UNDERSTANDING SEMANTIC GENERALIZATION IN KNOWLEDGE TRACING THROUGH DIFFICULTY SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge tracing (KT) predicts students' responses from interaction histories, but new questions have no training responses. Knowledge components (KCs) can be shared across questions, while pretrained question semantics add limited or mixed gains beyond them. A privileged-information experiment shows that empirical item difficulty can be highly predictive. However, measuring empirical difficulty requires student responses, which are unavailable for newly introduced questions. We therefore study whether difficulty can link response patterns from training questions to the content of unseen ones. To study this transfer, we adapt a pretrained semantic encoder with discrete difficulty classification and continuous correctness regression, then use its frozen representations for KT. We also test generated solutions as an augmentation to question content. Across three datasets and six KT backbones, adaptation improves unseen-item prediction, with controls linking the gains to content–difficulty correspondence. Probes find more accessible correctness information, while exposure experiments show that the readout advantage diminishes as item-specific responses accumulate. Together, these results show how difficulty supervision transfers information from observed student responses through question content to unseen items.

1 INTRODUCTION

Knowledge tracing (KT) models a student's evolving knowledge from past interactions to predict future responses, supporting adaptive practice and personalized learning (Corbett and Anderson, 1994; Piech et al., 2015). Item representations have developed from discrete identifiers learned through response sequences (Zhang et al., 2017; Ghosh et al., 2020) to exercise-text encoders and pretrained semantic embeddings (Su et al., 2018; Fu et al., 2024). This development brings question content into models traditionally driven by response histories: ID embeddings represent particular training items, whereas semantic encoders express properties shared across different questions.

Generalization becomes essential when educational platforms expand or refresh their question banks. New exercises must be recommended and their responses interpreted before sufficient item-specific data accumulate. Similar challenges arise in portable student modeling across platforms with different item inventories. We study unseen-item generalization within each dataset: evaluation items are excluded from KT training, while their content and knowledge components (KCs) remain available. Semantics offer a bridge because text can describe a new question before any student answers it, including differences between questions involving the same KC. KC-based prediction provides a baseline for evaluating this transfer. Inductive KT explores this opportunity (Fu et al., 2024).

Prior work connects content and difficulty through difficulty-level representations and retrieval-assisted estimation for new items (Liu et al., 2024; Ma et al., 2026). We study how this connection transfers response experience. Source responses already train KT to predict individual students' answers from their histories. Aggregating these responses into item-level difficulty targets additionally supervises the relation between question content and population performance. A new question can then access information learned from those responses through its content. We hypothesize that this supervision makes semantic representations more sensitive to performance-related differences between questions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

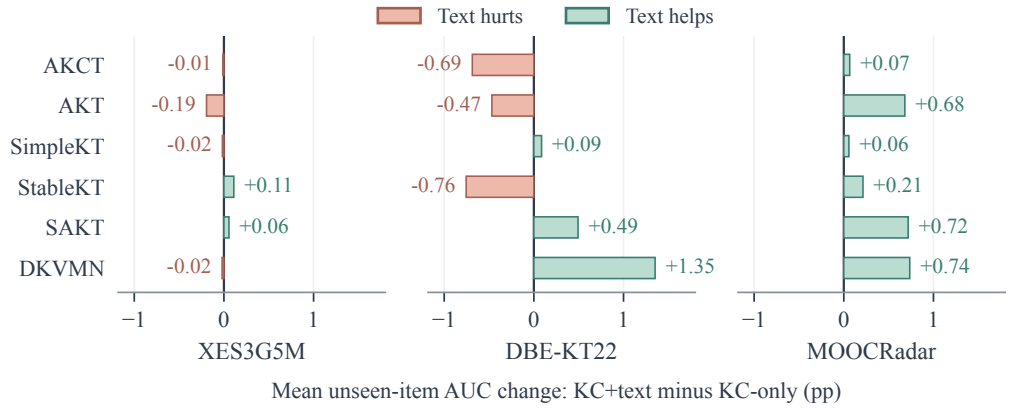

Figure 1: **Adding pretrained text gives limited or mixed unseen-item gains.** Bars show the change in mean unseen-item AUC for KC+text relative to KC-only (percentage points). Full scores and uncertainty: Table 2 and Appendix G.

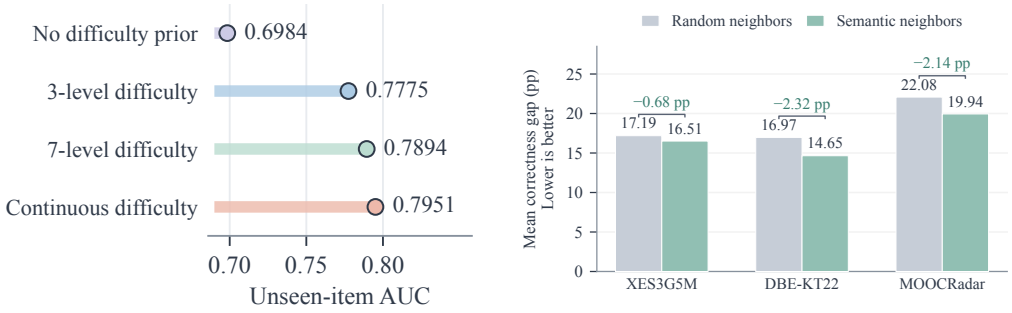

Figure 2: Left: privileged difficulty improves XES3G5M prediction using extra held-out-item responses. Right: semantic neighbors have smaller mean correctness gaps than random neighbors; brackets show reductions in percentage points. Protocols: Appendices B and H.2.

Our central question is: ***how do response experience and item semantics provide difficulty information for KT, and when can this information transfer to unseen items?***

Pretrained semantics contain information related to student performance, although their gains in unseen-item KT remain limited or mixed (Figure 1). In Figure 2 (right), we compare five semantic neighbors for each unseen question with random neighbors from the same eligible training-item pool. Semantic neighbors have modestly smaller mean correctness gaps on XES3G5M, DBE-KT22, and MOOCRadar. This preliminary observation suggests that question content already captures some performance-related structure. We study how response-derived supervision can make this information more useful for KT.

Empirical difficulty provides a direct connection to student responses. Figure 2 (left) shows a privileged-information experiment on XES3G5M: unseen-item AUC rises from 0.6984 to 0.7951 with continuous difficulty and to 0.7775 with three difficulty levels. Difficulty summarizes how a learner population responds to a question (Ghosh et al., 2020; Liu et al., 2024). These priors use extra responses, including held-out items, and are unavailable when a new question is introduced. Their predictive value motivates transferring difficulty information from training questions through semantic content.

We use difficulty-supervised encoder adaptation as an intervention to study this transfer. Across three datasets and six backbones, the resulting representations improve unseen-item prediction. Equal-budget label controls link this benefit to correct content–difficulty correspondence. Controlled exposure on XES3G5M further shows that an initial advantage in correctness readout diminishes as

the same items acquire training responses. Representation probes examine the accessibility of this information across datasets, and generated solutions provide a complementary content-enrichment comparison.

Our contributions are:

- **The role and limitations of question semantics.** We characterize how pretrained question semantics contribute to unseen-item generalization in knowledge tracing.
- **Difficulty connects response experience to semantic transfer.** We show how response-derived difficulty supervision transfers training experience through question content, improving unseen-item prediction across three datasets and six KT backbones.
- **How KT represents and uses difficulty information.** We investigate content–difficulty correspondence and the interaction between transferred difficulty information and item-specific response experience through representation analyses and controlled interventions.

2 PRELIMINARIES

For student u, let $\mathcal{H}_{u,t-1} = \{(q_i, \mathcal{K}_{q_i}, r_i)\}_{i<t}$ denote their preceding interactions, where q_i is an item, $\mathcal{K}_{q_i}$ its knowledge components (KCs), and $r_i \in \{0, 1\}$ the response correctness. Given query content x_{q_t}, KT estimates

$$\hat{y}_{u,t} = f_\theta(\mathcal{H}_{u,t-1}, \mathcal{K}_{q_t}, z_{q_t}), \qquad z_q = E_\phi(x_q). \tag{1}$$

We partition items into training (Source) items $\mathcal{Q}_{\text{src}}$, an unseen-item validation set, and an unseen-item test set $\mathcal{Q}_{\text{unseen}}$. Students are independently split into training, validation, and test populations. *Seen items* belong to the Source training set; *unseen items* are held out from KT training. Unseen items appear only as final evaluation queries, with their interactions removed from training sequences and evaluation histories. Both seen- and unseen-item test queries come from the test-student population. Base content x_q contains the question, existing KC text, and reference answer. The student's current response is masked.

We use the same colored condition labels throughout: **K** is the **K**C-only input, **T** is question **T**ext encoded by a pretrained language model, **S** is generated **S**olution content, and **D** is **D**ifficulty-supervised adaptation of the text encoder. Combined labels add these components while retaining KC inputs and response histories.

We evaluate response prediction and the accessibility of item correctness in learned representations. Difficulty supervision uses a group-smoothed empirical easiness target m_q (Section 3.1); probes use unsmoothed pooled correctness a_q (Section 4.5). Both characterize the observed learner population. Larger values indicate easier items.

3 METHOD

Figure 3 connects the KT prediction process (left) to difficulty supervision of the semantic encoder (right). A shared prediction head jointly classifies discrete difficulty levels and regresses continuous correctness using aggregated Source-item responses. After encoder adaptation, we discard this head and freeze the exported vectors for KT training. The fixed downstream interface lets us compare the same content before and after adaptation, with KC inputs available throughout.

3.1 DIFFICULTY SUPERVISION

We rank training students by their correctness on Source items and partition them into five equally sized groups. Let c_{qg} and n_{qg} be correct and total attempts on item q in group g, and let π_g be that group's correctness over Source items. The target is

$$p_{qg} = \frac{c_{qg} + 10\pi_g}{n_{qg} + 10}, \qquad m_q = \frac{1}{5}\sum_{g=1}^{5} p_{qg}. \tag{2}$$

The target is a group-balanced empirical easiness: larger values indicate easier items. Smoothing stabilizes groups with limited response support. We retain supervised items with at least 100 total

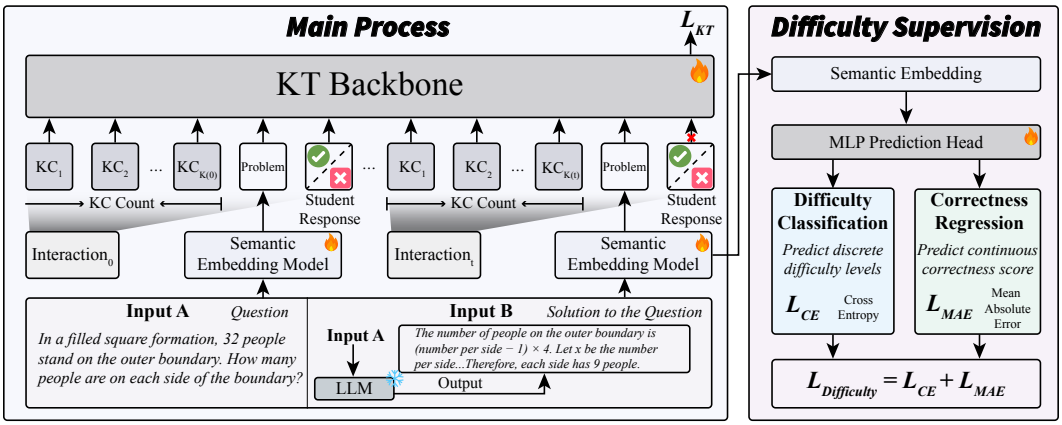

Figure 3: Left: KT uses item content, KCs, and historical responses to predict a masked query response. Right: a shared MLP provides classification and regression outputs, trained jointly with CE and MAE.

attempts and at least 10 per group. Three difficulty levels provide a companion target b_q, ordered from easiest to hardest using dataset-specific thresholds (Appendix C).

We use Qwen3-Embedding-0.6B (Zhang et al., 2025). Its final non-padding token yields an L2-normalized vector $z_q \in \mathbb{R}^{1024}$. A shared $1024 \to 128 \to 4$ MLP with GELU produces three classification logits ℓ_q and a scalar v_q. The objective is

$$\mathcal{L}_D = \frac{1}{|\mathcal{Q}^D_{\text{src}}|} \sum_{q \in \mathcal{Q}^D_{\text{src}}} \left[\text{CE}(\ell_q, b_q) + |\sigma(v_q) - m_q| \right], \tag{3}$$

where $\mathcal{Q}^D_{\text{src}}$ is the eligible Source supervision set. All encoder parameters are fine-tuned. Unseen-validation macro recall selects the epoch, with combined loss breaking ties. We then discard the task head and export frozen vectors for every item. This transfers the representation learned under difficulty supervision to downstream interaction modeling.

3.2 Knowledge Tracing

Each backbone receives KC information, frozen item vectors, and observed historical responses. Final-query binary cross-entropy trains the KT parameters and text projection. We keep the input interface and training budget fixed within each backbone across representation conditions.

Adaptive KC Transformer (AKCT). AKCT provides a standard Transformer encoder–decoder interface (Vaswani et al., 2017) that retains each annotated KC as a separate token, alongside text and response inputs (Appendix D). Within-item attention contextualizes these inputs; the decoder attends to the encoded history and predicts the masked query response. This interface accommodates variable KC counts while preserving the input information across experimental conditions.

Other backbones. We also use AKT (Ghosh et al., 2020), SimpleKT (Liu et al., 2023a), StableKT (Li et al., 2024), SAKT (Pandey and Karypis, 2019), and DKVMN (Zhang et al., 2017). Their public model cores are adapted to the shared KC/content setting, with problem-ID inputs disabled (NoPID). For AKT, the mean KC lookup and a projection of z_q form the item representation; a response embedding supplies the historical interaction stream. These experiments compare representations within each adapted backbone.

3.3 Input Conditions

Generated solutions can make the demands of an item more explicit: which concepts must be combined, which intermediate steps are required, and where a learner may encounter difficulty. Our intuition is that a large language model can draw on its pretrained subject knowledge to articulate

Table 1: Input conditions. **K**: KC-only; **T**: question text; **S**: generated solution; **D**: difficulty-supervised encoder adaptation. All groups receive KC lookup and historical responses; adapted conditions use frozen pre-head vectors.

Condition	Encoder content	Encoder adaptation	Auxiliary KT input
K	—	—	None
T	Base	Original pretrained encoder	Frozen vector
T+S	Base + solution	Original pretrained encoder	Frozen vector
T+D	Base	Source CE + MAE	Frozen adapted vector
T+S+D	Base + solution	Source CE + MAE	Frozen adapted vector

these demands, enriching the cues available to the difficulty-supervised encoder. Generated reasoning enriches the encoder content; empirical difficulty supervision comes from student responses. The teacher receives the question, KC text, and reference answer, while difficulty labels remain hidden. An unavailable-material marker handles invalid solutions. The same solution text is used in the adapted and unadapted conditions.

The four text conditions form a 2×2 design on all three datasets (Table 1). Holding content fixed measures the effect of the full adaptation recipe. Holding adaptation fixed measures the effect of solutions. All vectors have 1024 dimensions and occupy the same input slot within each backbone.

4 Experiments

We examine the transfer of source-item difficulty to unseen-item prediction, the role of content–difficulty correspondence, and changes in correctness readout as response experience accumulates.

4.1 Experimental Setup

We evaluate XES3G5M (Liu et al., 2023b), DBE-KT22 (Abdelrahman et al., 2023), and a 22-course multidisciplinary text subset of MOOCRadar (Yu et al., 2023). Each dataset uses six backbones (AKCT, AKT, SimpleKT, StableKT, SAKT, and DKVMN) and the five input conditions in Table 1. Interfaces omit problem-ID lookup and remain fixed across conditions. MOOCRadar solutions are usable for 1,848 of 2,490 items; the remainder retain an unavailable-material marker. Scalar-head and AKCT readout diagnostics cover all three datasets.

We primarily use locally deployed `Qwen3.6-35B-A3B` to generate solution content **S** (Appendix E). The teacher receives the question, KC text, and reference answer, with empirical difficulty labels withheld.

XES3G5M and DBE-KT22 use a 70/10/20 student split, item fold 0 for Unseen Test, and separately held-out Unseen Validation items. Test histories contain only preceding Source interactions. Training uses 10,000 steps, effective batch size 64, and sequences of at most 200 interactions. Unseen Validation AUC, checked every 500 steps, selects the checkpoint. MOOCRadar follows the same budget and selection rule, with within-course item holdout (Appendix B).

The main comparison uses KT seeds 13, 29, and 47 with frozen banks from upstream seed 13, giving 270 runs. We report means, sample standard deviations, and same-seed paired changes within each dataset/backbone. Unless stated otherwise, tables and error bars use these three KT seeds; upstream diagnostics use the fixed seed-13 encoder. The privileged-prior experiment uses one run.

4.2 Main Results on Three Datasets

In Table 2, difficulty adaptation improves unseen-item AUC over unadapted text in all 54 matched seed pairs (Δ_D). Mean gains are 0.42–1.35 percentage points on XES3G5M, 2.64–4.54 on DBE-KT22, and 1.78–2.62 on MOOCRadar. **T+D** also exceeds **K** in every dataset–backbone mean. The same checkpoints provide AP, LogLoss, and Brier comparisons (Appendix Table 28). Appendix G extends the comparison to seen items and solution ablations.

Table 2: Unseen-item AUC (mean ± standard deviation). Paired gains from **T** are reported separately for difficulty adaptation and solution-enriched adaptation (percentage points).

Backbone	K	T	T+S	T+D	T+S+D	Δ_D	Δ_{S+D}
XES3G5M							
AKCT	.6951±.0054	.6949±.0043	.6980±.0006	.7073±.0014	**.7136±.0006**	1.24±0.49	1.86±0.46
AKT	.7072±.0018	.7053±.0049	.7085±.0043	.7128±.0010	**.7225±.0017**	0.75±0.50	1.72±0.40
SimpleKT	.7052±.0023	.7050±.0019	.7052±.0026	.7092±.0014	**.7189±.0021**	0.42±0.33	1.38±0.24
StableKT	.7091±.0017	.7102±.0015	.7109±.0028	.7155±.0028	**.7246±.0024**	0.53±0.34	1.44±0.36
SAKT	.6667±.0013	.6673±.0015	.6671±.0016	.6808±.0013	**.6898±.0007**	1.35±0.13	2.26±0.10
DKVMN	.6855±.0043	.6853±.0018	.6876±.0051	.6952±.0010	**.7042±.0009**	0.99±0.16	1.89±0.12
DBE-KT22							
AKCT	.6662±.0031	.6594±.0036	.6646±.0125	.7048±.0048	**.7082±.0070**	4.54±0.30	4.88±0.85
AKT	.6773±.0051	.6726±.0051	.6793±.0124	**.6990±.0020**	.6960±.0058	2.64±0.33	2.35±0.18
SimpleKT	.6666±.0034	.6675±.0121	.6714±.0113	**.7011±.0084**	.6844±.0092	3.36±0.54	1.69±1.84
StableKT	.6930±.0171	.6855±.0254	.6849±.0262	**.7128±.0120**	.7054±.0235	2.73±1.39	1.99±1.19
SAKT	.6430±.0072	.6480±.0078	.6600±.0046	**.6808±.0033**	.6765±.0010	3.29±0.88	2.85±0.78
DKVMN	.6616±.0205	.6751±.0140	.6739±.0154	.7015±.0108	**.7045±.0066**	2.64±0.35	2.94±1.20
MOOCRadar							
AKCT	.7948±.0057	.7954±.0058	.8030±.0020	.8168±.0056	**.8251±.0038**	2.13±0.12	2.97±0.29
AKT	.7957±.0030	.8025±.0021	.8057±.0021	.8270±.0040	**.8324±.0032**	2.45±0.45	2.99±0.12
SimpleKT	.7975±.0027	.7981±.0042	.8070±.0021	.8243±.0046	**.8283±.0024**	2.62±0.48	3.02±0.21
StableKT	.8008±.0024	.8029±.0038	.8050±.0056	.8207±.0057	**.8321±.0038**	1.78±0.29	2.92±0.10
SAKT	.7694±.0055	.7766±.0050	.7832±.0052	.7987±.0014	**.8100±.0021**	2.21±0.49	3.34±0.61
DKVMN	.7823±.0026	.7896±.0016	.7934±.0056	**.8128±.0048**	.8113±.0068	2.32±0.52	2.17±0.71

$\Delta_D = (\mathbf{T}+\mathbf{D}) - \mathbf{T}$; $\Delta_{S+D} = (\mathbf{T}+\mathbf{S}+\mathbf{D}) - \mathbf{T}$. **Bold**: best mean in each row. Yellow/blue: better/worse than **T**.

Table 3: Equal-budget label control: unseen-item AUC, mean ± standard deviation. Δ compares correct with shuffled labels in percentage points.

Dataset	Original **T**	Shuffled labels	Correct labels	Δ
XES3G5M	0.6949±0.0043	0.6949±0.0051	**0.7085±0.0020**	+1.36±0.65
DBE-KT22	0.6594±0.0036	0.6580±0.0029	**0.7010±0.0045**	+4.31±0.69
MOOCRadar	0.7954±0.0058	0.7942±0.0045	**0.8083±0.0045**	+1.41±0.14

Δ = correct labels − shuffled labels. **Bold**: best mean per row. Yellow: higher AUC than the shuffled control.

The Δ_{S+D} column also shows gains over text alone for every dataset–backbone mean when solution enrichment is combined with adaptation. Without difficulty adaptation, adding generated solutions raises mean AUC over text alone on five of six XES3G5M backbones, four of six DBE-KT22 backbones, and all six MOOCRadar backbones. With difficulty adaptation, the combined input exceeds T+D on all six XES3G5M backbones, two on DBE-KT22, and five on MOOCRadar. Seen-item gains are less uniform than unseen-item gains. We therefore examine difficulty transfer and representation readout separately below.

4.3 Difficulty Supervision Controls

To test content–difficulty correspondence, we compare adaptation with correct labels against a control that randomly reassigns paired regression and classification targets among Source training items (Appendix H.1). Both arms share initialization, input content, minibatch order, and ten training epochs. AKCT uses the final-epoch frozen banks. The permutation preserves the label distribution and changes the target assigned to each item.

Correct labels improve AUC over shuffled labels on all three datasets (Table 3), linking the benefit to content–difficulty correspondence.

Table 4: Supervision-objective ablation with AKCT: unseen-item AUC gain over unadapted **T** (pp, mean ± standard deviation). Full results: Appendix I.2.

Dataset	Classification	Regression	Joint
XES3G5M	+1.32±0.35	**+1.89±0.50**	+1.36±0.51
DBE-KT22	**+4.26±0.36**	+0.06±0.10	+4.17±0.37
MOOCRadar	**+2.53±0.77**	−0.22±0.20	+1.29±0.28

Gain = objective AUC − matched **T** AUC. **Bold**: largest mean gain per dataset. Yellow/blue: higher/lower AUC than **T**.

We further separate the classification and regression objectives under the same adaptation budget (Table 4). Regression yields the largest gain on XES3G5M, while classification is more effective on DBE-KT22 and MOOCRadar. Joint supervision improves unseen-item AUC on all three datasets. The probe results follow the same ordering among adapted conditions, linking the effectiveness of the supervision objective to the accessibility of correctness information in the resulting representations.

4.4 Controlled Response Exposure

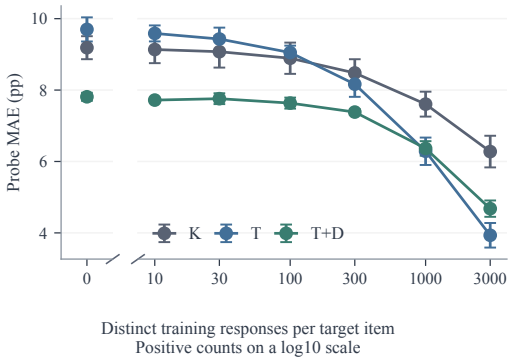

Figure 4: Probe MAE (mean ± standard deviation); zero is separate from the logarithmic axis. Protocol: Appendix H.3.

We study how transferred difficulty information relates to learning from a question's own responses. A linear Ridge probe (Hoerl and Kennard, 1970) predicts pooled item correctness from fixed, empty-history KT states; its MAE measures how readily that information can be read out. For 50 fixed XES3G5M targets, we compare AKCT with **K**, **T**, and **T**+**D** as exposure increases from 0 to 3,000 distinct training-student responses in nested subsets. A fixed correctness target for each item makes the readout errors comparable across exposure levels. Zero exposure corresponds to unseen items. Each selected response contributes once; background data and training steps remain fixed. The probe is fitted on Source items.

Figure 4 shows an initial readout benefit: at zero exposure, **T**+**D** reduces MAE from .0970 to .0781 relative to **T**, with a paired item-bootstrap 95% interval of [.0029,.0356] for the reduction. As responses accumulate, MAE falls to .0393 for **T** and .0468 for **T**+**D** at 3,000 responses; their endpoint difference has an interval spanning zero. **K** also improves, from .0919 to .0628. The initial readout advantage diminishes as response evidence accumulates: transferred difficulty information is most helpful before the model has learned from the target items themselves.

4.5 Representation Analysis

We extend the correctness-readout analysis to all three datasets. The upstream head predicts item correctness and difficulty level from content (Appendix C.1). A probe measures how readily correctness can be read from a fixed KT representation. KT metrics evaluate the final prediction of an individual student's response using their history.

A Ridge probe reads pooled correctness from fixed, empty-history AKCT states with masked responses, using items with at least 20 responses. Seen-item evaluation rotates five item folds, fitting on four and predicting the fifth. Unseen items use a Source-fitted probe (Appendix F). Empty history isolates item-conditioned representations, while keeping KT states fixed measures how readily a linear readout can recover correctness.

Table 5: Final-decoder probe MAE (pp) under empty history. Values and paired Δ versus **T** are mean ± standard deviation. Protocol: Appendix F.

		Seen items		Unseen items	
Dataset	Input	MAE ↓	Δ vs. **T**	MAE ↓	Δ vs. **T**
XES3G5M	**T**	13.43±0.28	–	13.41±0.10	–
	T+**D**	9.39±0.03	−4.04±0.28	13.07±0.05	−0.33±0.09
	T+**S**+**D**	9.92±0.08	−3.50±0.34	12.75±0.04	−0.66±0.10
DBE-KT22	**T**	6.17±1.91	–	14.74±0.94	–
	T+**D**	3.65±0.27	−2.53±1.97	12.62±0.41	−2.12±0.84
	T+**S**+**D**	3.49±0.18	−2.68±2.08	12.37±0.39	−2.37±0.94
MOOCRadar	**T**	11.52±2.54	–	14.71±0.72	–
	T+**D**	10.82±2.09	−0.70±0.48	13.31±0.61	−1.40±0.22
	T+**S**+**D**	11.41±1.59	−0.11±2.46	12.79±0.19	−1.93±0.60

Δ = row condition − **T**. Yellow: lower MAE than **T**.

Difficulty-adapted inputs reduce mean probe MAE on both item sets across all three datasets (Table 5). The unseen-item reductions for **T**+**D** are 0.33, 2.12, and 1.40 pp on XES3G5M, DBE-KT22, and MOOCRadar, respectively. The reductions indicate that difficulty supervision makes item correctness more accessible to a linear readout, including on items held out from KT training.

Solution enrichment further lowers unseen-item probe MAE on all three datasets, linking enriched content to more accessible correctness information on new questions.

4.6 Controlled Baseline Comparisons

We compare the adapted representations with difficulty-based reference implementations in Table 6. QDCKT (Liu et al., 2024) uses difficulty-level representations to share information across questions, while RAGKT (Ma et al., 2026) retrieves similar questions to guide an LLM's difficulty estimate for a new question. As official implementations are not publicly available, we implement targeted modules based on the core ideas of QDCKT and RAGKT (Appendix G.6) and evaluate them under a common protocol with the same AKCT backbone, student/item splits, training budget, and evaluation procedure. QDCKT receives a Source-supervised difficulty estimate; RAGKT uses five Source references and an LLM estimate, with a matched no-retrieval control. These scalars replace the text input.

The scalar baselines in Table 6 show dataset-dependent gains over **T**, while retrieval alone produces no consistent AUC improvement over the matched no-retrieval control. The shared interface lets us compare how these alternative difficulty inputs contribute to KT prediction. On AKCT, T+D has higher mean unseen-item AUC than every scalar baseline on all three datasets, with the largest margins on DBE-KT22 and MOOCRadar. **T**+**D** also lowers LogLoss and Brier relative to **T** on all three datasets. The gains therefore extend from ranking student responses to the accuracy of their predicted probabilities.

5 Related Work

Knowledge Tracing. KT has progressed from latent-skill and factorization models (Corbett and Anderson, 1994; Vie and Kashima, 2018) to recurrent, memory-based, and attentive models (Piech et al., 2015; Zhang et al., 2017; Pandey and Karypis, 2019; Ghosh et al., 2020; Choi et al., 2020; Liu et al., 2023a). Contrastive learning improves history representations (Lee et al., 2022); surveys and pyKT organize model development and evaluation (Abdelrahman et al., 2022; Shen et al., 2024; Liu et al., 2022).

Content and Difficulty Representations. EERNN/EKT and RKT connect exercise text to response histories and item relations (Su et al., 2018; Liu et al., 2021; Pandey and Srivastava, 2020). PEBG uses difficulty and question–skill relations for pretraining (Liu et al., 2020); QDCKT and DCL4KT

Table 6: Unseen-item performance with AKCT (mean ± standard deviation). Δ is paired AUC change from **T** in percentage points. † denotes controlled implementations.

Input / baseline	AUC ↑	Δ vs. **T**	LogLoss ↓	Brier ↓
XES3G5M				
K	0.6951±0.0054	+0.01±0.12	0.4741±0.0045	0.1519±0.0015
T	0.6949±0.0043	–	0.4744±0.0036	0.1523±0.0013
T+**S**	0.6980±0.0006	+0.31±0.49	0.4723±0.0016	0.1510±0.0002
T+**D**	0.7073±0.0014	+1.24±0.49	0.4719±0.0071	0.1501±0.0014
T+**S**+**D**	**0.7136±0.0006**	+1.86±0.46	0.4706±0.0064	**0.1493±0.0011**
QDCKT†	0.7065±0.0027	+1.16±0.48	**0.4692±0.0044**	0.1497±0.0008
RAGKT†	0.6955±0.0056	+0.06±0.14	0.4740±0.0046	0.1518±0.0016
LLM (no retrieval)	0.6954±0.0055	+0.04±0.13	0.4740±0.0045	0.1518±0.0015
DBE-KT22				
K	0.6662±0.0031	+0.69±0.24	0.5582±0.0142	0.1869±0.0020
T	0.6594±0.0036	–	0.5582±0.0074	0.1866±0.0043
T+**S**	0.6646±0.0125	+0.52±1.20	0.5655±0.0297	0.1882±0.0070
T+**D**	0.7048±0.0048	+4.54±0.30	0.5284±0.0161	0.1721±0.0030
T+**S**+**D**	**0.7082±0.0070**	+4.88±0.85	**0.5162±0.0079**	**0.1700±0.0023**
QDCKT†	0.6655±0.0027	+0.61±0.27	0.5584±0.0153	0.1868±0.0023
RAGKT†	0.6632±0.0066	+0.38±0.37	0.5683±0.0290	0.1898±0.0059
LLM (no retrieval)	0.6661±0.0022	+0.67±0.32	0.5578±0.0146	0.1866±0.0021
MOOCRadar				
K	0.7948±0.0057	–0.07±0.36	0.3916±0.0076	0.1251±0.0041
T	0.7954±0.0058	–	0.3948±0.0047	0.1260±0.0028
T+**S**	0.8030±0.0020	+0.75±0.40	0.3914±0.0044	0.1248±0.0028
T+**D**	0.8168±0.0056	+2.13±0.12	0.3755±0.0058	0.1187±0.0031
T+**S**+**D**	**0.8251±0.0038**	+2.97±0.29	**0.3649±0.0031**	**0.1151±0.0015**
QDCKT†	0.7946±0.0057	–0.08±0.37	0.3918±0.0080	0.1252±0.0042
RAGKT†	0.7948±0.0060	–0.06±0.38	0.3913±0.0071	0.1249±0.0038
LLM (no retrieval)	0.7949±0.0061	–0.06±0.38	0.3913±0.0071	0.1250±0.0038

Δ = row condition − **T**. **Bold**: best mean within each dataset/metric. Yellow/blue: better/worse than **T**.

learn difficulty-level representations and difficulty-aware embeddings (Liu et al., 2024; Lee et al., 2024). Item-level supervision summarizes how a learner population responds to question content, complementing the student-specific response modeling performed by KT. We examine how this supervision shapes semantic transfer through label controls, representation readouts, and response-exposure interventions.

Generalization. SINKT uses semantic graph structure for inductive KT (Fu et al., 2024); EAKT extracts LLM-derived item attributes (Guo et al., 2024); RAG-KT estimates difficulty with retrieved examples (Ma et al., 2026). MathBERT embeddings also support new-question prediction (Alotaibi et al., 2026). AdaptKT, relation-aware domain generalization, and CLST address cross-domain settings (Cheng et al., 2022; Xie et al., 2026; Jung et al., 2025). Our focus is how difficulty supervision makes semantics useful for unseen-item generalization.

6 Conclusion

We studied the role of difficulty in semantic transfer for unseen-item knowledge tracing. We adapted a pretrained text encoder with discrete difficulty and continuous correctness supervision from training-item responses, then transferred its representations to KT. Prediction experiments, supervision controls, representation analyses, and controlled response exposure show how this supervision improves new-item prediction and makes correctness information more accessible. The readout advantage diminishes as item-specific response experience accumulates. Together, these findings show how difficulty supervision helps semantic representations carry information from observed student responses to questions with no training responses of their own.

AI USE STATEMENT

AI tools assisted with translation, language polishing, manuscript organization, drafting sections of the paper, and retrieval and discovery during early investigation of related work.

ETHICS STATEMENT

This study uses publicly available educational datasets and does not involve personally identifiable student information.

REPRODUCIBILITY STATEMENT

The anonymous code repository for this paper is available at https://anonymous.4open.science/r/KT_Generalization-8056/.

REFERENCES

Ghodai Abdelrahman, Qing Wang, and Bernardo Pereira Nunes. Knowledge tracing: A survey, 2022. URL https://arxiv.org/abs/2201.06953.

Ghodai Abdelrahman, Sherif Abdelfattah, Qing Wang, and Yu Lin. Dbe-kt22: A knowledge tracing dataset based on online student evaluation, 2023. URL https://arxiv.org/abs/2208.12651.

Raghad Alotaibi, Samar Alkhuraiji, and Hajar Alharbi. Improving cold-question generalization in knowledge tracing via mathbert-based semantic embedding. *Applied Sciences*, 16(17), 2026. ISSN 2076-3417. doi: 10.3390/app16178351. URL https://www.mdpi.com/2076-3417/16/17/8351.

Song Cheng, Qi Liu, Enhong Chen, Kai Zhang, Zhenya Huang, Yu Yin, Xiaoqing Huang, and Yu Su. Adaptkt: A domain adaptable method for knowledge tracing. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 123–131, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498379. URL https://doi.org/10.1145/3488560.3498379.

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, page 341–344, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379519. doi: 10.1145/3386527.3405945. URL https://doi.org/10.1145/3386527.3405945.

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, Dec 1994. ISSN 1573-1391. doi: 10.1007/BF01099821. URL https://doi.org/10.1007/BF01099821.

Lingyue Fu, Hao Guan, Kounianhua Du, Jianghao Lin, Wei Xia, Weinan Zhang, Ruiming Tang, Yasheng Wang, and Yong Yu. Sinkt: A structure-aware inductive knowledge tracing model with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 632–642, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679760. URL https://doi.org/10.1145/3627673.3679760.

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, page 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.

Yuxiang Guo, Shuanghong Shen, Qi Liu, Zhenya Huang, Linbo Zhu, Yu Su, and Enhong Chen. Mitigating cold-start problems in knowledge tracing with large language models: An attribute-aware approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 727–736, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679664. URL https://doi.org/10.1145/3627673.3679664.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL https://doi.org/10.1080/00401706.1970.10488634.

Heeseok Jung, Jaesang Yoo, Yohaan Yoon, and Yeonju Jang. Clst: Cold-start mitigation in knowledge tracing by aligning a generative language model as a students' knowledge tracer. *Journal of Educational Data Mining*, 17(2):86–117, Dec. 2025. URL https://jedm.educationaldatamining.org/index.php/JEDM/article/view/854.

Unggi Lee, Sungjun Yoon, Joon Seo Yun, Kyoungsoo Park, YoungHoon Jung, Damji Stratton, and Hyeoncheol Kim. Difficulty-focused contrastive learning for knowledge tracing with a large language model-based difficulty prediction. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4891–4900, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.438/.

Wonsung Lee, Jaeyoon Chun, Youngmin Lee, Kyoungsoo Park, and Sungrae Park. Contrastive learning for knowledge tracing. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2330–2338, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3512105. URL https://doi.org/10.1145/3485447.3512105.

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 2024. doi: 10.24963/ijcai.2024/654. URL https://doi.org/10.24963/ijcai.2024/654. Main Track.

Guimei Liu, Huijing Zhan, and Jung-jae Kim. Question difficulty consistent knowledge tracing. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4239–4248, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645582. URL https://doi.org/10.1145/3589334.3645582.

Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115, 2021. doi: 10.1109/TKDE.2019.2924374.

Yunfei Liu, Yang Yang, Xianyu Chen, Jian Shen, Haifeng Zhang, and Yong Yu. Improving knowledge tracing via pre-training question embeddings. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1577–1583. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/219. URL https://doi.org/10.24963/ijcai.2020/219. Main track.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18542–18555. Curran Associates, Inc., 2022. doi: 10.52202/068431-1347. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/75ca2b23d9794f02a92449af65a57556-Paper-Datasets_and_Benchmarks.pdf.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. simpleKT: A simple but tough-to-beat baseline for knowledge tracing. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=9HiGqC9C-KA.

Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 32958–32970. Curran Associates, Inc., 2023b. doi: 10.52202/075280-1429. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/67fc628f17c2ad53621fb961c6bafcaf-Paper-Datasets_and_Benchmarks.pdf.

Hui Ma, Tiancheng Zhang, Wangyue Lu, Fan Li, Minghe Yu, Yifang Yin, and Ge Yu. Ragkt: Retrieval augmented generation based difficulty estimation for knowledge tracing. In Irfan Awan, Muhammad Younas, Yanchun Zhang, Mahmoud Barhamgi, and Hua Wang, editors, *Web Information Systems Engineering – WISE 2025*, pages 275–290, Singapore, 2026. Springer Nature Singapore. ISBN 978-981-95-7251-9.

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

Shalini Pandey and Jaideep Srivastava. Rkt: Relation-aware self-attention for knowledge tracing. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1205–1214, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411994. URL https://doi.org/10.1145/3340531.3411994.

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, Qi Liu, Zhenya Huang, Yonghe Zheng, Minghao Yin, Minjuan Wang, and Enhong Chen. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, 17:1858–1879, 2024. doi: 10.1109/TLT.2024.3383325.

Yu Su, Qingwen Liu, Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Chris Ding, Si Wei, and Guoping Hu. Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Jill-Jênn Vie and Hisashi Kashima. Knowledge tracing machines: Factorization machines for knowledge tracing, 2018. URL https://arxiv.org/abs/1811.03388.

Yuquan Xie, Shengtao Peng, Wanqi Yang, Ming Yang, and Yang Gao. Domain-generalizable relation-aware knowledge tracing for cold-start education system. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4041–4045, 2026. doi: 10.1109/ICASSP55912.2026.11463741.

Jifan Yu, Mengying Lu, Qingyang Zhong, Zijun Yao, Shangqing Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Hai-Tao Zheng, Juanzi Li, and Jie Tang. Moocradar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in moocs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2924–2934, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591898. URL https://doi.org/10.1145/3539618.3591898.

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, page 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.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

APPENDIX

A LIMITATIONS

The main backbone comparison uses one item split and one upstream seed. Additional AKCT experiments vary item partitions, encoder initialization, supervision objectives, and target construction (Appendix I). Evaluation remains within each dataset; cross-dataset transfer and broader benchmark comparisons remain open. The equal-budget control uses one label permutation and one upstream initialization. Additional permutations would test the stability of that control; comparisons with encoder adaptation through KT response loss would examine alternative supervision objectives.

The Oracle experiment uses extra held-out-item responses under an independent protocol. It measures the value of additional response-derived information. Probe diagnostics cover AKCT on all three datasets and measure linear decodability with empty history. Source OOF excludes items from probe fitting only. The single replay tracks probe R^2 during training. Further interventions could distinguish the contributions of different representation properties. Finally, the method assumes available reference answers, and empirical difficulty depends on the learner population.

B EVALUATION AND STATISTICAL PROTOCOL

B.1 STUDENT AND ITEM PARTITIONS

All experiments are conducted separately within each dataset. The Source set is the training-item collection in that dataset. XES3G5M/DBE-KT22 use the existing 70/10/20 student partition and item fold 0 as Unseen Test. Unseen Validation is independently held out from the remaining items. Table 7 distinguishes the item universe, the subset eligible for upstream supervision, and the subset eligible for the probe where evaluated. Exported text banks cover the full item universe, including items without an upstream supervision target. MOOCRadar uses a multidisciplinary text subset spanning 22 courses, with within-course item holdout, student-disjoint partitions, and fixed fold 0. All three datasets use six backbones and the five input conditions.

Table 7: Sample counts in the current protocol. A dash denotes a diagnostic split that is not used. Source/validation/test refer to items; student populations are independently disjoint.

Dataset	Stage or measurement	Source	Unseen Validation	Unseen Test
XES3G5M	All KT items	4,878	861	1,913
	Stage 1 eligible items	2,543	459	1,024
	Correctness-probe items	3,914	—	1,538
DBE-KT22	All KT items	143	16	53
	Stage 1 eligible items	143	16	53
	Correctness-probe items	143	—	53
MOOCRadar	All KT items	1,582	279	629
	Stage 1 eligible items	338	55	122
	Correctness-probe items	1,054	—	416

Unseen-validation and unseen-test interactions are removed from every training sequence and every evaluation history. An unseen item enters only as the final query of its corresponding evaluation example. Its current student response is hidden. A reference answer belongs to item metadata and is available in the base content. Seen queries are test-student interactions with Source items; unseen queries are test-student interactions with held-out items. Both query collections use the checkpoint selected by Unseen Validation. They contain different items and histories, so their AUC difference describes two evaluation populations.

MOOCRadar subset. The selected text subset contains 2,490 items, 14,211 students, and 897,287 responses across 22 courses. Items require a stem, reference answer, KC metadata, a response, and a course with at least eight eligible items; conservative text rules exclude detected image dependence. Within each course, fixed-hash ordering assigns four item folds, with fold 0 for unseen

test and approximately 15% of the remaining items for unseen validation. Student hashing gives 9,981/1,407/2,823 train/validation/test students, independently of the item split. The evaluation sets contain 9,629 unseen-validation queries, 100,000 seen-test queries, and 41,739 unseen-test queries. Item and student partitions are disjoint within their respective roles. Stage 1 thresholds use training items only; the two class cutoffs are 0.70824 and 0.82434. The task tests item holdout within courses, not transfer to new courses.

B.2 Sampling, optimization, and checkpoint selection

Each example contains up to 199 preceding interactions plus the final query. The model receives only the final-query response label for supervision. Training samples from the full eligible training pool for a fixed budget of 10,000 optimizer steps. Table 8 records the shared settings. This budget defines the training procedure independently of epoch terminology.

Table 8: Shared Stage 2 optimization and evaluation settings.

Setting	Value
KT seeds	13, 29, 47
Upstream encoder seed	13; one exported bank per content/adaptation arm
Optimizer / learning rate	Adam / 10^{-4}
Steps / gradient clipping	10,000 / 1
Effective batch / microbatch	64 / 16
Precision	BF16 autocast
Maximum sequence length	200, including final query
Validation interval / selection	500 steps / highest Unseen Validation AUC
Validation query cap / seed	20,000 / 2026
Test query cap / seed	100,000 / 2028
XES3G5M Seen / Unseen Test queries	100,000 / 100,000
DBE-KT22 Seen / Unseen Test queries	23,690 / 8,872

The smaller DBE-KT22 test sets contain all available candidates under the stated cap. Evaluation labels and ordering have been checked for consistency across the current conditions. Backbone-specific parameterizations are held fixed within each comparison. Source KC indices have learned lookup entries; KCs absent from training map to index zero.

B.3 Means, standard deviations, and paired changes

For a condition with KT-seed scores v_1, v_2, v_3, we report

$$\bar{v} = \frac{1}{3}\sum_{j=1}^{3} v_j, \qquad s_v = \sqrt{\frac{1}{2}\sum_{j=1}^{3}(v_j - \bar{v})^2}. \tag{4}$$

A paired difference between conditions A and B is $d_j = 100(v_j^A - v_j^B)$, measured in percentage points. We report its own mean and sample standard deviation. Paired standard deviations are computed from matched within-run changes. Small discrepancies between a displayed difference and a difference of displayed means arise from rounding. Positive-seed counts describe the direction of matched differences. The shared upstream bank makes these summaries measures of downstream training randomness conditional on that bank and item split.

B.4 Privileged difficulty experiment

C Difficulty Targets and Encoder Adaptation

Training students are ordered by correctness computed only from Source responses and divided into five stable, equally sized groups. Students with no Source response are excluded from grouping. Each group's prior π_g is estimated from Source items. The smoothed per-group rates and their mean are defined in Equation 2. This mean is population-dependent empirical easiness. Group weighting and

Table 9: Difficulty priors in an independent XES3G5M experiment. These inputs use additional response support for held-out items. Each row is a separately trained model; one seed is used throughout this block.

Prior input	Seen AUC ↑	Unseen AUC ↑	Unseen gain (pp)	Unseen LogLoss ↓
No difficulty	.7323	.6984	—	.4730
Three levels	.7822	.7775	+7.91	.4281
Seven levels	.7892	.7894	+9.09	.4190
Continuous mean	.7935	.7951	+9.67	.4147

smoothing distinguish it from the pooled target used by the probe. The present supervision uses the scalar mean and its three-level categorization.

The XES3G5M thresholds are rounded, approximately equal-frequency cutoffs at 0.75 and 0.85. DBE-KT22 cutoffs are fitted on its own eligible Source items, approximately 0.715794 and 0.863828. Categories are ordered from easiest (1) to hardest (3). The regression head predicts the original easiness direction, so larger outputs indicate easier items. The classification and regression heads share a hidden layer and are optimized jointly.

Table 10: Upstream encoder training. Both adapted content conditions use this configuration.

Setting	Value
Encoder	Qwen3-Embedding-0.6B
Representation	Last non-padding token; 1,024 dimensions; L2 normalization
Input processing	No added special tokens; limit 8,192; no silent truncation
Supervision eligibility	Total attempts ≥ 100 and each group ≥ 10
Head	Linear(1,024,128), GELU, Linear(128,4)
Outputs	Three class logits and one sigmoid regression output
Loss / relative weight	CE + MAE / 1:1
Trainable parameters	Full encoder and auxiliary head
Optimizer	AdamW
Encoder / head learning rate	2×10^{-5} / 10^{-4}
Weight decay / gradient clipping	0.01 / 1
Batch size / epochs / seed	32 / 10 / 13
Epoch selection	Unseen Validation macro recall, then lower CE + MAE
Chosen XES3G5M epochs (T+D / T+S+D)	9 / 6
Chosen DBE-KT22 epochs (T+D / T+S+D)	6 / 5
Chosen MOOCRadar epochs (T+D / T+S+D)	5 / 6
Stage 2 export	Discard head; freeze full-item vector bank

The two stages reuse Source response information through different optimization targets. Stage 1 presents item-level aggregate labels to the pretrained encoder. Stage 2 presents individual final-query response labels to a history-conditioned model. The unadapted **T** condition uses the same pretrained initialization as the adapted conditions, with its encoder frozen. The main comparison therefore evaluates a complete change in adaptation procedure. The equal-budget label-permutation experiment in Appendix H.1 separately tests whether the correspondence between item content and difficulty targets contributes to the gains.

C.1 Upstream Prediction Results

Adding solutions slightly reduces scalar MAE on all three datasets. Macro recall improves on XES3G5M and MOOCRadar, while Spearman improves only on DBE-KT22 (Table 11). Further upstream readouts appear in Appendix C.1.

Table 11: Upstream predictions on eligible unseen items. Macro recall evaluates three levels; MAE and Spearman evaluate the continuous sigmoid head against group-smoothed correctness. Item counts and additional metrics: Appendix C.1.

Dataset	Input	Macro recall ↑	MAE ↓	Spearman ↑
XES3G5M	T+D	.4491	.1015	.3666
	T+S+D	.4808	.0985	.2341
DBE-KT22	T+D	.4284	.2432	-.0701
	T+S+D	.4267	.2408	.0395
MOOCRadar	T+D	.4634	.1934	-.2292
	T+S+D	.4976	.1892	-.3168

Table 12 includes the complete upstream metrics, including regression R^2. Continuous regression and class-mapped estimates are separate outputs. The latter maps the predicted difficulty class to a representative correctness rate estimated from Source items. The main representation-transfer conditions use the pre-head encoder vector.

Table 12: Two upstream scalar readouts on eligible unseen items (seed 13). Both are evaluated against the same group-smoothed correctness target; neither scalar is the KT input in the main representation-transfer conditions.

Dataset	Input	Readout	Items	MAE ↓	Spearman ↑	R^2 ↑
XES3G5M	T+D	Continuous	1024	0.1015	0.3666	−0.0700
XES3G5M	T+D	Class-mapped	1024	0.1033	0.3652	−0.1083
XES3G5M	T+S+D	Continuous	1024	0.0985	0.2341	0.0278
XES3G5M	T+S+D	Class-mapped	1024	0.1098	0.3953	−0.2082
DBE-KT22	T+D	Continuous	53	0.2432	−0.0701	−2.1264
DBE-KT22	T+D	Class-mapped	53	0.1308	0.5234	−0.1488
DBE-KT22	T+S+D	Continuous	53	0.2408	0.0395	−2.0624
DBE-KT22	T+S+D	Class-mapped	53	0.1230	0.5667	−0.0291
MOOCRadar	T+D	Continuous	122	0.1934	−0.2292	−2.1680
MOOCRadar	T+D	Class-mapped	122	0.0873	0.5693	0.1785
MOOCRadar	T+S+D	Continuous	122	0.1892	−0.3168	−2.0946
MOOCRadar	T+S+D	Class-mapped	122	0.0912	0.5105	0.0919

MOOCRadar illustrates the distinction: the continuous head has MAE 0.1934 and Spearman −0.2292, whereas class-mapped estimates give MAE 0.0873 and Spearman 0.5693. For the base-content model, three-class macro recall is 0.4634, with zero recall for the hardest class. DBE-KT22 class-mapped estimates also have positive rank correlations (0.5234 and 0.5667). The class-mapped readout retains useful ranking information despite the weaker continuous head. For MOOCRadar solution adaptation, macro recall is 0.4976, continuous-head MAE is 0.1892, and Spearman is −0.3168. Its 122 eligible test items include 93 with usable generated solutions and 29 retained with the unavailable-material marker. The selected upstream epoch is 6.

DBE-KT22's continuous sigmoid-head estimates have MAE near 0.24 and weak rank correlation (Table 12). The corresponding adapted representations nevertheless improve KT AUC on unseen items across all six backbones (Table 2). This provides another case in which the quality of the scalar difficulty output gives an incomplete account of the downstream representation benefit.

D Backbone Interfaces

D.1 Adaptive KC Transformer (AKCT)

AKCT is the experimental encoder–decoder called `KC-E-D` in the original result exports. The renamed presentation uses the same runs and parameters. For an item with N_q knowledge components, the within-item sequence has $N_q + 2$ tokens: one for each KC, one auxiliary content token, and one

response token. All have dimension 128. The auxiliary vector is projected from 1,024 dimensions through a learned linear layer. Token-type embeddings mark KC, auxiliary, and response roles. All tokens of an item share its temporal position embedding. There is no KC-local positional encoding.

One within-item encoder layer contextualizes the item. Two decoder layers use the current encoded item tokens as queries and all valid encoded tokens of preceding items as memory. Current and future responses are excluded from the history memory. The query response token is MASK (index 3); its final decoder state is mapped to one logit by a linear prediction head. A learned empty-history vector provides memory when no preceding interaction exists. Each Transformer layer uses eight heads, feed-forward dimension 256, and dropout 0.1. Encoded auxiliary and MASK representations integrate content with KC information.

D.2 AKT

Let e_k denote a learned KC embedding and P the text projection. The adapted item and historical interaction inputs are

$$h_q = \frac{1}{|\mathcal{K}_q|} \sum_{k \in \mathcal{K}_q} e_k + P z_q, \qquad h_{q,r} = h_q + e_r. \tag{5}$$

The question and question–answer streams enter the pyKT AKT core. Causal history retrieval produces a representation that is concatenated with h_q for the official prediction MLP. Query-response masking and the attention mask prevent access to the current answer. Problem-ID difficulty is explicitly disabled.

The configuration has dimension 256, eight heads, feed-forward dimension 512, dropout 0.2, and four blocks. Internally this yields four question–answer attention layers and eight alternating question-self/history-retrieval layers. Here, "four blocks" names the core configuration. AKCT preserves separate KC tokens, whereas AKT averages KCs; we interpret improvements within each backbone's fixed interface.

D.3 Additional backbone interfaces

SimpleKT (Liu et al., 2023a), StableKT (Li et al., 2024), SAKT (Pandey and Karypis, 2019), and DKVMN (Zhang et al., 2017) extend the comparison to additional public model cores. The experimental implementations use NoPID input adaptations: KC and response inputs are shared across conditions, and item content is supplied through a frozen 1,024-dimensional bank. Input dimensions, banks, splits, and the 10,000-step budget are matched within each backbone. These results characterize the adapted cores under the shared protocol described here.

E Solution Content and Provenance

The base input contains the question, its existing KC text, and the reference answer. Solutions are generated with access to this same content and without item correctness or difficulty labels. Both raw-solution and adapted-solution conditions use matched input text. Invalid material uses the same unavailable-solution marker. The solution encoder is trained with the joint loss in Equation 3; the classifier and regressor share the encoder.

MOOCRadar solution generation retains usable material for 1,848 of 2,490 items. Items with unavailable solutions remain in the fixed evaluation sets and use the common unavailable-solution marker. The completed experiment includes both solution conditions across all six backbones.

E.1 Generation Instructions and Cached Examples

We present one cached example from each dataset, all from the unseen-test item set. The Chinese instructions and generated solutions are translated into English for display; the DBE-KT22 question was originally in English. The complete original messages and responses are preserved in the companion file `docs/appendix/solution_generation_records.md`. In each example, **S** consists of the returned `solution` field; the separately returned method list and answer are shown to document the output format.

Generation instructions (condensed English translation). Given the question, KCs, options when available, and reference answer, independently solve the problem and check the result. Return a JSON object with `status`, `solution`, `methods`, and `answer`. Preserve the necessary calculations and reasoning. Do not predict difficulty or student correctness, invent missing visual information, or alter a derivation to match the reference answer. Use `insufficient` when essential information is missing and `answer_conflict` when the result conflicts with the reference. XES3G5M and DBE-KT22 request a short but complete solution and one to four concrete methods; MOOCRadar requests one to three sentences containing the necessary justification. The XES3G5M request also includes four preceding demonstration exchanges, including a missing-image example returning `insufficient`; these are retained in the original record.

E.1.1 XES3G5M: Counting Line Intersections

Item: `1853`; unseen test.
Question: What is the maximum number of intersection points formed by 13 straight lines in the same plane?
KCs: Extended thinking → Combinatorics → Mathematical puzzles → Point–line arrangements → Lines and intersections.
Reference answer: 78.

Cached solution used as S (English translation). To maximize the number of intersections, every pair of lines must intersect, and no three lines may pass through the same point. Each pair determines one intersection. Thus, the maximum number of intersections for 13 lines equals the number of ways to choose 2 of the 13 lines: $\binom{13}{2} = 13 \times 12/2 = 78$.

Other returned fields: Status: `ok`. Methods: combinatorial counting; analyzing the conditions for the maximum number of intersections; counting pairwise intersections. Answer: 78.

E.1.2 DBE-KT22: Identifying an SQL Command

Item: `dbe_kt22:question:59`; unseen test.
Question: In SQL, which command is used to add new rows to a table?
Options: (1) Alter Table; (2) Add row; (3) Insert; (4) Append.
KC: INSERT: The INSERT statement is used to add tuples into a relation.
Reference answer: Insert (option 3).

Cached solution used as S (English translation). In SQL, INSERT is the command used to add new rows (tuples) to a table. ALTER TABLE modifies the table structure; "ADD row" and "Append" are not standard SQL commands. By the definitions of basic database operations, INSERT is specifically used to insert data.

Other returned fields: Status: `ok`. Methods: identifying the functions of standard SQL commands; distinguishing data definition language (DDL) from data manipulation language (DML). Answer: Option 3.

E.1.3 MOOCRadar: Validity and Truth

Item: `Pm_2046136`; unseen test.
Question: Is the following statement correct? "Formally valid reasoning can guarantee the truth of the conclusion."
Options: (A) Correct; (B) Incorrect.
KCs: Deductive reasoning; reasoning; conclusion.
Reference answer: B (Incorrect).

Cached solution used as S (English translation). Validity guarantees that the conclusion is true when the premises are true, but the correctness (validity) of the inference does not guarantee that the premises themselves are true. If the premises are false, the conclusion may also be false even when the form of the inference is correct.

Other returned fields: Status: `ok`. Method: analyzing definitions in logic. Answer: B.

F Representation Readout Protocol and Full Results

F.1 Extraction points

The diagnostic examines AKCT for **T**, **T+D**, and **T+S+D** on all three datasets. MOOCRadar additionally includes **K**, whose readout omits language and projection layers. The unadapted solution condition and other backbones are outside this probe analysis. We evaluate each item at position zero with an empty real history and a MASK response. Decoder memory is the model's learned empty-history placeholder. This setting isolates the item-side computation under empty history. Extracted values have been checked against the actual forward computation.

Table 13: Exact readout locations. The language vector is upstream of KT; the other locations are part of the KT forward computation.

Name	Definition	Dimension
language	Last non-padding language-encoder token, L2-normalized	1,024
projection	Linear projection of the auxiliary vector, before type/position embeddings	128
encoder_aux	Penultimate within-item encoder output: auxiliary token	128
encoder_mask	Last within-item encoder output: MASK response token	128
decoder1_mask	Last MASK token after the first decoder layer	128
decoder2_mask	Last MASK token after the second decoder layer	128
logit	Scalar linear prediction from the final MASK state	1

F.2 Targets and fitting

The readout target is unsmoothed pooled correctness $a_q = c_q/n_q$, computed from training-student responses, with $n_q \geq 20$. Source/Unseen probe sets contain 3,914/1,538 XES3G5M items, 143/53 DBE-KT22 items, and 1,054/416 MOOCRadar items. Unseen response statistics are used only to assign offline diagnostic targets, separate from encoder and KT training.

Source readout uses five-fold shuffled item-level CV with seed 2031. Inside each outer training fold, 20% of its items select the Ridge regularization by MAE, with seed $2031 + \text{fold}$. Candidates are $\{0.1, 1, 10, 100, 1000, 10000\}$. StandardScaler is fitted only on the probe training subset. Unseen readout uses a scaler and Ridge model fitted from Source items, with a Source-only validation split of seed 2131 for regularization selection. Predictions are left unclipped. Recorded metrics include MAE, R^2, Pearson correlation, and Spearman correlation.

The Source OOF label refers specifically to the readout split: the underlying KT model has already trained on Source items, and the adapted upstream encoder has already used eligible Source difficulty labels. Thus, Source OOF measures statistical accessibility within a Source-trained representation. Unseen readout evaluates extrapolation of that accessibility to held-out items.

Table 14 compares initialization with the validation-selected checkpoint. The shared upstream seed and frozen banks mean that across-seed variation arises downstream of the language vector. All layer-wise metrics and seed standard deviations are recomputed from the per-item prediction caches. The MOOCRadar extension uses the same observational readout protocol. Other current KT backbones were not probed under this protocol.

For **T** on DBE-KT22, seen-item R^2 rises from –0.0227 to 0.8080, and unseen-item R^2 changes from –0.0670 to –0.2245. XES3G5M shows a milder pattern: seen-item readout increases from 0.0064 to 0.0429 and unseen-item readout from 0.0136 to 0.0213. For XES3G5M **T+D**, unseen-item R^2 changes from 0.0152 to 0.0093. The effects of KT training on readout are therefore dataset and condition dependent.

Table 5 reports MAE at the selected checkpoints, comparing input conditions. MAE is the probe's regularization-selection criterion and expresses error directly in correctness-rate units; R^2 measures squared error relative to the target variance. We compute each seed's MAE over all eligible items, then report the seed mean and sample standard deviation. This gives every item equal weight regardless of bin occupancy. The complete export includes initialization and selected checkpoints for all seven

Table 14: Final decoder MASK readout R^2, mean ± standard deviation over three KT seeds. Initialization changes only the KT checkpoint; the language bank is fixed within a condition.

		Source OOF		Unseen transfer	
Dataset	Input	Init	Selected	Init	Selected
XES3G5M	T	0.0064±0.0053	0.0429±0.0310	0.0136±0.0016	0.0213±0.0053
XES3G5M	T+D	0.0102±0.0079	0.3804±0.0025	0.0152±0.0003	0.0093±0.0080
XES3G5M	T+S+D	0.0086±0.0065	0.3483±0.0036	0.0148±0.0015	0.0691±0.0091
DBE-KT22	T	−0.0227±0.1783	0.8080±0.1070	−0.0670±0.2660	−0.2245±0.1685
DBE-KT22	T+D	−0.0809±0.2099	0.9327±0.0085	−0.0597±0.2534	0.0776±0.0496
DBE-KT22	T+S+D	−0.0593±0.1897	0.9353±0.0093	−0.0586±0.2504	0.1132±0.0725
MOOCRadar	K	0.0051±0.0161	0.1899±0.2557	0.0406±0.0136	0.0505±0.0069
MOOCRadar	T	0.0126±0.0088	0.4180±0.2320	0.0298±0.0128	0.0018±0.0960
MOOCRadar	T+D	0.0125±0.0089	0.4696±0.1837	0.0290±0.0147	0.1340±0.0577
MOOCRadar	T+S+D	0.0128±0.0089	0.4236±0.1412	0.0292±0.0144	0.2406±0.0267

layers. The following tables report both stages under the same fitting protocol; language vectors remain fixed within a condition.

On DBE-KT22, both adapted conditions improve selected-checkpoint MAE on both item sets (Table 5). Unseen-item error remains substantially higher than seen-item error. These differences concern a scalar readout under empty history; the KT predictor conditions on real student histories.

Table 15: XES3G5M initialization layer-wise correctness-probe MAE (pp), mean ± standard deviation over three KT seeds. Language banks are fixed within each condition.

	T		T+D		T+S+D	
Layer	Seen	Unseen	Seen	Unseen	Seen	Unseen
`language`	13.30±0.00	13.14±0.00	9.38±0.00	12.88±0.00	9.90±0.00	12.65±0.00
`projection`	13.58±0.04	13.32±0.00	9.43±0.02	12.95±0.04	10.04±0.04	12.68±0.02
`encoder_aux`	13.77±0.02	13.54±0.02	10.18±0.14	13.36±0.04	10.88±0.04	12.95±0.04
`encoder_mask`	13.75±0.02	13.56±0.02	13.25±0.86	13.49±0.08	13.49±0.47	13.50±0.11
`decoder1_mask`	13.75±0.05	13.51±0.05	13.65±0.13	13.45±0.09	13.69±0.09	13.48±0.10
`decoder2_mask`	13.77±0.04	13.50±0.04	13.70±0.11	13.49±0.04	13.71±0.09	13.49±0.03
`logit`	13.82±0.01	13.62±0.00	13.83±0.00	13.62±0.00	13.83±0.00	13.62±0.00

Table 16: XES3G5M trained-model layer-wise correctness-probe MAE (pp), mean ± standard deviation over three KT seeds. Language banks are fixed within each condition.

	T		T+D		T+S+D	
Layer	Seen	Unseen	Seen	Unseen	Seen	Unseen
`language`	13.30±0.00	13.14±0.00	9.38±0.00	12.88±0.00	9.90±0.00	12.65±0.00
`projection`	13.38±0.08	13.32±0.02	9.37±0.01	12.95±0.07	9.90±0.02	12.75±0.01
`encoder_aux`	13.51±0.20	13.42±0.13	9.49±0.03	13.11±0.01	10.06±0.10	12.70±0.04
`encoder_mask`	13.52±0.26	13.46±0.09	9.50±0.05	13.09±0.01	10.07±0.03	12.73±0.05
`decoder1_mask`	13.44±0.29	13.40±0.11	9.39±0.03	13.08±0.04	9.91±0.03	12.72±0.00
`decoder2_mask`	13.43±0.28	13.41±0.10	9.39±0.03	13.07±0.05	9.92±0.08	12.75±0.04
`logit`	13.43±0.31	13.49±0.11	9.48±0.01	13.17±0.06	10.03±0.10	12.86±0.04

F.3 A SINGLE REPLAY ACROSS TRAINING STEPS

The replay follows XES3G5M/T/seed 13 and saves steps 0, 500, 1,000, 2,000, 5,000, and 10,000. Its step-zero parameters match the original seed reconstruction elementwise. Step 5,000 is the selected

Table 17: DBE-KT22 initialization layer-wise correctness-probe MAE (pp), mean ± standard deviation over three KT seeds. Language banks are fixed within each condition.

	T		T+D		T+S+D	
Layer	Seen	Unseen	Seen	Unseen	Seen	Unseen
language	14.42±0.00	10.42±0.00	7.36±0.00	10.41±0.00	6.86±0.00	10.72±0.00
projection	14.08±0.57	10.88±0.91	7.95±0.40	10.68±0.10	7.94±0.18	11.29±0.19
encoder_aux	15.35±0.61	12.62±0.49	13.77±0.37	12.36±0.82	13.92±0.96	12.40±0.84
encoder_mask	15.66±0.68	13.01±0.21	15.52±0.74	12.99±0.19	15.69±0.62	12.99±0.18
decoder1_mask	15.40±0.86	13.93±1.34	15.76±1.19	13.89±1.26	15.53±0.91	13.91±1.29
decoder2_mask	15.50±1.18	13.98±1.48	15.80±1.37	13.96±1.42	15.69±1.26	13.96±1.40
logit	16.26±0.14	14.06±0.07	16.24±0.11	14.06±0.08	16.24±0.12	14.06±0.07

Table 18: DBE-KT22 trained-model layer-wise correctness-probe MAE (pp), mean ± standard deviation over three KT seeds. Language banks are fixed within each condition.

	T		T+D		T+S+D	
Layer	Seen	Unseen	Seen	Unseen	Seen	Unseen
language	14.42±0.00	10.42±0.00	7.36±0.00	10.41±0.00	6.86±0.00	10.72±0.00
projection	10.78±0.55	10.95±0.44	5.73±0.38	11.23±0.81	6.03±0.46	11.43±0.44
encoder_aux	10.10±0.53	11.94±0.36	5.68±0.40	11.34±0.47	5.42±0.43	11.71±0.73
encoder_mask	9.12±1.44	13.85±1.75	5.54±0.63	11.98±0.73	6.16±0.71	11.98±0.92
decoder1_mask	6.60±2.01	14.52±0.75	3.97±0.28	12.64±0.40	3.86±0.26	12.26±0.19
decoder2_mask	6.17±1.91	14.74±0.94	3.65±0.27	12.62±0.41	3.49±0.18	12.37±0.39
logit	6.24±0.64	14.98±0.78	5.01±0.56	11.89±0.74	4.75±0.36	12.12±0.46

Table 19: MOOCRadar initialization layer-wise correctness-probe MAE (pp), mean ± standard deviation over three KT seeds. Language banks are fixed within each condition.

	T		T+D		T+S+D	
Layer	Seen	Unseen	Seen	Unseen	Seen	Unseen
language	14.18±0.00	12.74±0.00	12.37±0.00	12.24±0.00	11.84±0.00	11.75±0.00
projection	14.37±0.09	13.55±0.16	13.14±0.19	13.14±0.22	12.56±0.03	12.49±0.16
encoder_aux	15.89±0.15	14.68±0.22	15.20±0.16	13.75±0.16	14.81±0.32	13.36±0.25
encoder_mask	15.99±0.03	15.11±0.16	16.00±0.03	15.09±0.21	16.00±0.03	15.09±0.20
decoder1_mask	15.92±0.04	15.18±0.13	15.92±0.04	15.19±0.14	15.92±0.04	15.18±0.14
decoder2_mask	15.91±0.08	15.28±0.18	15.91±0.09	15.29±0.20	15.91±0.08	15.28±0.19
logit	16.05±0.02	15.54±0.09	16.05±0.02	15.55±0.09	16.05±0.02	15.55±0.09

Table 20: MOOCRadar trained-model layer-wise correctness-probe MAE (pp), mean ± standard deviation over three KT seeds. Language banks are fixed within each condition.

	T		T+D		T+S+D	
Layer	Seen	Unseen	Seen	Unseen	Seen	Unseen
language	14.18±0.00	12.74±0.00	12.37±0.00	12.24±0.00	11.84±0.00	11.75±0.00
projection	13.16±0.20	13.27±0.29	12.71±0.13	12.86±0.11	12.20±0.34	12.52±0.20
encoder_aux	13.44±0.53	13.82±0.26	13.01±0.68	13.06±0.50	12.53±0.40	12.62±0.21
encoder_mask	14.04±0.77	14.11±0.80	12.92±0.67	13.30±0.51	12.67±0.63	12.83±0.41
decoder1_mask	11.75±2.44	14.79±0.43	11.05±1.98	13.23±0.67	11.63±1.60	12.79±0.29
decoder2_mask	11.52±2.54	14.71±0.72	10.82±2.09	13.31±0.61	11.41±1.59	12.79±0.19
logit	11.72±2.43	15.14±0.35	11.18±1.81	13.68±0.62	11.65±1.38	13.29±0.20

replay checkpoint. The saved terminal step is a separate checkpoint. The replay provides a temporal

view of this training run. Seen-item R^2 increases from 0.0114 to 0.1213 at step 10,000. Unseen-item R^2 reaches 0.0257 at step 5,000 and ends at 0.0158 (Figure 5). Increasing seen-item decodability can therefore coexist with a late reduction in unseen-item decodability.

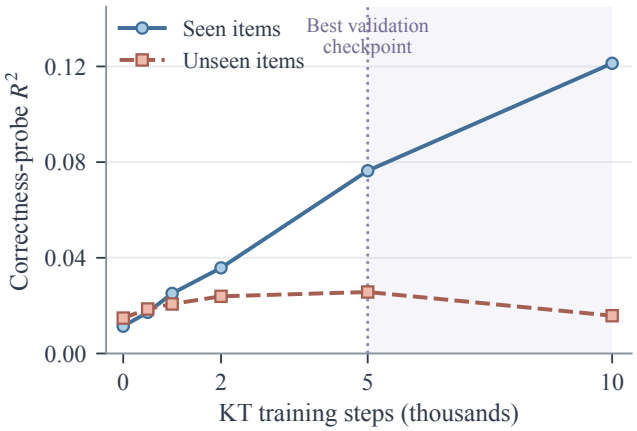

Figure 5: **Readout during KT training.** XES3G5M/T/seed 13, final decoder under empty history. Unseen-item R^2 declines after the validation-selected checkpoint at step 5,000; seen-item R^2 continues to rise.

Table 21: Trajectory values for the final decoder MASK probe in Figure 5. These are correctness-probe R^2 values under empty history.

KT step	Source OOF R^2	Unseen transfer R^2
0	.0114	.0148
500	.0172	.0186
1,000	.0251	.0207
2,000	.0358	.0239
5,000	.0764	.0257
10,000	.1213	.0158

G Complete Knowledge-Tracing Results

All tables group the same six backbones and three datasets. All five input conditions are available on all three datasets. Scores are reported within each dataset, without pooling AUC across datasets. The fixed protocol uses three KT seeds and one upstream seed.

Uncertainty. All $\pm$ terms are sample standard deviations across KT seeds. Condition scores and within-seed changes are recomputed from full-precision per-run metrics; paired standard deviations are computed from the within-seed differences. Yellow/blue means improvement/deterioration relative to the stated baseline, not statistical significance.

G.1 Scores and Changes on Three Datasets

Seen and unseen AUC are complete for all 90 trained conditions (270 runs). Tables report the same six backbones and three datasets, with all means, sample standard deviations, and paired changes available.

Table 22: Unseen AUC on three datasets, mean ± standard deviation over three KT seeds. Full uncertainty for Table 2.

Backbone	Dataset	K	T	T+S	T+D	T+S+D
AKCT	XES3G5M	0.6951±0.0054	0.6949±0.0043	0.6980±0.0006	0.7073±0.0014	0.7136±0.0006
	DBE-KT22	0.6662±0.0031	0.6594±0.0036	0.6646±0.0125	0.7048±0.0048	0.7082±0.0070
	MOOCRadar	0.7948±0.0057	0.7954±0.0058	0.8030±0.0020	0.8168±0.0056	0.8251±0.0038
AKT	XES3G5M	0.7072±0.0018	0.7053±0.0049	0.7085±0.0043	0.7128±0.0010	0.7225±0.0017
	DBE-KT22	0.6773±0.0051	0.6726±0.0051	0.6793±0.0124	0.6990±0.0020	0.6960±0.0058
	MOOCRadar	0.7957±0.0030	0.8025±0.0021	0.8057±0.0021	0.8270±0.0040	0.8324±0.0032
SimpleKT	XES3G5M	0.7052±0.0023	0.7050±0.0019	0.7052±0.0026	0.7092±0.0014	0.7189±0.0021
	DBE-KT22	0.6666±0.0034	0.6675±0.0121	0.6714±0.0113	0.7011±0.0084	0.6844±0.0092
	MOOCRadar	0.7975±0.0027	0.7981±0.0042	0.8070±0.0021	0.8243±0.0046	0.8283±0.0024
StableKT	XES3G5M	0.7091±0.0017	0.7102±0.0015	0.7109±0.0028	0.7155±0.0028	0.7246±0.0024
	DBE-KT22	0.6930±0.0171	0.6855±0.0254	0.6849±0.0262	0.7128±0.0120	0.7054±0.0235
	MOOCRadar	0.8008±0.0024	0.8029±0.0038	0.8050±0.0056	0.8207±0.0057	0.8321±0.0038
SAKT	XES3G5M	0.6667±0.0013	0.6673±0.0015	0.6671±0.0016	0.6808±0.0013	0.6898±0.0007
	DBE-KT22	0.6430±0.0072	0.6480±0.0078	0.6600±0.0046	0.6808±0.0033	0.6765±0.0010
	MOOCRadar	0.7694±0.0055	0.7766±0.0050	0.7832±0.0052	0.7987±0.0014	0.8100±0.0021
DKVMN	XES3G5M	0.6855±0.0043	0.6853±0.0018	0.6876±0.0051	0.6952±0.0010	0.7042±0.0009
	DBE-KT22	0.6616±0.0205	0.6751±0.0140	0.6739±0.0154	0.7015±0.0108	0.7045±0.0066
	MOOCRadar	0.7823±0.0026	0.7896±0.0016	0.7934±0.0056	0.8128±0.0048	0.8113±0.0068

Table 23: Unseen AUC changes from T (pp). Paired means ± standard deviation over three KT seeds.

Backbone	Dataset	K	T+S	T+D	T+S+D
AKCT	XES3G5M	+0.01±0.12	+0.31±0.49	+1.24±0.49	+1.86±0.46
	DBE-KT22	+0.69±0.24	+0.52±1.20	+4.54±0.30	+4.88±0.85
	MOOCRadar	−0.07±0.36	+0.75±0.40	+2.13±0.12	+2.97±0.29
AKT	XES3G5M	+0.19±0.45	+0.32±0.37	+0.75±0.50	+1.72±0.40
	DBE-KT22	+0.47±0.32	+0.68±1.03	+2.64±0.33	+2.35±0.18
	MOOCRadar	−0.68±0.10	+0.32±0.28	+2.45±0.45	+2.99±0.12
SimpleKT	XES3G5M	+0.02±0.16	+0.02±0.16	+0.42±0.33	+1.38±0.24
	DBE-KT22	−0.09±1.47	+0.40±0.31	+3.36±0.54	+1.69±1.84
	MOOCRadar	−0.06±0.21	+0.89±0.48	+2.62±0.48	+3.02±0.21
StableKT	XES3G5M	−0.11±0.08	+0.06±0.39	+0.53±0.34	+1.44±0.36
	DBE-KT22	+0.76±0.83	−0.06±0.16	+2.73±1.39	+1.99±1.19
	MOOCRadar	−0.21±0.14	+0.22±0.20	+1.78±0.29	+2.92±0.10
SAKT	XES3G5M	−0.06±0.03	−0.02±0.03	+1.35±0.13	+2.26±0.10
	DBE-KT22	−0.49±0.25	+1.21±0.38	+3.29±0.88	+2.85±0.78
	MOOCRadar	−0.72±0.15	+0.66±0.03	+2.21±0.49	+3.34±0.61
DKVMN	XES3G5M	+0.02±0.35	+0.23±0.49	+0.99±0.16	+1.89±0.12
	DBE-KT22	−1.35±0.64	−0.12±0.26	+2.64±0.35	+2.94±1.20
	MOOCRadar	−0.74±0.39	+0.38±0.40	+2.32±0.52	+2.17±0.71

Δ = column condition − T. Yellow/blue: better/worse than T.

Table 24: Seen AUC on three datasets, mean ± standard deviation over three KT seeds.

Backbone	Dataset	K	T	T+S	T+D	T+S+D
AKCT	XES3G5M	0.7119±0.0204	0.7094±0.0149	0.7288±0.0145	0.7748±0.0019	0.7686±0.0027
	DBE-KT22	0.7918±0.0066	0.7940±0.0059	0.7957±0.0084	0.8050±0.0013	0.8064±0.0010
	MOOCRadar	0.8307±0.0333	0.8586±0.0190	0.8703±0.0011	0.8635±0.0121	0.8662±0.0102
AKT	XES3G5M	0.7422±0.0021	0.7370±0.0204	0.7486±0.0042	0.7798±0.0023	0.7772±0.0002
	DBE-KT22	0.8054±0.0024	0.8024±0.0019	0.8083±0.0035	0.8079±0.0031	0.8082±0.0014
	MOOCRadar	0.8677±0.0060	0.8685±0.0066	0.8693±0.0064	0.8762±0.0022	0.8765±0.0021
SimpleKT	XES3G5M	0.7443±0.0033	0.7469±0.0086	0.7503±0.0046	0.7837±0.0011	0.7774±0.0012
	DBE-KT22	0.7966±0.0014	0.7977±0.0036	0.7968±0.0030	0.8047±0.0007	0.8008±0.0038
	MOOCRadar	0.8393±0.0184	0.8463±0.0171	0.8733±0.0085	0.8692±0.0151	0.8649±0.0066
StableKT	XES3G5M	0.7406±0.0174	0.7446±0.0202	0.7561±0.0062	0.7834±0.0017	0.7826±0.0025
	DBE-KT22	0.8000±0.0030	0.7967±0.0067	0.7964±0.0061	0.8049±0.0011	0.8026±0.0088
	MOOCRadar	0.8346±0.0017	0.8598±0.0229	0.8603±0.0232	0.8568±0.0192	0.8742±0.0099
SAKT	XES3G5M	0.7139±0.0060	0.7199±0.0071	0.7197±0.0070	0.7600±0.0019	0.7522±0.0019
	DBE-KT22	0.7923±0.0036	0.8000±0.0008	0.7998±0.0006	0.7985±0.0019	0.8004±0.0009
	MOOCRadar	0.8469±0.0059	0.8504±0.0063	0.8517±0.0062	0.8567±0.0003	0.8578±0.0014
DKVMN	XES3G5M	0.7174±0.0045	0.7207±0.0087	0.7261±0.0046	0.7724±0.0009	0.7637±0.0012
	DBE-KT22	0.7869±0.0207	0.7798±0.0212	0.7807±0.0263	0.8039±0.0054	0.8088±0.0021
	MOOCRadar	0.8383±0.0186	0.8658±0.0021	0.8560±0.0211	0.8675±0.0020	0.8561±0.0074

Table 25: Seen AUC changes from T (pp), paired mean ± standard deviation.

Backbone	Dataset	K	T+S	T+D	T+S+D
AKCT	XES3G5M	+0.26±0.62	+1.94±2.26	+6.54±1.61	+5.92±1.71
	DBE-KT22	−0.22±0.45	+0.17±0.68	+1.09±0.70	+1.24±0.69
	MOOCRadar	−2.79±2.75	+1.17±1.90	+0.49±0.69	+0.75±0.88
AKT	XES3G5M	+0.52±1.91	+1.16±1.83	+4.28±2.02	+4.02±2.02
	DBE-KT22	+0.29±0.20	+0.59±0.26	+0.54±0.24	+0.58±0.25
	MOOCRadar	−0.07±0.08	+0.08±0.04	+0.77±0.88	+0.81±0.58
SimpleKT	XES3G5M	−0.26±0.65	+0.34±0.57	+3.68±0.79	+3.06±0.92
	DBE-KT22	−0.11±0.43	−0.09±0.20	+0.70±0.43	+0.31±0.44
	MOOCRadar	−0.70±0.94	+2.70±2.49	+2.29±0.61	+1.86±1.98
StableKT	XES3G5M	−0.40±0.28	+1.15±1.87	+3.88±2.12	+3.80±1.88
	DBE-KT22	+0.32±0.41	−0.03±0.09	+0.82±0.67	+0.59±0.50
	MOOCRadar	−2.52±2.15	+0.06±0.03	−0.30±2.53	+1.44±1.37
SAKT	XES3G5M	−0.60±0.13	−0.02±0.01	+4.01±0.53	+3.23±0.60
	DBE-KT22	−0.77±0.33	−0.02±0.02	−0.15±0.23	+0.04±0.06
	MOOCRadar	−0.35±0.06	+0.13±0.02	+0.63±0.64	+0.75±0.69
DKVMN	XES3G5M	−0.33±0.60	+0.54±0.78	+5.17±0.93	+4.30±0.96
	DBE-KT22	+0.72±2.76	+0.09±0.56	+2.42±1.73	+2.90±2.01
	MOOCRadar	−2.75±1.79	−0.99±1.91	+0.17±0.06	−0.97±0.77

Δ = column condition − T. Yellow/blue: better/worse than T.

G.2 SOLUTION-CONTENT CONTRASTS

The solution contrasts cover all three datasets. With solutions fixed, difficulty adaptation improves mean unseen AUC for all eighteen settings, with 53 of 54 seed differences positive; DBE-KT22/SimpleKT is the exception to seed-wise consistency. With difficulty adaptation fixed, adding solutions improves all 18 XES3G5M unseen pairs, but lowers seen means for all six XES3G5M backbones (17 of 18 seed differences are negative). On DBE-KT22, the unseen solution effect varies by backbone; DKVMN has two positive seed differences out of three. MOOCRadar has 15 positive unseen solution increments in 18 pairs, with a negative DKVMN mean.

Table 26: Unseen factorial AUC changes (pp), paired mean ± standard deviation. Counts show positive changes among three seeds.

Dataset	Backbone	$(\mathbf{T}+\mathbf{S}+\mathbf{D})-(\mathbf{T}+\mathbf{S})$	Positive	$(\mathbf{T}+\mathbf{S}+\mathbf{D})-(\mathbf{T}+\mathbf{D})$	Positive
XES3G5M	AKCT	+1.55±0.07	3/3	+0.62±0.19	3/3
DBE-KT22	AKCT	+4.36±0.71	3/3	+0.33±1.12	2/3
MOOCRadar	AKCT	+2.21±0.28	3/3	+0.83±0.21	3/3
XES3G5M	AKT	+1.40±0.26	3/3	+0.97±0.12	3/3
DBE-KT22	AKT	+1.67±0.85	3/3	−0.30±0.42	1/3
MOOCRadar	AKT	+2.67±0.32	3/3	+0.55±0.46	3/3
XES3G5M	SimpleKT	+1.37±0.15	3/3	+0.97±0.31	3/3
DBE-KT22	SimpleKT	+1.30±1.88	2/3	−1.67±1.31	0/3
MOOCRadar	SimpleKT	+2.13±0.38	3/3	+0.41±0.51	2/3
XES3G5M	StableKT	+1.37±0.06	3/3	+0.91±0.15	3/3
DBE-KT22	StableKT	+2.05±1.10	3/3	−0.74±1.63	1/3
MOOCRadar	StableKT	+2.70±0.17	3/3	+1.14±0.39	3/3
XES3G5M	SAKT	+2.28±0.10	3/3	+0.91±0.14	3/3
DBE-KT22	SAKT	+1.65±0.50	3/3	−0.43±0.44	0/3
MOOCRadar	SAKT	+2.68±0.62	3/3	+1.13±0.12	3/3
XES3G5M	DKVMN	+1.66±0.55	3/3	+0.90±0.04	3/3
DBE-KT22	DKVMN	+3.06±1.42	3/3	+0.30±1.00	2/3
MOOCRadar	DKVMN	+1.79±0.96	3/3	−0.15±0.20	1/3

Yellow/blue: better/worse for the stated contrast.

Table 27: Seen factorial AUC changes (pp), paired mean ± standard deviation. Counts show positive changes among three seeds.

Dataset	Backbone	$(\mathbf{T}+\mathbf{S}+\mathbf{D})-(\mathbf{T}+\mathbf{S})$	Positive	$(\mathbf{T}+\mathbf{S}+\mathbf{D})-(\mathbf{T}+\mathbf{D})$	Positive
XES3G5M	AKCT	+3.98±1.61	3/3	−0.62±0.10	0/3
DBE-KT22	AKCT	+1.07±0.90	3/3	+0.14±0.07	3/3
MOOCRadar	AKCT	−0.42±1.03	2/3	+0.26±0.19	3/3
XES3G5M	AKT	+2.87±0.40	3/3	−0.26±0.23	0/3
DBE-KT22	AKT	−0.02±0.28	2/3	+0.03±0.25	2/3
MOOCRadar	AKT	+0.72±0.57	3/3	+0.03±0.39	1/3
XES3G5M	SimpleKT	+2.72±0.45	3/3	−0.62±0.13	0/3
DBE-KT22	SimpleKT	+0.40±0.54	2/3	−0.39±0.38	1/3
MOOCRadar	SimpleKT	−0.84±1.19	1/3	−0.43±1.97	1/3
XES3G5M	StableKT	+2.64±0.39	3/3	−0.08±0.26	1/3
DBE-KT22	StableKT	+0.63±0.57	3/3	−0.23±0.94	2/3
MOOCRadar	StableKT	+1.39±1.39	3/3	+1.74±1.62	3/3
XES3G5M	SAKT	+3.25±0.59	3/3	−0.78±0.12	0/3
DBE-KT22	SAKT	+0.05±0.04	3/3	+0.19±0.18	2/3
MOOCRadar	SAKT	+0.62±0.68	3/3	+0.12±0.16	2/3
XES3G5M	DKVMN	+3.76±0.58	3/3	−0.87±0.04	0/3
DBE-KT22	DKVMN	+2.81±2.50	3/3	+0.48±0.35	3/3
MOOCRadar	DKVMN	+0.01±2.12	1/3	−1.14±0.82	0/3

Yellow/blue: better/worse for the stated contrast.

G.3 Secondary Metrics on Three Datasets

The same selected checkpoints support the AUC and secondary metrics. The following tables compare **T**+**D** with **T** in both query populations. Changes are kept in native metric units; higher AP is better, whereas lower LogLoss, Brier, and ECE is better. Calibration is evaluated separately from discrimination.

Table 28: Unseen AP, LogLoss, and Brier for **T**+**D** and paired changes from **T** (native units). Changes report mean ± standard deviation.

		AP ↑		LogLoss ↓		Brier ↓	
Backbone	Dataset	**T**+**D**	Δ_D	**T**+**D**	Δ_D	**T**+**D**	Δ_D
AKCT	XES3G5M	0.8915	+0.0045±0.0018	0.4719	−0.0025±0.0102	0.1501	−0.0023±0.0027
	DBE-KT22	0.8871	+0.0216±0.0009	0.5284	−0.0298±0.0114	0.1721	−0.0145±0.0035
	MOOCRadar	0.9500	+0.0051±0.0001	0.3755	−0.0193±0.0043	0.1187	−0.0073±0.0020
AKT	XES3G5M	0.8914	+0.0040±0.0027	0.4677	−0.0050±0.0031	0.1487	−0.0012±0.0010
	DBE-KT22	0.8810	+0.0079±0.0032	0.5273	−0.0432±0.0060	0.1749	−0.0122±0.0023
	MOOCRadar	0.9529	+0.0065±0.0015	0.3676	−0.0433±0.0150	0.1152	−0.0165±0.0067
SimpleKT	XES3G5M	0.8913	+0.0018±0.0013	0.4714	+0.0006±0.0010	0.1497	−0.0002±0.0005
	DBE-KT22	0.8834	+0.0185±0.0028	0.5268	−0.0454±0.0022	0.1743	−0.0106±0.0013
	MOOCRadar	0.9522	+0.0075±0.0026	0.3705	−0.0188±0.0044	0.1167	−0.0072±0.0010
StableKT	XES3G5M	0.8934	+0.0023±0.0018	0.4638	−0.0039±0.0014	0.1476	−0.0013±0.0007
	DBE-KT22	0.8896	+0.0130±0.0098	0.5124	−0.0250±0.0189	0.1694	−0.0054±0.0030
	MOOCRadar	0.9507	+0.0039±0.0012	0.3685	−0.0220±0.0075	0.1164	−0.0088±0.0027
SAKT	XES3G5M	0.8798	+0.0052±0.0003	0.4862	−0.0004±0.0038	0.1551	−0.0011±0.0007
	DBE-KT22	0.8711	+0.0234±0.0045	0.5484	−0.0346±0.0132	0.1793	−0.0084±0.0027
	MOOCRadar	0.9424	+0.0068±0.0026	0.4011	−0.0257±0.0085	0.1262	−0.0090±0.0046
DKVMN	XES3G5M	0.8846	+0.0044±0.0011	0.4734	−0.0032±0.0026	0.1512	−0.0013±0.0006
	DBE-KT22	0.8839	+0.0108±0.0041	0.5225	−0.0126±0.0228	0.1728	−0.0063±0.0084
	MOOCRadar	0.9487	+0.0061±0.0016	0.3750	−0.0253±0.0050	0.1181	−0.0092±0.0016

$\Delta_D = (\mathbf{T} + \mathbf{D}) - \mathbf{T}$. Yellow/blue: better/worse than **T**.

G.4 Course-Level and Secondary Metrics

The following course-level metrics concern MOOCRadar's within-course item holdout. They supplement the common three-dataset AUC and secondary-metric comparisons above. Course-level and secondary metrics provide complementary views of the MOOCRadar intervention. AKCT's mean Unseen course-macro AUC increases from 0.6737 to 0.6840, evaluated on the 16 courses whose Unseen tests contain both response classes. Its mean LogLoss decreases from 0.3948 to 0.3755, Brier score from 0.1260 to 0.1187, and expected calibration error (ECE) from 0.0439 to 0.0381. SimpleKT's Unseen course-macro AUC increases from 0.6815 to 0.6961; StableKT's increases from 0.6908 to 0.6998. StableKT's Seen course-macro AUC decreases from 0.7285 to 0.7271. All six backbones improve mean Unseen AP, LogLoss, and Brier score. Calibration has a different pattern: SimpleKT's mean ECE increases from 0.0344 to 0.0384, and AKCT's ECE change varies across runs. Calibration effects are therefore dataset- and model-dependent.

SAKT and DKVMN extend the same comparison: their mean Unseen course-macro AUC increases from 0.6510 to 0.6833 and from 0.6362 to 0.6575, respectively. Across all six MOOCRadar backbones, the 18 paired Unseen AUC changes are positive (minimum 1.51 points); 16 of 18 Seen changes are positive. Calibration remains mixed: Seen ECE increases for SAKT and slightly for DKVMN, whose Seen course-macro AUC also decreases.

Table 29: Unseen metric changes, $(\mathbf{T}+\mathbf{D})-\mathbf{T}$, paired mean ± standard deviation in native units.

Backbone	Dataset	AP ↑	LogLoss ↓	Brier ↓	ECE ↓
AKCT	XES3G5M	+.0045±.0018	−.0025±.0102	−.0023±.0027	+.0003±.0310
	DBE-KT22	+.0216±.0009	−.0298±.0114	−.0145±.0035	−.0190±.0133
	MOOCRadar	+.0051±.0001	−.0193±.0043	−.0073±.0020	−.0058±.0090
AKT	XES3G5M	+.0040±.0027	−.0050±.0031	−.0012±.0010	−.0068±.0040
	DBE-KT22	+.0079±.0032	−.0432±.0060	−.0122±.0023	−.0401±.0159
	MOOCRadar	+.0065±.0015	−.0433±.0150	−.0165±.0067	−.0366±.0210
SimpleKT	XES3G5M	+.0018±.0013	+.0006±.0010	−.0002±.0005	+.0085±.0053
	DBE-KT22	+.0185±.0028	−.0454±.0022	−.0106±.0013	−.0324±.0095
	MOOCRadar	+.0075±.0026	−.0188±.0044	−.0072±.0010	+.0040±.0085
StableKT	XES3G5M	+.0023±.0018	−.0039±.0014	−.0013±.0007	−.0052±.0032
	DBE-KT22	+.0130±.0098	−.0250±.0189	−.0054±.0030	−.0172±.0071
	MOOCRadar	+.0039±.0012	−.0220±.0075	−.0088±.0027	−.0194±.0206
SAKT	XES3G5M	+.0052±.0003	−.0004±.0038	−.0011±.0007	+.0089±.0079
	DBE-KT22	+.0234±.0045	−.0346±.0132	−.0084±.0027	−.0152±.0063
	MOOCRadar	+.0068±.0026	−.0257±.0085	−.0090±.0046	−.0125±.0137
DKVMN	XES3G5M	+.0044±.0011	−.0032±.0026	−.0013±.0006	+.0062±.0110
	DBE-KT22	+.0108±.0041	−.0126±.0228	−.0063±.0084	−.0006±.0114
	MOOCRadar	+.0061±.0016	−.0253±.0050	−.0092±.0016	−.0123±.0030

$\Delta_D = (\mathbf{T}+\mathbf{D})-\mathbf{T}$. Yellow/blue: better/worse than $\mathbf{T}$.

Table 30: Seen metric changes, $(\mathbf{T}+\mathbf{D})-\mathbf{T}$, paired mean ± standard deviation in native units.

Backbone	Dataset	AP ↑	LogLoss ↓	Brier ↓	ECE ↓
AKCT	XES3G5M	+.0283±.0077	−.0362±.0100	−.0124±.0031	−.0028±.0277
	DBE-KT22	+.0050±.0032	−.0100±.0081	−.0039±.0030	−.0094±.0081
	MOOCRadar	+.0013±.0018	−.0043±.0053	−.0017±.0021	+.0016±.0034
AKT	XES3G5M	+.0193±.0106	−.0283±.0120	−.0089±.0040	−.0157±.0101
	DBE-KT22	+.0024±.0013	−.0079±.0052	−.0021±.0009	−.0102±.0079
	MOOCRadar	+.0025±.0027	−.0136±.0082	−.0051±.0028	−.0181±.0116
SimpleKT	XES3G5M	+.0152±.0038	−.0228±.0052	−.0076±.0017	−.00004±.0032
	DBE-KT22	+.0029±.0014	−.0091±.0020	−.0024±.0008	−.0070±.0065
	MOOCRadar	+.0063±.0015	−.0220±.0109	−.0085±.0037	+.0020±.0189
StableKT	XES3G5M	+.0168±.0099	−.0247±.0130	−.0082±.0045	−.0088±.0061
	DBE-KT22	+.0033±.0026	−.0105±.0087	−.0029±.0027	−.0088±.0121
	MOOCRadar	−.0010±.0072	+.0039±.0284	+.0014±.0107	+.0041±.0127
SAKT	XES3G5M	+.0174±.0023	−.0204±.0008	−.0071±.0006	+.0083±.0093
	DBE-KT22	−.0003±.0008	+.0031±.0040	+.0013±.0016	+.0102±.0112
	MOOCRadar	+.0018±.0019	−.0053±.0051	−.0019±.0017	+.0042±.0080
DKVMN	XES3G5M	+.0221±.0042	−.0294±.0051	−.0100±.0017	−.0029±.0119
	DBE-KT22	+.0115±.0071	−.0198±.0122	−.0067±.0043	−.0086±.0088
	MOOCRadar	+.0002±.0003	−.0018±.0006	−.0007±.0002	+.00004±.0027

$\Delta_D = (\mathbf{T}+\mathbf{D})-\mathbf{T}$. Yellow/blue: better/worse than $\mathbf{T}$.

Table 31: MOOCRadar course-macro AUC, mean ± standard deviation over three seeds. Courses without both classes are excluded from AUC; unseen evaluation includes 16 eligible courses.

	Seen		Unseen	
Backbone	T	T+D	T	T+D
AKCT	0.7266±0.0239	0.7319±0.0145	0.6737±0.0046	0.6840±0.0044
AKT	0.7432±0.0228	0.7694±0.0152	0.6964±0.0060	0.7122±0.0103
SimpleKT	0.6973±0.0326	0.7452±0.0429	0.6815±0.0033	0.6961±0.0259
StableKT	0.7285±0.0612	0.7271±0.0496	0.6908±0.0046	0.6998±0.0032
SAKT	0.7298±0.0123	0.7407±0.0083	0.6510±0.0148	0.6833±0.0276
DKVMN	0.7511±0.0133	0.7496±0.0100	0.6362±0.0098	0.6575±0.0263

G.5 Seen Results for Controlled Baseline Implementations

These comparisons use the same AKCT interface and implementation definitions as Section 4.6.

Table 32: Seen-item baseline comparisons (mean ± standard deviation, three seeds). Δ: paired AUC change from **T** (pp). †: controlled implementations, as in Section 4.6.

Input / method	AUC ↑	Δ	LogLoss ↓	Brier ↓
XES3G5M				
K	0.7119±0.0204	+0.26±0.62	0.4661±0.0117	0.1495±0.0040
T	0.7094±0.0149	–	0.4677±0.0079	0.1502±0.0025
T+S	0.7288±0.0145	+1.94±2.26	0.4563±0.0068	0.1462±0.0024
T+D	**0.7748±0.0019**	+6.54±1.61	**0.4315±0.0024**	**0.1378±0.0007**
T+S+D	0.7686±0.0027	+5.92±1.71	0.4357±0.0023	0.1393±0.0007
QDCKT†	0.7687±0.0015	+5.94±1.63	0.4345±0.0006	0.1390±0.0002
RAGKT†	0.7122±0.0205	+0.28±0.63	0.4661±0.0117	0.1494±0.0040
LLM (no retrieval)	0.7121±0.0204	+0.27±0.62	0.4660±0.0116	0.1494±0.0039
DBE-KT22				
K	0.7918±0.0066	–0.22±0.45	0.4395±0.0057	0.1417±0.0024
T	0.7940±0.0059	–	0.4381±0.0076	0.1410±0.0030
T+S	0.7957±0.0084	+0.17±0.68	0.4389±0.0068	0.1412±0.0030
T+D	0.8050±0.0013	+1.09±0.70	**0.4281±0.0032**	**0.1372±0.0007**
T+S+D	**0.8064±0.0010**	+1.24±0.69	0.4293±0.0040	0.1382±0.0018
QDCKT†	0.7916±0.0068	–0.24±0.45	0.4397±0.0056	0.1418±0.0024
RAGKT†	0.7907±0.0060	–0.33±0.52	0.4431±0.0029	0.1430±0.0016
LLM (no retrieval)	0.7920±0.0068	–0.21±0.45	0.4394±0.0057	0.1416±0.0024
MOOCRadar				
K	0.8307±0.0333	–2.79±2.75	0.3707±0.0275	0.1181±0.0098
T	0.8586±0.0190	–	0.3454±0.0180	0.1089±0.0067
T+S	**0.8703±0.0011**	+1.17±1.90	**0.3336±0.0045**	**0.1045±0.0018**
T+D	0.8635±0.0121	+0.49±0.69	0.3411±0.0128	0.1072±0.0046
T+S+D	0.8662±0.0102	+0.75±0.88	0.3365±0.0124	0.1053±0.0045
QDCKT†	0.8306±0.0332	–2.81±2.74	0.3709±0.0272	0.1181±0.0098
RAGKT†	0.8307±0.0332	–2.80±2.74	0.3707±0.0275	0.1181±0.0099
LLM (no retrieval)	0.8306±0.0332	–2.80±2.74	0.3708±0.0274	0.1181±0.0098

Δ = row condition − **T**. **Bold**: best mean within each dataset/metric. Yellow/blue: better/worse than **T**.

G.6 Reference Baseline Implementations

We did not locate official author-released code for QDCKT or RAGKT when preparing these experiments. We implement the difficulty-transfer components of QDCKT (Liu et al., 2024) and RAGKT (Ma et al., 2026) within AKCT, using the same item split, student split, training budget, and unseen-item validation rule as the main comparisons. Table 6 reports these shared-backbone component implementations. Each supplies a predicted correctness scalar in place of the text auxiliary token, for both history and query items in all splits. The scalar occupies the first coordinate of a six-dimensional compatibility vector, with the remaining coordinates zero, followed by a learned projection to 128 dimensions. Thus, the three scalar conditions share an input interface, but their projection parameter count differs from that of the 1024-dimensional text conditions.

Our QDCKT implementation uses the existing Source-supervised difficulty predictor, selecting its scalar readout by Source validation MAE: the expected class representative for XES3G5M and DBE-KT22, and the predicted class representative for MOOCRadar. The implementation uses the scalar-input interface above; the original difficulty-level lookup, Hann smoothing, and difficulty-consistency regularizer are omitted. Our RAGKT implementation trains a Source-only contrastive adapter over frozen text embeddings and retrieves five Source reference questions with their difficulty labels, excluding the query itself. Qwen3.6-35B-A3B then estimates the query's correctness probability; a matched control uses the same model and prompt without retrieved references. The original RAGKT full configuration was unavailable, so the retriever, prompts, and downstream architecture

are specified by our protocol. Upstream fitting and generation use seed 13; only KT varies over seeds 13, 29, and 47.

H Additional Mechanism Analyses

H.1 Equal-Budget Difficulty-Label Control

For each dataset, we train two encoders from the same seed-13 initialization for ten epochs using the same inputs, minibatch order, optimizer, learning rate, effective batch size of 32, and joint CE+MAE loss. One arm uses the original Source targets; the other uses one fixed permutation (seed 20260917) of the paired continuous and categorical targets among Source training items. The joint label distribution is preserved, and validation/test labels are unchanged. The permutation spans the full eligible Source set in each dataset. Initialization and per-epoch batch-order hashes agree between arms.

The eligible training sets contain 2,543, 143, and 338 items for XES3G5M, DBE-KT22, and MOOCRadar, corresponding to 800, 50, and 110 optimizer steps. Both arms export the final-epoch encoder. The six banks each train AKCT for 10,000 steps with seeds 13, 29, and 47, giving 18 downstream runs; the usual unseen-validation AUC rule selects the KT checkpoint. Reported standard deviations describe downstream seeds conditional on this upstream initialization and single label permutation.

Table 33: Probability errors for the equal-budget label control (mean ± standard deviation over three KT seeds). Lower is better.

Dataset	Metric	Original $\mathbf{T}$	Shuffled labels	Correct labels
XES3G5M	LogLoss	0.4744±0.0036	0.4735±0.0041	0.4730±0.0084
	Brier	0.1523±0.0013	0.1517±0.0014	0.1503±0.0017
DBE-KT22	LogLoss	0.5582±0.0074	0.5652±0.0186	0.5336±0.0140
	Brier	0.1866±0.0043	0.1871±0.0057	0.1735±0.0024
MOOCRadar	LogLoss	0.3948±0.0047	0.3913±0.0004	0.3842±0.0057
	Brier	0.1260±0.0028	0.1246±0.0006	0.1207±0.0026

Correct labels reduce mean LogLoss and Brier on all three datasets. All three seeds improve over shuffled labels for both metrics on DBE-KT22 and MOOCRadar; XES3G5M has one reverse-direction seed for each metric. AUC improves in all nine paired comparisons.

H.2 Neighborhood Difficulty Alignment

Each unseen item retrieves five Source neighbors sharing at least one KC; MOOCRadar additionally requires the same course. We compare cosine neighbors from normalized original and adapted encoder vectors with 100 random five-neighbor samples from the same eligible pool. All methods use identical query sets and candidate pools, and retrieval uses no evaluation correctness labels. Validation-item analysis uses Validation students; test-item analysis uses Test students. Each student's first valid response per item contributes to raw empirical correctness. Queries and candidates require at least 20 distinct students, with 50 as a sensitivity threshold.

For each query, we average the absolute correctness difference from its five neighbors, then average equally over queries. Table 34 reports the test results. Intervals use 2,000 paired query-item bootstrap samples, conditional on the representations and estimated correctness. Validation has 333/9/38 eligible queries at the main threshold; test has 995/31/146. The corresponding 50-student test sets contain 670/31/52 items.

Original semantic neighbors improve over random neighbors in all three main test comparisons. At the 50-student threshold, adapted-minus-original gaps are −.0170, −.0165, and +.0050 for XES3G5M, DBE-KT22, and MOOCRadar; the respective intervals are [−.0218,−.0121], [−.0294,−.0030], and [−.0115,.0216]. Using the neighbors' mean correctness as a direct predictor gives smaller adapted mean MAE at the 20-student threshold, but all three paired improvement intervals include zero.

Table 34: Test-item neighborhood correctness gaps; lower is better. Adapted representations use the original main-experiment checkpoint selection. Brackets give paired 95% intervals.

Dataset	Random	T	T+D	Adapted − original
XES3G5M	.1719	.1651	.1509	−.0142[−.0187,−.0096]
DBE-KT22	.1697	.1465	.1301	−.0165[−.0294,−.0030]
MOOCRadar	.2208	.1994	.1883	−.0111[−.0238,.0019]

We also repeat the neighborhood analysis with the final-epoch banks from the equal-budget control. Correct-minus-shuffled gaps are −.0143 on XES3G5M, −.0066 on DBE-KT22, and +.0025 on MOOCRadar, with intervals [−.0189,−.0100], [−.0346,.0210], and [−.0096,.0143]. Neighborhood alignment improves most clearly on XES3G5M, whereas the KT AUC benefit of correct labels occurs on all three datasets. Local correctness similarity captures one aspect of the learned representation; its changes vary across datasets.

H.3 Controlled Exposure Protocol and Uncertainty

We fix 50 high-support XES3G5M items and vary distinct training-student responses per item over $n \in \{0, 10, 30, 100, 300, 1000, 3000\}$. AKCT with **K**, **T**, or **T**+**D** and seeds 13, 29, and 47 gives 63 runs. All runs end at step 30,000 without target-dependent checkpoint selection. Within each condition and seed, initialization, background examples, and evaluation labels remain fixed. Response subsets are nested across doses: a shared response retains the same training step and slot, and the final exposure step for each target remains fixed across positive doses. Every selected response is from a distinct student for that item and contributes once; targets never appear in history. Only zero exposure leaves the target items unseen.

Each step reserves 48 background and 16 target slots, with a fixed loss denominator of 64; unused target slots contribute zero loss. Background supervision therefore remains at 1,440,000 examples. Target supervision increases to 150,000 examples at the highest dose, raising total supervision by 10.42%; the design fixes steps and background evidence while varying target supervision. The reported seven-dose results use a fixed 50-item cohort and 30,000 training steps.

A standardized Ridge probe reads the 128-dimensional second-decoder MASK-token representation at position zero with empty history and unknown current response. It is fitted on 1,233 Source items, tuned on 308 others over regularization strengths $\{0.1, 1, 10, 100, 1000, 10000\}$, and refitted on their union. The 50 targets are used only for evaluation. Labels come from the original test-student population, using each student's first valid response per item without smoothing, and remain fixed across doses. Neither KT nor upstream adaptation uses this label population. The probe measures transfer of a Source-fitted linear readout.

Increasing exposure from zero to 3,000 reduces MAE from .0919 to .0628 for **K**, .0970 to .0393 for **T**, and .0781 to .0468 for **T**+**D**. Spearman correlation rises from .2202 to .6195, .1330 to .8407, and .3757 to .8374, respectively. The paired item-bootstrap 95% intervals for the MAE reductions are [.0191,.0401], [.0429,.0744], and [.0216,.0424]. The zero-exposure MAE reduction from adaptation is .0189 with interval [.0029,.0356]. At 3,000 responses, unadapted text has the lower mean MAE, but the adapted advantage interval [-.0162,.0014] spans zero. All three conditions exceed the Source-median constant at the highest dose throughout their corresponding item-bootstrap intervals.

Error bars report the sample standard deviation across three KT seeds. Item-bootstrap intervals resample the 50 items jointly across paired seeds, condition on the fixed checkpoints and probes, and are not corrected for multiple comparisons. All target items receive the same dose, so their own responses and transfer through shared KCs can both contribute. The analysis concerns high-support items and the accessibility of correctness information under an empty-history linear readout.

I Supervision Ablations and Generalization Stability

I.1 Protocol

All experiments use AKCT with question text (T) on the three datasets. We hold the encoder architecture, supervision items within a split, optimizer, and ten-epoch upstream budget fixed. For every objective, frozen pre-head encoder representations from the final epoch are used as KT inputs. The classification and regression heads provide supervision during encoder adaptation. KT is trained for 10,000 steps and selected by unseen-validation AUC. Downstream seeds are 13, 29, and 47. The original-split T and joint-loss endpoints reuse the matched runs from Appendix H.1; the other conditions add 24 encoder adaptations and 90 KT runs, yielding 36 settings and 108 KT results in total. This fixed-endpoint comparison is separate from the upstream validation-selected representations in Table 2.

We report within-setting means and sample standard deviations, with AUC changes paired by downstream seed against T on the same split. Item-split and upstream-seed variations are reported separately. All probe values use the final decoder MASK state and the protocol in Appendix F.

I.2 Classification and Regression Supervision

CE-only and MAE-only retain the joint prediction-head architecture and change only the active loss. The original item split, upstream seed 13, input content, supervised item set, initialization, and minibatch order are shared. Table 35 compares both individual objectives with joint supervision and unadapted text.

Regression alone gives the highest unseen-item AUC on XES3G5M. Classification alone gives the highest AUC on DBE-KT22 and MOOCRadar, where regression alone is close to or below unadapted text. Joint supervision improves mean AUC on all three datasets, while the preferred individual objective varies with the dataset. The probe results follow the same ordering among the adapted conditions: regression gives the lowest MAE on XES3G5M, and classification gives the lowest on DBE-KT22 and MOOCRadar.

Table 35: Supervision objectives on unseen items. Probe MAE is in percentage points.

Condition	AUC ↑	Δ (pp)	LogLoss ↓	Brier ↓	Probe MAE ↓
XES3G5M					
T	0.6949±0.0043	–	0.4744±0.0036	0.1523±0.0013	13.41±0.10
CE-only	0.7081±0.0015	+1.32±0.35	0.4667±0.0021	0.1495±0.0009	13.30±0.03
MAE-only	0.7138±0.0009	+1.89±0.50	0.4689±0.0022	0.1493±0.0004	12.79±0.03
CE+MAE	0.7085±0.0020	+1.36±0.51	0.4730±0.0084	0.1503±0.0017	13.03±0.06
DBE-KT22					
T	0.6594±0.0036	–	0.5582±0.0074	0.1866±0.0043	14.74±0.94
CE-only	0.7020±0.0055	+4.26±0.36	0.5306±0.0128	0.1742±0.0023	12.41±0.22
MAE-only	0.6600±0.0028	+0.06±0.10	0.5721±0.0234	0.1877±0.0067	15.57±0.21
CE+MAE	0.7010±0.0045	+4.17±0.37	0.5336±0.0140	0.1735±0.0024	12.54±0.10
MOOCRadar					
T	0.7954±0.0058	–	0.3948±0.0047	0.1260±0.0028	14.71±0.72
CE-only	0.8207±0.0019	+2.53±0.77	0.3680±0.0014	0.1156±0.0006	13.00±0.14
MAE-only	0.7932±0.0038	−0.22±0.20	0.3925±0.0011	0.1252±0.0006	15.06±0.38
CE+MAE	0.8083±0.0045	+1.29±0.28	0.3842±0.0057	0.1207±0.0026	13.75±0.38

Δ = row AUC − matched T AUC (pp). Yellow/blue: higher/lower than T. Values are mean ± standard deviation over KT seeds.

I.3 Difficulty-Target Construction

We compare the five-group balanced and smoothed target with pooled correctness and pooled correctness smoothed toward the Source prior with strength 10. All variants use exactly the same eligible supervision items and joint loss. XES3G5M uses the fixed 0.75/0.85 class boundaries; DBE-

KT22 and MOOCRadar use Source-training terciles for each target. Only Source labels determine these boundaries.

All three constructions improve mean unseen-item AUC over unadapted text. Pooled correctness performs best on XES3G5M and MOOCRadar, while smoothed pooled correctness performs best on DBE-KT22. The benefit of difficulty supervision therefore extends to simple pooled response statistics. Probability-error metrics vary across constructions; for example, smoothed pooled supervision on XES3G5M improves AUC while slightly increasing LogLoss.

Table 36: Difficulty-target construction on unseen items, using the joint loss. Probe MAE is in percentage points.

Condition	AUC ↑	Δ (pp)	LogLoss ↓	Brier ↓	Probe MAE ↓
XES3G5M					
T	0.6949±0.0043	–	0.4744±0.0036	0.1523±0.0013	13.41±0.10
Balanced	0.7085±0.0020	+1.36±0.51	0.4730±0.0084	0.1503±0.0017	13.03±0.06
Pooled	0.7093±0.0004	+1.43±0.46	0.4686±0.0028	0.1494±0.0004	12.59±0.03
Smoothed pooled	0.7062±0.0003	+1.12±0.40	0.4763±0.0047	0.1514±0.0011	12.92±0.02
DBE-KT22					
T	0.6594±0.0036	–	0.5582±0.0074	0.1866±0.0043	14.74±0.94
Balanced	0.7010±0.0045	+4.17±0.37	0.5336±0.0140	0.1735±0.0024	12.54±0.10
Pooled	0.7026±0.0024	+4.33±0.49	0.5483±0.0111	0.1750±0.0020	13.02±0.42
Smoothed pooled	0.7052±0.0039	+4.58±0.28	0.5366±0.0147	0.1736±0.0022	12.54±0.12
MOOCRadar					
T	0.7954±0.0058	–	0.3948±0.0047	0.1260±0.0028	14.71±0.72
Balanced	0.8083±0.0045	+1.29±0.28	0.3842±0.0057	0.1207±0.0026	13.75±0.38
Pooled	0.8168±0.0019	+2.14±0.71	0.3754±0.0017	0.1185±0.0013	12.63±0.28
Smoothed pooled	0.8161±0.0018	+2.07±0.75	0.3757±0.0011	0.1180±0.0005	12.62±0.31

Δ = row AUC − matched T AUC (pp). Yellow/blue: higher/lower than T. Values are mean ± standard deviation over KT seeds.

I.4 Item-Split Stability

Two additional item partitions use split seeds 20260922 and 20260923, denoted splits 1 and 2. Student partitions and the number of items in each partition remain fixed; MOOCRadar also preserves within-course partition sizes. Each split recomputes Source-based student groups, smoothing priors, eligible supervision items, and difficulty thresholds, and retrains the encoder with upstream seed 13. Joint supervision has positive mean unseen-item AUC gains on all nine dataset–split combinations (Table 37). The six new comparisons range from 0.71 to 3.84 percentage points.

Table 37: Item-split stability: unseen-item AUC under joint supervision.

Dataset	Item split	T	CE+MAE	Δ (pp)
XES3G5M	0	0.6949±0.0043	0.7085±0.0020	+1.36±0.51
	1	0.7059±0.0028	0.7144±0.0016	+0.85±0.38
	2	0.6861±0.0002	0.6932±0.0014	+0.71±0.15
DBE-KT22	0	0.6594±0.0036	0.7010±0.0045	+4.17±0.37
	1	0.6833±0.0174	0.7217±0.0030	+3.84±1.63
	2	0.6511±0.0087	0.6696±0.0084	+1.85±1.06
MOOCRadar	0	0.7954±0.0058	0.8083±0.0045	+1.29±0.28
	1	0.7854±0.0013	0.8120±0.0031	+2.67±0.18
	2	0.7867±0.0029	0.8092±0.0025	+2.25±0.43

Δ = row AUC − matched T AUC (pp). Yellow/blue: higher/lower than T. Values are mean ± standard deviation over KT seeds.

I.5 Upstream-Initialization Stability

We vary the encoder-adaptation seed over 13, 29, and 47 on the original item split, retaining three downstream KT seeds for each bank. The same unadapted text baseline is reused across encoder seeds. Mean AUC gains remain positive in all nine dataset–encoder-seed combinations, ranging from 1.29 to 4.17 percentage points (Table 38). Each row's standard deviation describes downstream training with that particular bank.

Table 38: Upstream-initialization stability on the original item split: unseen-item AUC.

Dataset	Encoder seed	T	CE+MAE	Δ (pp)
XES3G5M	13	0.6949±0.0043	0.7085±0.0020	+1.36±0.51
	29	0.6949±0.0043	0.7146±0.0004	+1.97±0.39
	47	0.6949±0.0043	0.7128±0.0011	+1.78±0.53
DBE-KT22	13	0.6594±0.0036	0.7010±0.0045	+4.17±0.37
	29	0.6594±0.0036	0.6841±0.0038	+2.48±0.18
	47	0.6594±0.0036	0.7004±0.0044	+4.10±0.10
MOOCRadar	13	0.7954±0.0058	0.8083±0.0045	+1.29±0.28
	29	0.7954±0.0058	0.8178±0.0024	+2.24±0.40
	47	0.7954±0.0058	0.8201±0.0008	+2.47±0.63

Δ = row AUC − matched T AUC (pp). Yellow/blue: higher/lower than T. Values are mean ± standard deviation over KT seeds.

I.6 Seen-Item Results

The corresponding seen-item metrics are reported below. Conditions share the same identifiers as the preceding ablations. These tables retain variation across settings explicitly, including cases where unseen-item and seen-item changes differ.

Table 39: XES3G5M: seen-item results for all ablation settings.

Setting	AUC ↑	Δ (pp)	LogLoss ↓	Brier ↓
CE-only	0.7652±0.0062	+5.58±0.91	0.4360±0.0039	0.1398±0.0013
Joint / seed 13	0.7763±0.0020	+6.69±1.60	0.4310±0.0021	0.1376±0.0006
Joint / seed 29	0.7775±0.0016	+6.81±1.61	0.4304±0.0022	0.1374±0.0004
Joint / seed 47	0.7746±0.0018	+6.52±1.55	0.4317±0.0023	0.1379±0.0006
MAE-only	0.7823±0.0004	+7.29±1.51	0.4263±0.0023	0.1357±0.0006
Original T	0.7094±0.0149	–	0.4677±0.0079	0.1502±0.0025
Pooled	0.7712±0.0007	+6.18±1.55	0.4326±0.0004	0.1384±0.0000
Smoothed pooled	0.7750±0.0011	+6.56±1.59	0.4320±0.0021	0.1379±0.0006
Split 1 / joint	0.7771±0.0009	+3.77±0.15	0.4341±0.0033	0.1392±0.0013
Split 1 / T	0.7393±0.0023	–	0.4577±0.0030	0.1470±0.0010
Split 2 / joint	0.7780±0.0034	+5.96±2.20	0.4259±0.0056	0.1356±0.0018
Split 2 / T	0.7184±0.0207	–	0.4600±0.0118	0.1469±0.0037

Δ = row AUC − matched T AUC (pp). Yellow/blue: higher/lower than T. Values are mean ± standard deviation over KT seeds.

Table 40: DBE-KT22: seen-item results for all ablation settings.

Setting	AUC ↑	Δ (pp)	LogLoss ↓	Brier ↓
CE-only	0.8060±0.0012	+1.20±0.70	0.4279±0.0033	0.1371±0.0006
Joint / seed 13	0.8066±0.0014	+1.26±0.70	0.4270±0.0031	0.1367±0.0006
Joint / seed 29	0.8035±0.0004	+0.95±0.63	0.4284±0.0022	0.1372±0.0004
Joint / seed 47	0.8047±0.0002	+1.07±0.59	0.4281±0.0025	0.1370±0.0006
MAE-only	0.7942±0.0018	+0.02±0.69	0.4377±0.0016	0.1406±0.0003
Original T	0.7940±0.0059	–	0.4381±0.0076	0.1410±0.0030
Pooled	0.8058±0.0006	+1.17±0.57	0.4276±0.0032	0.1368±0.0007
Smoothed pooled	0.8062±0.0010	+1.22±0.65	0.4277±0.0033	0.1369±0.0006
Split 1 / joint	0.7863±0.0023	+2.00±1.40	0.4351±0.0034	0.1394±0.0009
Split 1 / T	0.7663±0.0119	–	0.4474±0.0074	0.1432±0.0029
Split 2 / joint	0.8086±0.0033	+0.27±0.22	0.4395±0.0043	0.1429±0.0014
Split 2 / T	0.8059±0.0026	–	0.4403±0.0026	0.1431±0.0007

Δ = row AUC − matched **T** AUC (pp). Yellow/blue: higher/lower than **T**. Values are mean ± standard deviation over KT seeds.

Table 41: MOOCRadar: seen-item results for all ablation settings.

Setting	AUC ↑	Δ (pp)	LogLoss ↓	Brier ↓
CE-only	0.8700±0.0025	+1.13±1.95	0.3322±0.0034	0.1038±0.0012
Joint / seed 13	0.8647±0.0110	+0.60±0.80	0.3396±0.0112	0.1067±0.0040
Joint / seed 29	0.8620±0.0091	+0.33±1.33	0.3406±0.0101	0.1069±0.0037
Joint / seed 47	0.8670±0.0065	+0.84±2.18	0.3352±0.0075	0.1050±0.0028
MAE-only	0.8545±0.0165	−0.42±0.57	0.3481±0.0166	0.1098±0.0062
Original T	0.8586±0.0190	–	0.3454±0.0180	0.1089±0.0067
Pooled	0.8657±0.0068	+0.70±2.09	0.3382±0.0075	0.1062±0.0028
Smoothed pooled	0.8668±0.0073	+0.81±2.27	0.3357±0.0079	0.1053±0.0029
Split 1 / joint	0.8720±0.0011	+0.31±0.47	0.3264±0.0041	0.1018±0.0015
Split 1 / T	0.8689±0.0051	–	0.3278±0.0051	0.1022±0.0016
Split 2 / joint	0.8720±0.0026	−0.05±0.28	0.3201±0.0025	0.0986±0.0007
Split 2 / T	0.8725±0.0002	–	0.3200±0.0016	0.0991±0.0006

Δ = row AUC − matched **T** AUC (pp). Yellow/blue: higher/lower than **T**. Values are mean ± standard deviation over KT seeds.

J Relationship to Prior Difficulty and Generalization Work

Table 42 summarizes the research questions and representation interfaces most relevant to the present study.

Table 42: Selected related directions and the dimension emphasized in this study. This is a comparison of research questions and interfaces, not a numerical benchmark.

Work	Relevant prior contribution	Dimension examined here
PEBG (Liu et al., 2020)	Pretraining question embeddings from difficulty and question–skill relations	Content-encoder transfer to held-out items under a fixed protocol
DCL4KT (Lee et al., 2024)	Difficulty-focused KT and language-model difficulty prediction for unseen questions	Pre-head vectors; factorial content/adaptation controls; three separate measurements
QDCKT (Liu et al., 2024)	Difficulty-consistent representations for unseen-item KT	Difficulty-oriented content adaptation and empirical readout analysis
AdaptKT / DGKT (Cheng et al., 2022; Xie et al., 2026)	Cross-domain adaptation and generalization	Same-dataset item holdout with independent student partitions
CLST (Jung et al., 2025)	Generative-language-model KT in limited-data and cross-domain conditions	Item-specific absence of training response support
MathBERT unseen-item KT (Alotaibi et al., 2026)	Semantic embeddings for unseen-item prediction	Difficulty adaptation and the distinction between scalar and vector utility