# Bayesian Deep Knowledge Tracing: A Hybrid Bayesian–Neural Approach to Student Modeling

M. E. Narovanjanahary, J. C. Ralaivao, A. Ratovondrahona,
T. Mahatody, and N. M. V. Ravonimanantsoa

***Abstract*— Adaptive tutoring systems depend on accurate, trustworthy estimates of student knowledge—yet most existing knowledge tracing models produce point predictions that carry no measure of their own reliability. When a system cannot distinguish a confident prediction from an uncertain one, it risks triggering interventions that are pedagogically inappropriate, particularly in early or data-scarce stages of a learning sequence. This work addresses that limitation by introducing Bayesian Deep Knowledge Tracing (BDKT), a hybrid framework in which latent knowledge states are represented as posterior distributions over a multi-skill space and propagated through time via an LSTM-based transition model. Unlike deterministic tracers such as DKT, and unlike attention- or graph-based extensions that improve predictive accuracy without addressing calibration, BDKT couples sequential neural modeling with Monte Carlo dropout to quantify both epistemic uncertainty—stemming from limited data—and aleatoric uncertainty inherent in student responses. A graph-informed multivariate Gaussian covariance further encodes prerequisite dependencies among skills, a structural dimension absent from prior neural KT models. Evaluated on a large-scale public dataset of over 500 000 interactions spanning 30 mathematical skills, BDKT yields statistically significant improvements over DKT in discrimination (AUC $0.87$ vs. $0.83$, $p<0.01$) and reduces expected calibration error by nearly half (ECE $0.04$ vs. $0.07$, $p<0.01$), supporting the practical value of uncertainty-aware student modeling.**

## I. INTRODUCTION

Knowledge tracing (KT) estimates and forecasts learner mastery from sequences of learning interactions. Classical Bayesian Knowledge Tracing (BKT) models mastery as a binary latent variable updated through Bayesian inference [2], while Deep Knowledge Tracing (DKT) replaces handcrafted assumptions with recurrent neural networks learned from data [4]. Although DKT improves predictive accuracy, its deterministic outputs carry no information about confidence, leaving adaptive systems unable to distinguish between reliable predictions and uncertain ones.

Recent extensions address complementary shortcomings of DKT yet none resolves its calibration deficit. SAKT [6] improves sensitivity to relevant history via self-attention but outputs deterministic logits; DKVMN [7] separates concept keys from dynamic mastery values for better interpretability, yet reads those values deterministically; GKT [8] encodes prerequisite structure as a directed graph but likewise produces point estimates. Each improves *what* is predicted while leaving open *how confident* that prediction should be—a distinction consequential in adaptive tutoring, where a system that cannot quantify uncertainty risks triggering interventions that are inappropriate precisely when the learner's state is most ambiguous [16]. BDKT addresses this gap by representing latent knowledge states as posterior distributions that propagate uncertainty through every step of the sequential reasoning process. The importance of reliable, calibrated predictions in applied educational AI is further underlined by recent work on explainable predictive modeling [24], Bayesian calibration methods [18], and AI-driven human-computer interaction [23].

This paper makes three key contributions:

1) A probabilistic BDKT formulation that maintains a multivariate Gaussian over skill states, informed by a prerequisite graph, and employs Monte Carlo dropout [17] to quantify both epistemic and aleatoric uncertainty;
2) A comprehensive empirical evaluation on a large-scale public dataset [22] (500 k interactions across 30 skills) with rigorous statistical analysis including ablation studies;
3) Demonstration of statistically significant improvements in both discrimination (AUC) and calibration (ECE) relative to BKT, DKT, IRT, SAKT, DKVMN, and GKT.

## II. RELATED WORK

This section reviews KT models with emphasis on their modeling assumptions, interpretability, and ability to quantify uncertainty.

### A. Item Response Theory (IRT)

Item Response Theory estimates the probability of a correct response as a function of learner ability and item difficulty. In the Rasch model [1], the probability is expressed as:

$$P(\text{correct}) = \frac{1}{1 + e^{-(\theta - b)}} \tag{1}$$

where $\theta$ is learner ability and $b$ is item difficulty. A common extension introduces discrimination $a$:

$$P(\text{correct}) = \frac{1}{1 + e^{-a(\theta - b)}}. \tag{2}$$

IRT is typically static and does not explicitly model temporal learning dynamics.

### B. Bayesian Knowledge Tracing

Introduced in 1994 [2], BKT models mastery as a binary latent state that transitions probabilistically with practice.

The model parameters include initial mastery $P(L_0)$, learning rate $P(T)$, slip $P(S)$, and guess $P(G)$ [3]. Posterior updates follow standard Bayes rules:

$$P(L_n \mid \text{Correct}) = \frac{P(L_{n-1})(1-P(S))}{P(L_{n-1})(1-P(S))+(1-P(L_{n-1}))P(G)} \quad (3)$$

$$P(L_n \mid \text{Incorrect}) = \frac{P(L_{n-1})P(S)}{P(L_{n-1})P(S)+(1-P(L_{n-1}))(1-P(G))} \quad (4)$$

followed by a transition step accounting for learning over time.

### C. Deep Knowledge Tracing

DKT [4] uses recurrent neural networks (typically LSTMs [5]) to capture nonlinear dependencies in learning sequences. At each interaction $X_n$, the model updates its latent state:

$$H_n = f(X_n, H_{n-1}) \quad (5)$$

and predicts the probability of success:

$$\hat{y}_n = \sigma(WH_n). \quad (6)$$

Subsequent work [19] explored deeper LSTM stacking and showed that gains plateau quickly, suggesting that architectural depth alone does not resolve the fundamental limitation of deterministic outputs.

### D. Extensions of DKT

Modern KT models incorporate attention and structured representations. SAKT [6] applies self-attention to weight diagnostically relevant past interactions, but produces deterministic scores without any confidence measure and ignores inter-skill dependencies. DKVMN [7] separates static concept keys from dynamic mastery values in an external memory, improving per-skill interpretability; yet mastery values are updated and read deterministically, with no distributional representation of uncertainty. GKT [8] encodes prerequisite relationships as a directed graph and propagates evidence across related skills—improving prediction for poorly observed concepts—but equally outputs point estimates. Yin et al. [21] further showed that graph-based tracers are sensitive to missing interaction data, a limitation that calibrated uncertainty estimates would help flag. Deeper LSTM stacking [19] yields only marginal gains, confirming that architectural depth alone does not compensate for the absence of calibrated confidence.

*Gap:* No existing KT model simultaneously provides (i) a full distributional knowledge-state representation, (ii) a deep sequential transition model for nonlinear dynamics, and (iii) calibrated uncertainty at each prediction step. Probabilistic student modeling has been explored in adjacent domains [9], [12], but not unified into a sequential neural KT framework. BDKT is designed to satisfy all three requirements jointly.

TABLE I
COMPARATIVE ANALYSIS OF REPRESENTATIVE KT MODELS

| Model | Attention | Uncertainty | Graph | Interpretability |
|---|---|---|---|---|
| BKT | No | Partial | No | High |
| SAKT | Yes | No | No | Moderate |
| DKVMN | No | No | No | High |
| GKT | No | No | Yes | Moderate |
| **BDKT** | Yes | **Yes** | Yes | Moderate |

Table I makes this gap visible at a glance. BKT maintains a binary mastery posterior—a form of uncertainty—but cannot represent continuous, correlated, multi-skill knowledge states. SAKT, DKVMN, and GKT all improve discrimination through richer architectures, yet the *Uncertainty* column remains empty for each. BDKT extends BKT's probabilistic commitment to the full depth of a neural sequential model, combining what each prior approach handles well while addressing what all of them leave aside.

## III. BAYESIAN DEEP KNOWLEDGE TRACING

This section presents a hybrid approach that combines the uncertainty quantification of Bayesian methods with the representational power of deep neural networks. The core probabilistic framework is first established, then neural networks for modeling temporal transitions are described.

### A. Notation and Problem Setup

Let $t \in \{1, \ldots, T\}$ index student interactions in chronological order. Each interaction $t$ consists of an observation $\mathbf{O}_t$ (e.g., correctness, response time) and contextual information $\mathbf{A}_t$ (e.g., item and skill identifiers).

The central idea of BDKT is to represent student knowledge as a probability distribution rather than a single point estimate. Specifically, we model the latent knowledge state as a real-valued vector $\mathbf{K}_t \in \mathbb{R}^n$ over $n$ skills, where each element represents knowledge on the logit scale. Actual mastery probabilities are recovered through sigmoid transformation: $P(\text{mastery of skill } i) = \sigma(K_{t,i})$.

### B. Probabilistic Knowledge Representation

Unlike traditional neural knowledge tracers that output deterministic predictions, BDKT maintains a full probability distribution over student knowledge states. This distribution captures both the estimated student mastery and the uncertainty about that estimate.

Formally, we model the posterior distribution over knowledge states given all past observations as a multivariate Gaussian:

$$P(\mathbf{K}_t \mid \mathbf{O}_{1:t}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t). \quad (7)$$

The mean vector $\boldsymbol{\mu}_t$ represents the estimated student proficiencies, while the covariance matrix $\boldsymbol{\Sigma}_t$ encodes both the uncertainty in the estimates and correlations between different skills.

*Covariance informed by a prerequisite graph:* Let $G = (\mathcal{C}, E)$ be a prerequisite graph. We define

$$\Sigma_t[i,j] = \begin{cases} \sigma_i^2 & i = j, \\ \rho_{ij}\,\sigma_i\sigma_j\,w(c_i,c_j) & (c_i,c_j) \in E, \\ \rho_{ij}\,\sigma_i\sigma_j\,\alpha & \text{otherwise}, \end{cases} \quad (8)$$

where $\sigma_i$ represents the per-skill uncertainty scale, $\rho_{ij}$ is a learned correlation coefficient, $w(c_i, c_j) \in [0.5, 1]$ weights prerequisite relationship strength, and $\alpha \ll 1$ controls background correlations between unrelated skills.

The initial state follows

$$P(\mathbf{K}_0) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad (9)$$

where $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are population priors.

### C. Dynamic Transitions with Forgetting

To model forgetting and learning, we use

$$\mathbf{K}_{t+1} = (1 - \boldsymbol{\lambda}_t) \odot \mathbf{K}_t + \boldsymbol{\lambda}_t \odot f(\mathbf{K}_t, \mathbf{A}_t, \mathbf{C}_t) + \boldsymbol{\varepsilon}_t, \quad (10)$$

where $\boldsymbol{\lambda}_t \in [0,1]^n$ balances retention and update, $\mathbf{C}_t$ is a learner/context feature vector, and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$ is process noise.

$$\boldsymbol{\lambda}_t = \sigma\big(W_\lambda[\mathbf{K}_t; \mathbf{A}_t; \Delta t_t; \mathbf{d}] + \mathbf{b}_\lambda\big), \quad (11)$$

where $\Delta t_t$ is time since last practice and $\mathbf{d}$ denotes difficulty-related features.

Here, $\mathbf{A}_t$ represents the encoded interaction vector at time $t$ (including item identifier and response), $\mathbf{C}_t$ contains auxiliary contextual features (e.g., session information, learning objectives), and $\mathbf{Q}_t$ specifies the transition noise covariance matrix.

*Sequential transition model:* The transition function employs an LSTM to capture complex learning dynamics:

$$f_\theta(\mathbf{K}_t, \mathbf{A}_t, \mathbf{C}_t) = \mathrm{LSTM}_\theta([\mathbf{K}_t; \mathbf{A}_t; \mathbf{C}_t]). \quad (12)$$

*Uncertainty quantification through Monte Carlo dropout:* BDKT distinguishes between two types of uncertainty. *Epistemic uncertainty* reflects our ignorance about the true model parameters $\theta$ and decreases as we observe more data. We approximate this using Monte Carlo dropout [17], where we randomly mask different neural network connections during each forward pass. *Aleatoric uncertainty* captures inherent randomness in student responses and learning processes, modeled through observation and process noise terms in our probabilistic framework.

### D. Multimodal Observation Model

Binary correctness alone is an incomplete observation: response time signals cognitive effort (a slow correct answer may reflect reconstruction rather than mastery), hint requests reveal metacognitive difficulty perception, and temporal gaps modulate expected forgetting [10], [20]. Treating these signals as equivalent forces the model to conflate observationally distinct interactions. BDKT therefore factorizes the observation likelihood across $M$ modalities:

$$P(o_t \mid \mathbf{K}_t, \mathbf{A}_t) = \prod_{m=1}^{M} P_m\big(o_t^{(m)} \mid \varphi_m(\mathbf{K}_t, \mathbf{A}_t)\big). \quad (13)$$

In the present implementation, the three modalities are correctness (Bernoulli), response time (log-normal), and behavioral/contextual signals including hint counts [10].

### E. Variational Objective with Regularisers

Exact inference is intractable, so we maximize an evidence lower bound (ELBO) with additional regularization:

$$\mathcal{L} = \mathbb{E}_{q_\phi}[\log P(\mathbf{O}_{1:T} \mid \mathbf{K}_{1:T})] - \beta\, \mathrm{D}_{\mathrm{KL}}\big[q_\phi(\mathbf{K}_{1:T}) \| P(\mathbf{K}_{1:T})\big] \quad (14)$$

$$- \gamma\, \mathcal{R}_{\text{mono}} - \delta\, \mathcal{R}_{\text{transfer}}. \quad (15)$$

The regularization terms $\mathcal{R}_{\text{mono}}$ penalize implausible short-term mastery regressions, while $\mathcal{R}_{\text{transfer}}$ encourage learning transfer consistent with prerequisite relationships.

### F. Hierarchical Attention Mechanism

Hierarchical attention can be optionally applied across concepts and input sources [11]:

$$\boldsymbol{\alpha}_t^{\text{concept}} = \mathrm{softmax}\big(\mathrm{score}(\mathbf{h}_t, \{c_i\})\big), \quad (16)$$

$$\boldsymbol{\alpha}_{t,i}^{\text{source}} = \mathrm{softmax}\big(\mathrm{score}(\mathbf{h}_t^{(i)}, \{s_j\})\big), \quad (17)$$

leading to

$$\mathbf{h}_t^{\text{final}} = \sum_i \alpha_{t,i}^{\text{concept}} \Big( \sum_j \alpha_{t,i,j}^{\text{source}}\, s_j^{(i)} \Big). \quad (18)$$

### G. BDKT Architecture

Figure 1 summarizes the information flow. At each step $t$, multimodal inputs ($r_t$, $\tau_t$, $h_t^{\text{hint}}$) are fused into $x_t$ and fed with context $\mathbf{A}_t$ into the Gaussian knowledge state $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. An `LSTM+MC` cell updates hidden state $h_t$ under Monte Carlo dropout, sampling 20 masks per forward pass; the mean output across passes gives $\hat{y}_t$ while variance quantifies epistemic uncertainty.

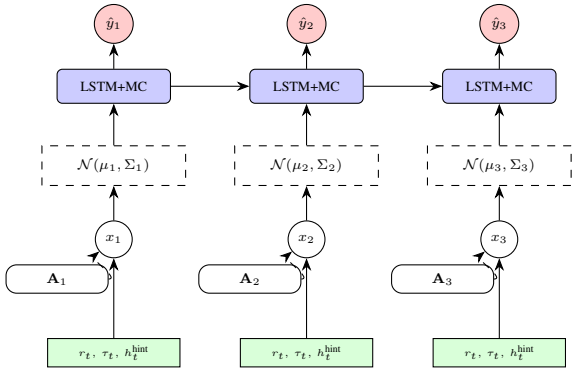


Fig. 1. BDKT architecture. Green boxes: multimodal inputs (correctness $r_t$, response time $\tau_t$, hints $h_t^{\text{hint}}$). Dashed boxes: probabilistic knowledge states $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. Blue boxes: LSTM cells with Monte Carlo (MC) dropout for epistemic uncertainty. Red circles: predictive distributions $\hat{y}_t$.

## IV. DATASET

### A. Data Source and Collection

BDKT is evaluated on a large-scale educational dataset publicly available on Kaggle [22]. The dataset was originally collected from an adaptive learning platform and represents authentic student-system interactions over a period of six months. The corpus contains 500 952 timestamped interactions from 4 000 students across 6 000 unique learning items covering 30 distinct mathematical skills.

### B. Data Structure and Features

Each record includes anonymized student and session identifiers, item ID with skill tag(s) and estimated difficulty, binary correctness, attempt number (1–5), hint requests, response timestamp, and response time in milliseconds. Students contribute 50–200 interactions each (median: 125) across 2 100 sessions; core arithmetic skills represent 40% of interactions and advanced topics 15%. Table II summarizes corpus dimensions.

TABLE II
DATASET STATISTICS

| Dimension | Count |
|---|---|
| Students | 4,000 |
| Items | 6,000 |
| Skills | 30 |
| Interactions | 500,952 |
| Sessions | 2,100 |
| Processed sequences | 3,847 |
| Correct responses | 312,595 (62.4%) |

### C. Preprocessing Pipeline

Duplicates and records with missing critical fields are removed (<0.5%). Items spanning multiple skills receive multi-hot encoding [15] yielding 30-dimensional skill vectors. Interactions are sorted chronologically per student, segmented into windows of length $L$=100 with 20% overlap, and zero-padded for shorter sequences. Response times and temporal gaps are log-normalized: $t' = \log(1 + t/1000)$. This yields 3 847 sequences (avg. length 89.3), with train/validation/test splits preserving temporal ordering within each student's history.

## V. EXPERIMENTAL SETUP

### A. Model Architecture

BDKT employs a two-layer LSTM with hidden dimensionality $d_h = 128$ and input embedding dimensionality $d_e = 64$. The covariance matrix $\mathbf{\Sigma}_t$ has rank $r = 32$ to balance expressiveness and computational efficiency. Monte Carlo dropout with rate $p = 0.2$ is applied to both LSTM layers and the final prediction layer.

### B. Training Procedure

A stratified five-fold cross-validation split by student ID ensures no data leakage. All hyperparameters are selected by grid search on a held-out 20% validation partition. Training uses Adam ($\beta_1$=0.9, $\beta_2$=0.999, $\epsilon$=$10^{-8}$) with learning rate $\eta$=$3\times10^{-4}$ (cosine annealing), batch size 256 (32 sequences $\times$ 8 gradient accumulation steps), weight decay $\lambda$=$10^{-4}$ on non-bias parameters, and $\ell_2$ gradient clipping at 5.0. The full objective is:

$$\mathcal{L}_{\text{total}} = -\text{ELBO} + \beta \cdot \text{D}_{\text{KL}} + \gamma \cdot \mathcal{R}_{\text{mono}} + \delta \cdot \mathcal{R}_{\text{transfer}} \quad (19)$$

with $\beta$=1.0, $\gamma$=0.05, $\delta$=0.1. Training stops after 20 epochs or when validation AUC plateaus for 5 consecutive epochs.

### C. Inference Details

Predictive probabilities are computed by averaging 20 stochastic forward passes with dropout enabled. For computational efficiency during evaluation, we use a fixed random seed across all Monte Carlo samples to ensure reproducible uncertainty estimates. Expected Calibration Error uses 15 equally-spaced confidence bins $[0, 0.067, 0.133, \ldots, 1.0]$ following [16].

### D. Baseline Implementation

Baselines share identical preprocessing and evaluation: BKT (classical four-parameter EM [2]), DKT (two-layer LSTM, $d_h$=200, dropout $p$=0.1 [4]), and IRT (Rasch model, L-BFGS). DKT uses the same optimizer as BDKT but without uncertainty quantification or regularization terms.

## VI. RESULTS

We report performance across five-fold cross-validation, including means, standard deviations, and statistical significance tests using paired t-tests against the DKT baseline.

TABLE III
PERFORMANCE RESULTS (MEAN $\pm$ STD, 5-FOLD CV)

| Model | AUC | ACC | RMSE | ECE |
|---|---|---|---|---|
| BKT | 0.69 ± 0.02 | 0.63 ± 0.01 | 0.46 ± 0.03 | 0.08 ± 0.01 |
| DKT | 0.83 ± 0.01 | 0.76 ± 0.02 | 0.37 ± 0.02 | 0.07 ± 0.01 |
| IRT | 0.71 ± 0.02 | 0.65 ± 0.02 | 0.44 ± 0.02 | — |
| **BDKT** | **0.87 ± 0.01**$^{**}$ | **0.79 ± 0.02**$^{*}$ | **0.34 ± 0.02**$^{*}$ | **0.04 ± 0.01**$^{**}$ |

$^{*}p < 0.05$, $^{**}p < 0.01$ vs. DKT (paired t-test)

### A. Discrimination and Calibration

BDKT achieves statistically significant improvements over DKT in both discrimination (AUC: $0.87 \pm 0.01$ vs. $0.83 \pm 0.01$, $p < 0.01$) and calibration (ECE: $0.04 \pm 0.01$ vs. $0.07 \pm 0.01$, $p < 0.01$). The improvement in calibration is particularly noteworthy, as high discrimination does not guarantee well-calibrated confidence estimates [16].

To assess calibration quality beyond scalar ECE, we examined reliability diagrams across all folds. BDKT consistently produces confidence estimates that align more closely with empirical success rates across different confidence bins, with the largest improvements observed in intermediate confidence ranges (0.3-0.7) where uncertainty quantification is most critical for adaptive interventions.

### B. Experimental Comparison with Modern Knowledge Tracing Baselines

To provide a comprehensive evaluation, we compare BDKT against several representative modern knowledge tracing models beyond the classical baselines reported above. This comparison includes attention-based approaches and graph-enhanced architectures that have demonstrated strong performance in recent literature. All models were implemented using identical experimental protocols and hyperparameter optimization procedures. The evaluation encompasses both traditional metrics and uncertainty-related measures to ensure fair assessment across different architectural paradigms.

TABLE IV
COMPARISON WITH MODERN KT BASELINES (MEAN $\pm$ STD, 5-FOLD CV)

| Model | Architecture | AUC | ACC |
|---|---|---|---|
| DKT | LSTM | 0.83 ± 0.01 | 0.76 ± 0.02 |
| SAKT | Self-attention | 0.84 ± 0.02 | 0.77 ± 0.01 |
| DKVMN | Memory network | 0.82 ± 0.02 | 0.75 ± 0.02 |
| GKT | Graph neural | 0.85 ± 0.01 | 0.78 ± 0.02 |
| **BDKT (Ours)** | **Bayesian-neural** | **0.87 ± 0.01**$^{*}$ | **0.79 ± 0.02**$^{*}$ |

$^{*}$Best performance across all modern baselines

Modern attention-based and graph-enhanced models achieve notable improvements over classical approaches, with SAKT and GKT showing particular strength in capturing complex interaction patterns and skill dependencies respectively. SAKT's self-attention mechanism effectively weights historical interactions, while GKT leverages prerequisite relationships to improve prediction accuracy. However, these architectures focus primarily on improving discrimination without addressing prediction confidence. On the evaluated dataset, SAKT and GKT show the strongest discrimination among neural baselines (+0.01–0.02 AUC over DKT). BDKT achieves a further +0.02–0.03 AUC gain by combining their architectural strengths with principled uncertainty estimation. Whether this ranking holds on datasets with different skill-graph structures or sparser interactions requires further benchmarking. The calibration advantage—ECE halved relative to DKT—is most pronounced in low-data regimes, where point-estimate models are most likely to be overconfident.

### C. Ablation Study

To isolate the contribution of each architectural component, we conducted a systematic ablation study. Table V reports AUC and ECE for different model variants.

TABLE V
ABLATION STUDY RESULTS

| Model Variant | AUC | ECE |
|---|---|---|
| BDKT (full) | $0.87 \pm 0.01$ | $0.04 \pm 0.01$ |
| w/o MC dropout | $0.85 \pm 0.02$ | $0.06 \pm 0.01$ |
| w/o graph cov. | $0.86 \pm 0.01$ | $0.05 \pm 0.01$ |
| w/o multimodal | $0.84 \pm 0.02$ | $0.05 \pm 0.01$ |
| w/o attention | $0.86 \pm 0.01$ | $0.04 \pm 0.01$ |
| DKT baseline | $0.83 \pm 0.01$ | $0.07 \pm 0.01$ |

Monte Carlo dropout provides the largest contribution to both discrimination and calibration improvements. Removing stochastic inference reduces AUC by 0.02 and increases ECE by 0.02, highlighting uncertainty quantification importance.

Multimodal observations contribute substantially to discrimination ($\Delta$ AUC = 0.03), while graph-informed covariance provides moderate improvements. The attention mechanism offers minimal gains, suggesting simpler sequential dependencies suffice for this experimental setting.

### D. Diagnostic Analyses

BDKT shows consistent performance across skill frequency quartiles (AUC: 0.85–0.89) and item difficulty levels. Performance is stable across dropout rates $p \in [0.1, 0.3]$ and MC sample counts 10–50, with diminishing returns beyond 20 samples; higher $\beta$ improves calibration at a small discrimination cost.

### E. Computational Efficiency

BDKT incurs a $2.5\times$ training overhead over DKT (150 vs. 380 sequences/second on V100). At inference, reducing to 5–10 MC samples degrades AUC and ECE by less than 0.005 each, making a reduced-sample mode viable for latency-sensitive deployment. Alternatively, full 20-sample inference can be reserved for interactions where the model's uncertainty exceeds a threshold, leaving mean-prediction for the remainder.

## VII. CONCLUSIONS

BDKT addresses a specific, unresolved limitation of sequential KT models: the absence of calibrated uncertainty. By maintaining a graph-structured multivariate Gaussian posterior over skill states and propagating epistemic uncertainty via Monte Carlo dropout through the LSTM transition, BDKT produces predictions that are simultaneously more discriminative and better calibrated than all evaluated baselines. The ablation study identifies MC dropout as the dominant contributor to the ECE improvement, with multimodal observations adding substantially to discrimination and graph-informed covariance providing moderate gains. These findings hold across five cross-validation folds on a single mathematical dataset and should be interpreted within that empirical scope.

Deployment of BDKT is feasible with five MC samples at inference, recovering most calibration benefit at a fraction of full latency. Future work should validate generalization on ASSISTments [13] and EdNet [14], and investigate lighter-weight alternatives such as deep ensembles or mean-field variational inference to reduce training overhead while preserving calibration.

## REFERENCES


[1] G. Rasch, "Probabilistic models for some intelligence and attainment tests," *The Danish Institute for Educational Research*, 1960.

[2] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Modeling and User-Adapted Interaction*, vol. 4, no. 4, pp. 253–278, 1994.

[3] R. S. J. d. Baker, A. T. Corbett, and V. Aleven, "More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing," in *Proc. 9th Int. Conf. Intell. Tutoring Syst.*, 2008, pp. 406–415.

[4] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *Proc. NIPS*, 2015, pp. 505–513.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] S. Pandey and G. Karypis, "A self-attentive model for knowledge tracing," in *Proc. 12th Int. Conf. Educ. Data Mining*, 2019, pp. 384–389.

[7] J. Zhang, X. Shi, I. King, and D. Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 765–774.

[8] H. Nakagawa, Y. Iwasawa, and Y. Matsuo, "Graph-based knowledge tracing: Modeling student proficiency using graph neural network," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, 2019, pp. 156–163.

[9] M. V. Michael, A. Brown, and L. Carter, "Probabilistic cognitive modeling in learning analytics," in *Proc. Int. Conf. Learning Analytics*, 2020, pp. 120–129.

[10] M. Mingming, Y. Chen, and H. Wang, "Multimodal indicators for forecasting learning gains," *IEEE Trans. Learning Technologies*, vol. 15, no. 3, pp. 350–362, 2022.

[11] J. Zhang and Z. Shi, "Hierarchical attention toward metacognitive modeling," in *Proc. Int. Conf. Artificial Intelligence in Education (AIED)*, 2019, pp. 240–251.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2014.

[13] J. C. Stamper, T. Lin, N. T. Heffernan, and J. A. Smith, "The ASSISTments data mining competition 2015," in *Proc. EDM*, 2015, pp. 712–715.

[14] Y. J. Choi, S. Lee, J. Shin, and B. Kim, "EdNet: A large-scale hierarchical dataset in education," *IEEE Trans. Learning Technologies*, vol. 13, no. 4, pp. 788–798, 2020.

[15] M. Khajah, R. V. Lindsey, and M. C. Mozer, "How deep is knowledge tracing?" in *Proc. EDM*, 2014, pp. 123–130.

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learning (ICML)*, 2017, pp. 1321–1330.

[17] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learning (ICML)*, 2016, pp. 1050–1059.

[18] S. Nadarajah, J. C. Mba, N. M. V. Ravonimanantsoa, P. Rakotomarolahy, and H. T. J. E. Ratolojanahary, "Empirical calibration of XGBoost model hyperparameters using the Bayesian optimisation method: The case of Bitcoin volatility," *Journal of Risk and Financial Management*, vol. 18, no. 9, p. 487, 2025.

[19] J. Xiong, E. Vijayaraghavan, and C. Piech, "Going deeper with deep knowledge tracing," in *Proc. Int. Conf. Educational Data Mining (EDM)*, 2016, pp. 545–550.

[20] H. Ebbinghaus, *Memory: A Contribution to Experimental Psychology*. New York, NY, USA: Teachers College, Columbia Univ., 1913.

[21] Y. Yin, Z. Wang, and G. Karypis, "Robust knowledge tracing with missing data," in *Proc. Int. Conf. Educational Data Mining (EDM)*, 2020, pp. 444–455.

[22] MandaNR, "Bayesian deep knowledge dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/mandasoanr/bayesian-deep-knowledge

[23] A. Ratovondrahona, H. Rakotozanany, T. Mahatody, and V. Manantsoa, "Human like programming using SPADE BDI agents and the GPT-3-based Transformer," in *Human Interaction and Emerging Technologies (IHIET-AI 2023): Artificial Intelligence and Future Applications*, T. Ahram and R. Taiar, Eds. AHFE International, USA, 2023, vol. 70, AHFE Open Access.

[24] R. E. Ralinirina, J. C. Ralaivao, N. M. Ralaivao, A. J. Ratovondrahona, and T. Mahatody, "Unveiling the potential of explainable artificial intelligence in predictive modeling, exploring food security and nutrition in Madagascar," 2025.