat{P}_\theta(r \mid \mathcal{F}_t)}\Big]. \tag{A.5}$$

The first term in (A.5) is $H(R_{t+1} \mid \mathcal{F}_t)$ and the second term is $\mathbb{E}[\mathrm{KL}(P_t \| \hat{P}_\theta)]$, which proves (A.3). □

Fact A.2 (Binary Bayes error lower bound)**.** *Let $P_e^*(t)$ be the Bayes* 0-1 *error for predicting R_{t+1} from $\mathcal{F}_t$. Then*

$$P_e^*(t) \geq h_2^{-1}\big(H(R_{t+1} \mid \mathcal{F}_t)\big), \tag{A.6}$$

where $h_2(p) = -p\log_2 p - (1-p)\log_2(1-p)$ and h_2^{-1} denotes its inverse on $[0, 1/2]$.

Proof. For binary targets, the conditional form of Fano's inequality gives

$$H(R_{t+1} \mid \mathcal{F}_t) \leq h_2(P_e^*(t)). \tag{A.7}$$

Because $P_e^*(t) \in [0, 1/2]$ and h_2 is strictly increasing on that interval by (A.2), applying h_2^{-1} to (A.7) yields (A.6). □

These facts explain why the appendix uses

$$\hat{u}_{\mathrm{ctw},t} = H_2(\hat{p}_{\mathrm{ctw},t}) \tag{A.8}$$

as the operational LIU proxy. The quantity in (A.8) depends only on the one-step-ahead Bernoulli law implied by the causal anchor, it is invariant to threshold choices, and it is directly aligned with the cross-entropy decomposition in (A.3).

A.2 CTW algorithm details and cold-start behavior

The real-data implementation conditions binary response predictions on past item-response tokens and a current-item query token, as described in Section 3.2. Item identity is therefore part of its context. For clarity, the equations and Algorithm 1 below review classical binary-source CTW on $R = (r_1, \ldots, r_T)$; they are an explanatory special case, not a specification of the item-token context builder used for the real benchmarks. Exact reconstruction of that builder requires the implementation and its input artifacts. Let D be the maximum context depth and let λ denote the empty context. For each binary context s with $|s| \leq D$, let r^s be the subsequence of responses whose preceding suffix equals s. The Krichevsky-Trofimov block probability at node s is

$$P_e^s(r^s) = \prod_{j=1}^{|r^s|} \frac{n_{s,j-1}^{r_j^s} + \frac{1}{2}}{n_{s,j-1}^0 + n_{s,j-1}^1 + 1}, \tag{A.9}$$

where $n_{s,j-1}^a$ counts how many times symbol $a \in \{0, 1\}$ has appeared in context s before the j-th symbol of r^s.

The CTW mixture recursively combines the local KT model with the two child contexts:

$$P_w^s(r^s) = \begin{cases} P_e^s(r^s), & |s| = D, \\ \frac{1}{2}P_e^s(r^s) + \frac{1}{2}P_w^{0s}(r^{0s})P_w^{1s}(r^{1s}), & |s| < D. \end{cases} \tag{A.10}$$

The one-step predictive probability at the root is then

$$\hat{p}_{\mathrm{ctw},t} = \Pr_{\mathrm{CTW}}(R_{t+1} = 1 \mid r_{1:t}) = \frac{P_w^\lambda(r_{1:t}1)}{P_w^\lambda(r_{1:t}0) + P_w^\lambda(r_{1:t}1)}. \tag{A.11}$$

Cold-start behavior. At the beginning of a student sequence, most contexts are unseen. In that regime, the KT prior in (A.9) keeps $\hat{p}_{\mathrm{ctw},t}$ close to 0.5, so $\hat{u}_{\mathrm{ctw},t}$ in (A.8) stays close to one bit rather than becoming spuriously overconfident. As more evidence accumulates, longer contexts contribute through (A.10). For stationary finite-depth tree sources, the classical CTW per-symbol redundancy decays as $O(\log t/t)$ for a fixed source class [Willems et al., 1995], so the cold-start penalty vanishes with sequence length. In this paper we use that result only as a property of the reference source model: student learning itself is not assumed stationary, but the reference guarantee clarifies why CTW is a stable causal anchor.

Algorithm 1 Classical binary-source CTW: an illustrative special case

Input: Binary response stream $R = (r_1, \ldots, r_T)$, maximum depth D
Initialize a depth-D context tree with KT priors at every node
for $t = 0$ **to** $T - 1$ **do**
    Traverse the suffix path induced by the observed prefix $r_{1:t}$
    Compute the root mixture probability $\hat{p}_{\mathrm{ctw},t}$ by the CTW recursion in (A.10) and the predictive ratio in (A.11)
    Convert the anchor probability into uncertainty: $\hat{u}_{\mathrm{ctw},t} \leftarrow H_2(\hat{p}_{\mathrm{ctw},t})$
    Observe r_{t+1} and update the node statistics on the same suffix path
end for
Output: Anchor probabilities $\{\hat{p}_{\mathrm{ctw},t}\}_{t=0}^{T-1}$ and uncertainty profile $\{\hat{u}_{\mathrm{ctw},t}\}_{t=0}^{T-1}$

A.3 Mutual-information decomposition into RU and IG

For an item q, define the item-marginal entropy and the oracle conditional entropy by

$$H_q^{\mathrm{marg}} := H(R_{t+1} \mid q_{t+1} = q), \qquad H_q^{\mathrm{cond}} := H(R_{t+1} \mid X_t, q_{t+1} = q). \tag{A.12}$$

The theoretical remaining uncertainty and information gain are

$$\mathrm{RU}^\star(q) := H_q^{\mathrm{cond}}, \qquad \mathrm{IG}^\star(q) := H_q^{\mathrm{marg}} - H_q^{\mathrm{cond}}. \tag{A.13}$$

Fact A.3 (RU/IG identity)**.** *For every item q,*

$$H_q^{\mathrm{marg}} = \mathrm{RU}^\star(q) + \mathrm{IG}^\star(q), \qquad \mathrm{IG}^\star(q) = I(R_{t+1}; X_t \mid q_{t+1} = q) \geq 0. \tag{A.14}$$

Proof. By the definition of conditional mutual information,

$$I(R_{t+1}; X_t \mid q_{t+1} = q) = H(R_{t+1} \mid q_{t+1} = q) - H(R_{t+1} \mid X_t, q_{t+1} = q). \tag{A.15}$$

Substituting (A.12) into (A.15) gives the second identity in (A.14); the first follows by rearrangement, and non-negativity is a standard property of mutual information. □

The paper uses CTW to obtain operational plug-in diagnostics:

$$\mathrm{RU}(q) := H_q^{\mathrm{ctw}}, \qquad \mathrm{IG}(q) := H_q^{\mathrm{marg}} - H_q^{\mathrm{ctw}}. \tag{A.16}$$

Because (A.16) replaces the oracle conditional entropy with an estimated one, negative $\mathrm{IG}(q)$ values may appear because of finite-sample estimation or anchor misspecification; the non-negativity guarantee applies to the oracle mutual information, not to this plug-in diagnostic.

Venn-style interpretation. It is useful to read (A.14) as a verbal Venn diagram. The full circle is the item-marginal entropy H_q^{marg}. The portion explained by student history is $\mathrm{IG}^\star(q)$, while the unexplained remainder is $\mathrm{RU}^\star(q)$. High IG therefore means that historical context removes a large fraction of the original uncertainty; high RU means that substantial uncertainty remains even after conditioning on the observable history.

Interpretation guide. If RU is high and IG is low, the item is a warning case: the item-conditioned anchor leaves high uncertainty relative to item-marginal entropy. This motivates checking guessing, item ambiguity, label quality, missing variables, and anchor misspecification; it does not prove intrinsic randomness. If IG is high, the benchmark still contains exploitable contextual structure, so stronger temporal, prerequisite, or item-context modeling is warranted.

B Data and Preprocessing Details

B.1 Dataset coverage

Table 3 summarizes the dataset coverage used in the appendix. The highest-band mass refers to the CTW-entropy band $[0.8, 1.0]$ used in the main benchmark-diagnosis figures.

Table 3: Dataset coverage used by the appendix. Support-interaction counts are reported for NIPS Task 3/4 and Algebra 2005.

Dataset	Protocol	Support interactions	Aligned evaluation points	Highest-band mass
NIPS Task 3/4	question-level	898,674	1,354,735	74.8%
Algebra 2005	question-level	598,954	464,725	55.6%
ASSIST2015	$concept = question$	–	672,960 for CTW+NN/KT; 673,200 for CTW before KT-side intersection	–

Table 4: Implementation settings used by the appendix.

Component	Setting
CTW anchor	Item-response token stream (past `I:item|R:resp` tokens plus the current `Q:item` query token), maximum context depth $D = 6$, KT priors, causal one-step-ahead prediction. The token budget includes one query token and at most $D - 1 = 5$ historical interaction tokens (effective response depth 5).
Uncertainty bands	Five fixed CTW-entropy bands: $[0.0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, $[0.8, 1.0]$.
CTW+NN probe	Fused logit correction $\sigma(\mathrm{logit}(\hat{p}_{\mathrm{ctw},t}) + f_\phi(X_{\leq t}, Q_{t+1}))$ used only as an anchor-side neural probe.
Synthetic sanity check	64 trials, sequence length 400, one-step warmup, CTW depth 6.
Item diagnostics	Minimum support 50 per item; diagnostic guide lines at $RU = 0.8$ and $IG = 0.1$.

B.2 Preprocessing protocol and alignment rules

The main benchmark-diagnosis figures use question-level preprocessing for NIPS Task 3/4 and Algebra 2005. ASSIST2015 is reported as an auxiliary consistency check under the convention $concept = question$, which makes the prediction unit compatible with the question-level protocol used by the two primary benchmarks. All reported comparisons use five-fold aligned held-out points: CTW-side anchor probabilities and task-facing KT predictions are joined only when they share the same fold, student/history position, item identifier, and held-out response.

The diagnostic stage operates on binary correctness labels and does not introduce a separate appendix-specific prediction task beyond the benchmark sequence protocol used by the main paper.

B.3 Knowledge-component mapping and Q-matrix status

The CTW anchor itself does not require a Q-matrix. It uses historical item-response tokens and the next queried item's token as predictor inputs, in addition to using item identifiers for prediction alignment and item-level diagnostics. Prediction precedes the current-response update. The appendix does not introduce an additional inferred KC mapping layer.

C Implementation and Reproducibility

C.1 Implementation settings

Table 4 collects the implementation settings used by the appendix.

C.2 Training and evaluation details

All global metrics and local diagnostic summaries are reported as mean $\pm$ standard deviation over five folds. The CTW anchor and the derived bucketization are always computed out-of-fold. The shared aligned subset contains 1,354,735 points on NIPS Task 3/4, 464,725 points on Algebra 2005, and 672,960 points on ASSIST2015 under the convention $concept = question$.

Table 5: Recovered fold-specific hyperparameter summaries for representative KT baselines in the archived run exports. Each cell reports the values appearing across the five retained folds for that dataset. Abbreviations: d for hidden width, d_{ff} for feed-forward width, heads for attention heads, B for transformer blocks, and mem for memory size.

Model	NIPS Task 3/4	Algebra 2005	ASSIST2015
DKT	emb 256, drop $0.3/0.5$, lr 10^{-3}	emb 256, drop $0.1/0.3/0.5$, lr 10^{-3}	emb 256, drop $0.05/0.3/0.5$, lr 10^{-3}
AKT	$d = 64$, $d_{ff} = 64/256$, heads $4/8$, $B = 2/4$, drop $0.1/0.3$, lr 10^{-3}	$d = 64/256$, $d_{ff} = 256$, heads 4, $B = 4$, drop $0.05/0.1$, lr 10^{-4}	$d = 64/256$, $d_{ff} = 64/256$, heads $4/8$, $B = 4$, drop $0.05/0.1$, lr $10^{-4}/10^{-5}$
simpleKT	$d = 64$, $d_{ff} = 256$, layers 2, $B = 4$, drop 0.3, lr 10^{-3}	$d = 256$, $d_{ff} = 64/256$, layers 2, $B = 4$, drop $0.05/0.1$, lr $10^{-4}/10^{-5}$	$d = 256$, $d_{ff} = 256$, layers 2, $B = 4$, drop $0.05/0.1$, lr $10^{-4}/10^{-5}$
stableKT	$d = 64$, $d_{ff} = 64/256$, layers 2, heads 4, drop 0.3, lr 10^{-3}, $r = 0.1/0.7/0.9/1.0$, $\gamma = 0.7/0.9/1.0/5.0$	$d = 256$, $d_{ff} = 64/256$, layers 2, heads 4, drop $0.05/0.1$, lr $10^{-4}/10^{-5}$, $r = 0.5/0.9/5.0$, $\gamma = 0.1/0.3/0.5/5.0$	$d = 256$, $d_{ff} = 64/256$, layers 2, heads 4, drop $0.05/0.1$, lr 10^{-5}, $r = 0.5/1.0/2.0$, $\gamma = 0.1/0.5/0.7$
Deep-IRT	$\dim_s = 256$, mem $32/64$, drop $0.05/0.3$, lr 10^{-3}	$\dim_s = 256$, mem 32, drop $0.05/0.1$, lr 10^{-3}	$\dim_s = 64$, mem 32, drop $0.05/0.1$, lr 10^{-3}
ATKT	skill 64, ans $64/256$, hidden 256, attn $64/256$, drop $0.05/0.1$, lr $10^{-4}/10^{-3}$	skill 64, ans 64, hidden 256, attn 256, drop $0.05/0.1/0.5$, lr 10^{-3}	skill 64, ans $64/256$, hidden 256, attn $64/256$, drop 0.05, lr $10^{-4}/10^{-3}$

The archived run exports used for the retained KT comparison table preserve fold-specific hyperparameter fields for each task-facing baseline. Table 5 summarizes representative recovered settings for the main baselines discussed in the paper. Each entry reports the set of values that appears across the five retained folds for that dataset, so multi-valued cells indicate genuine fold-level variation rather than post hoc averaging.

The same archive also preserves explicit training-length metadata for DIMKT on the two main benchmark datasets: batch size 64, num_steps 199, learning rate 10^{-3}, and weight decay 10^{-4} on both NIPS Task 3/4 and Algebra 2005, with dropout varying over $\{0.05, 0.1, 0.3\}$. We did not find optimizer, early-stopping, or wall-clock fields in the retained run exports, so those quantities are not supplied for the original baseline branch. The separately reported CTW+NN-SKT branch has validation-NLL selection and early-stopping settings described in Appendix E.

C.3 Compute-cost discussion

The CTW anchor is intentionally lightweight. For a fixed depth D, each prediction/update step touches only the suffix path of length at most D, so the per-step symbolic overhead is linear in D and small in practice. In our experiments we use $D = 6$, which is sufficient to separate the i.i.d. and persistent synthetic regimes while keeping the causal anchor inexpensive. This is one of the main practical advantages of the framework: the uncertainty analysis adds a modest symbolic pass on top of predictions that are already being generated for KT evaluation, whereas the expensive part of the pipeline remains the training of the task-facing neural KT models.

D Additional Experimental Results

D.1 Cross-benchmark diagnostic differences

Table 6 summarizes the strongest cross-benchmark contrasts reported in the paper. The main point is not only that the high-entropy region exists, but that it carries different mass and different residual headroom across datasets.

The table reinforces the same diagnosis from two angles. On NIPS Task 3/4, the highest-entropy region dominates the benchmark mass and still admits the largest accuracy gain over CTW, showing residual predictive structure relative to this anchor in its difficult regime. On Algebra 2005, the hard region is smaller but still substantial, and the largest proper-score improvement also appears there. Thus RU/IG and band-level gains should be read together: high entropy can still hide exploitable structure, but the gain pattern differs by dataset.

Table 6: Cross-benchmark diagnostic contrasts. Positive ΔACC means CTW+NN improves over CTW; negative ΔBrier means the proper score improves.

Dataset	Highest CTW band mass	ΔACC in highest CTW band	ΔBrier in highest CTW band	Refined-coordinate note
NIPS Task 3/4	74.8%	$+0.0886$	-0.0378	Under refined repartitioning, the peak CTW+NN gain shifts to the upper-middle bucket at $+0.0868$.
Algebra 2005	55.6%	$+0.0283$	-0.0681	Under refined repartitioning, the largest CTW+NN gain appears in the fourth bucket at $+0.0335$, while gains remain positive in the highest bucket.

E Robustness to Anchor Information, Band Resolution, and Cold Start

E.1 Analysis branches and reporting conventions

The original reported Table 2 branch is identified as a single SimpleKT residual export, `simplekt_residual_ctw_local_bayes`, with the aligned subsets specified in the main text. The additional analysis in this appendix uses a single SimpleKT residual probe, CTW+NN-SKT, and its common evaluation cohort. These are separate analysis branches: differences between their point estimates should not be interpreted as reruns of an identical predictor or cohort. Within each additional comparison, evaluated prediction rows are held fixed across the reference and probe. Equal-frequency CTW quintiles contain approximately $271{,}804$ NIPS predictions or $93{,}171$ Algebra predictions per quintile. These approximate counts are not a replacement for the original aligned-subset totals.

All deltas are model minus reference on matched predictions. Positive ΔACC and negative ΔBrier or ΔLog-loss indicate improvement. Additional log-loss values use natural logarithms and are reported in nats, whereas uncertainty coordinates are in bits. Dividing log-loss by $\ln 2$ converts it to bits without changing comparisons. The theoretical decomposition applies to log-loss; Brier, accuracy, and calibration are complementary empirical checks.

E.2 Independent item-aware causal references

ItemMean is an out-of-fold empirical-Bayes reference. With c_q correct responses and n_q observations for item q in the four training folds, training-fold global correctness $\bar{p}$, and prior strength α, it predicts

$$\hat{p}_{\mathrm{item}}(q) = \frac{c_q + \alpha\bar{p}}{n_q + \alpha}. \tag{E.1}$$

The default $\alpha = 20$ supplies 20 global-rate pseudo-observations; unseen items fall back to $\bar{p}$. Held-out labels are excluded from these item counts.

ItemRasch adds a one-dimensional causal online student state to the item logit. It is a 1PL-style online reference, not a jointly fitted IRT model. For each sequence sorted by prediction position, initialize $\theta_0 = 0$ and precision $\lambda_0 = 1$. Before observing response r_t, compute

$$b_t = \mathrm{logit}(\hat{p}_{\mathrm{item}}(q_t)), \qquad p_t = \sigma(b_t + \theta_{t-1}). \tag{E.2}$$

After observing the response, update

$$\lambda_t = \lambda_{t-1} + \max\{p_t(1-p_t), 10^{-6}\}, \qquad \theta_t = \mathrm{clip}\left(\theta_{t-1} + \frac{r_t - p_t}{\lambda_t}, -4, 4\right). \tag{E.3}$$

The item-prior probability is clipped to $[10^{-12}, 1 - 10^{-12}]$ before taking its logit. The argument passed to the sigmoid is clipped to $[-30, 30]$; there is no additional 10^{-12} clamp after the sigmoid, whose extreme outputs are approximately 9.36×10^{-14} and $1 - 9.36 \times 10^{-14}$. The reference has no neural parameters, offline student-ability fitting, optimizer, batch size, epochs, or validation-based selection. Its effective update step is $1/\lambda_t$.

Table 7: Independent reference scores on the additional common cohort. Lower is better; log-loss is in nats.

Dataset	Reference	Brier	Log-loss
NIPS Task 3/4	CTW	0.2111	0.6096
	ItemRasch	0.1923	0.5649
Algebra 2005	CTW	0.1684	0.5016
	ItemRasch	0.1585	0.4853

E.3 Highest-band gains and uncertainty estimates

Table 8 reports five-fold mean gains with 95% t intervals. The ItemRasch rows use ItemRasch entropy to define the band and ItemRasch as the reference; they therefore test localization under an alternative

Table 8: CTW+NN-SKT gains in the highest-entropy band, $[0.8, 1.0]$, on each reference's own coordinate. Entries are five-fold means [95% t intervals]; log-loss is in nats.

Dataset	Reference	ΔACC	ΔBrier	ΔLog-loss
NIPS	CTW	$+0.0905$	-0.0388	-0.0902
		$[0.0898, 0.0913]$	$[-0.0389, -0.0386]$	$[-0.0906, -0.0899]$
Algebra	CTW	$+0.0285$	-0.0685	-0.1578
		$[0.0252, 0.0317]$	$[-0.0701, -0.0669]$	$[-0.1624, -0.1533]$
NIPS	ItemRasch	$+0.0420$	-0.0175	-0.0394
		$[0.0411, 0.0429]$	$[-0.0178, -0.0172]$	$[-0.0400, -0.0388]$
Algebra	ItemRasch	$+0.0788$	-0.0430	-0.1009
		$[0.0765, 0.0812]$	$[-0.0450, -0.0411]$	$[-0.1065, -0.0952]$

item-and-ability reference, rather than adding item identity to a response-only CTW reference. AKT, simpleKT, and stableKT also improve accuracy, Brier, and log-loss in the highest ItemRasch-entropy band on both datasets.

For the CTW coordinate, two-sided paired tests over folds compare the highest and lowest bands. For ΔACC, $p = 1.16 \times 10^{-8}$ on NIPS and $p = 1.71 \times 10^{-5}$ on Algebra; for ΔBrier, $p = 2.83 \times 10^{-6}$ and $p = 2.92 \times 10^{-8}$; for ΔLog-loss, $p = 7.77 \times 10^{-4}$ and $p = 2.89 \times 10^{-8}$. Under the ItemRasch coordinate, high-versus-low differences are significant for all three metrics, with $p \leq 4.05 \times 10^{-5}$ on NIPS and $p \leq 4.78 \times 10^{-6}$ on Algebra. These are fold-level descriptive comparisons rather than guarantees for new populations; the folds share training data.

Within-band expected calibration error (ECE) uses 15 fixed probability bins over $[0, 1]$. In the highest CTW-entropy band, CTW+NN-SKT has ECE 0.0058 on NIPS and 0.0264 on Algebra. The corresponding ECE changes are -0.0043 $[-0.0056, -0.0030]$ and -0.1496 $[-0.1565, -0.1428]$. ECE depends on binning and is interpreted alongside Brier and log-loss.

E.4 Equal-frequency resolution check

Table 9 resolves the large fixed-width high-entropy population with balanced CTW quintiles. On NIPS, accuracy gain increases across quintiles; on Algebra, accuracy gains concentrate in the upper two quintiles, while the largest log-loss reduction appears in the hardest quintile. Thus the evidence supports non-uniform localization rather than a universal monotonic accuracy curve.

Table 9: CTW+NN-SKT minus CTW on equal-frequency CTW-entropy quintiles. Q1 has the lowest entropy and Q5 the highest; log-loss is in nats.

Dataset	Metric	Q1	Q2	Q3	Q4	Q5
NIPS	ΔACC	$+0.0094$	$+0.0301$	$+0.0637$	$+0.0926$	$+0.1570$
	ΔLog-loss	-0.0665	-0.0705	-0.0806	-0.0816	-0.1227
Algebra	ΔACC	-0.0001	$+0.0011$	$+0.0157$	$+0.0433$	$+0.0206$
	ΔLog-loss	-0.0087	-0.0176	-0.0404	-0.1443	-0.2590

Using independent ItemRasch quintiles gives the same directional check: CTW+NN-SKT's accuracy gain over ItemRasch rises from $+0.0034$ to $+0.0853$ across NIPS quintiles and from $+0.0012$ to $+0.1036$ across Algebra quintiles. Brier and log-loss improve in every ItemRasch quintile.

E.5 Reference-parameter sensitivity

We vary item prior strength over $\{1, 5, 20, 100, 200\}$ and initial Rasch precision over $\{0.25, 0.5, 1, 2, 4\}$. Across these 25 configurations, CTW+NN-SKT remains favorable over ItemRasch in the highest ItemRasch-entropy band. On NIPS, ΔACC ranges from $+0.0410$ to $+0.0449$, ΔBrier from -0.0191 to -0.0174, and ΔLog-loss from -0.0432 to -0.0391. On Algebra, the corresponding ranges are $+0.0747$ to $+0.0990$, -0.0553 to -0.0411, and -0.1332 to -0.0952. This sensitivity check addresses two reference settings, not every possible uncertainty estimator.

E.6 Interaction position, item support, and RU/IG regions

The additional common cohort's highest CTW-entropy band contains $1{,}015{,}423$ NIPS predictions and $258{,}740$ Algebra predictions. Only 6.61% and 2.38%, respectively, are in the first 20 interactions. Equivalently, 93.39% and 97.62% occur after position 20; median positions are 168 and 600. High ItemRasch entropy is similarly persistent, with 92.85% and 97.81% after position 20. The high-entropy region is therefore not dominated by initial CTW cold start.

Zero training-item support accounts for 1.22% of the NIPS test intersection and 47.33% of the Algebra test intersection. Within the high ItemRasch-entropy band of that slice, CTW+NN-SKT improves over ItemRasch by $+0.1056$ ACC, -0.0353 Brier, and -0.0757 log-loss on NIPS, and by $+0.1165$, -0.0609, and -0.1399 on Algebra. These are support-slice checks, not a full out-of-distribution generalization claim.

For model diagnosis, the additional NIPS high-IG region contains $250{,}315$ predictions. stableKT improves over CTW by $+0.0638$ ACC $[0.0628, 0.0649]$, -0.0325 Brier $[-0.0328, -0.0321]$, and -0.0843 log-loss $[-0.0857, -0.0829]$. This supported example complements the original scatter plots; it does not supply inferential estimates for every plotted cell.

E.7 Single-probe architecture and training protocol

CTW+NN-SKT adds a SimpleKT residual to the CTW logit. The strict checkpoints use 2 or 4 SimpleKT blocks, four attention heads, and feed-forward width 256. Model selection uses validation negative log-likelihood, with at most 50 epochs and early-stopping patience 4. NIPS settings use hidden width 64 or 128, dropout 0.10 or 0.20, learning rate 5×10^{-4} or 10^{-3}, and batch size 64 or 128. Algebra settings use hidden width 128 or 256, dropout 0.05, 0.10, or 0.20, learning rate 10^{-4}, and batch size 64 or 128.

These reported ranges document the additional branch but do not identify every selected dataset-fold configuration. Complete selected configurations, seeds, best epochs, validation NLL values, and optimizer metadata are required for exact retraining; they are not available in the provided archive. The archived aggregate results and the reported additional comparisons support the diagnostic analysis, while exact retraining remains a reproducibility limitation.

Data and code availability

The study uses the existing educational benchmarks cited in the paper. Residual-probe training entry points are available in the supplementary material of the accepted OpenReview record at `https://openreview.net/forum?id=qISxwVTG3L`. Benchmark access remains subject to the original providers' terms.

Responsible use and writing assistance

This study analyzes existing educational benchmarks and does not collect new human-subject data. The diagnostic scores should guide benchmark inspection and model evaluation; they should not be used alone to make decisions about individual students. High anchor entropy can reflect missing variables or model mismatch rather than an intrinsic limitation of a learner. Data access and reuse remain subject to each original provider's terms. LLM tools were used for writing, editing, and formatting assistance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the uncertainty-conditioned evaluation objective and the benchmark/model diagnosis scope, and these are matched by the reported experiments and analyses.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: A dedicated "Discussion and Limitations" section states the main scope boundaries, including uncertainty-coordinate choice, dataset-specific coverage differences, and the RU/IG-region-level granularity of the model-diagnosis view.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: Appendix A now states the assumptions and gives complete proofs for the entropy-risk decomposition, the Bayes-error lower bound, and the RU/IG decomposition used by the paper.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No].

   Justification: The paper documents protocols, reference formulas, reported hyperparameter ranges, and archived settings. Exact retraining of the additional branch requires selected per-fold configurations, seeds, and checkpoints not present in the available archive.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: Implementation code is available through the accepted OpenReview supplementary material, but complete instructions and artifacts for faithfully reproducing every additional robustness result are not available.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No].

   Justification: The retained baseline settings and the additional probe search ranges, validation metric, and early-stopping rule are documented. Complete per-fold selections and optimizer metadata for the additional branch remain unavailable.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

Justification: The original comparison reports five-fold mean and standard deviation. Appendix E separately reports five-fold confidence intervals and paired comparisons for the additional branch, with a limitation concerning overlapping training folds.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: The original checklist records two Intel Xeon E5-2683 v4 CPUs and two NVIDIA Tesla V100 SXM2 32GB GPUs. Complete execution times and memory consumption by experiment were not retained.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is benchmark/model diagnosis on existing educational datasets and does not involve manipulative deployment or harmful experimental procedures.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The responsible-use discussion identifies benchmark auditing benefits and the risk of interpreting uncertainty as an intrinsic learner limitation or using it alone for individual student decisions.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release high-risk generative models or newly scraped high-risk datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No].

    Justification: Benchmark and toolkit sources are cited, but the available materials do not establish complete license and terms compliance for every reused asset. The paper license is not treated as a license for all datasets.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not claim release of new public datasets, codebases, or model assets as a core contribution.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

Justification: The study uses existing benchmark logs and does not conduct crowdsourcing or new human-subject data collection.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: No new human-subject study is performed; the work analyzes pre-existing benchmark data.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs are not part of the core methodology or experimental pipeline described in this paper.