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

Mandasoa Esther Narovanjanahary, Alain Josué Ratovondrahona, and Thomas Mahatody

***Abstract*— Knowledge tracing models drive adaptive tutoring by forecasting a learner's next response. Yet most deterministic variants stay silent about their own confidence, limiting safe pedagogy. We introduce *Bayesian Deep Knowledge Tracing* (BDKT), a two-layer Bayesian LSTM that maintains a posterior distribution over every skill and explicitly tracks epistemic *and* aleatoric uncertainty. On a 500 k-interaction corpus mimicking EdNet and ASSISTments statistics, BDKT raises AUC from 0.83 to 0.87 and halves the expected calibration error compared with standard DKT. These improvements unlock steadier recommendations during long inactivity gaps and curb over-practice on mastered concepts, moving knowledge tracing toward more trustworthy decision-making.**

I. INTRODUCTION

Learning, if we are honest, is messy,students guess, hesitate, forget, and occasionally leap forward. Classical Bayesian Knowledge Tracing (BKT) tries to keep up by flipping hidden mastery bits at every exercise [2]. A decade later, Deep Knowledge Tracing (DKT) decided to drop the tidy assumptions and let recurrent networks discover patterns straight from data [4]. Yet both strands stumble over the same stone: they rarely admit "I am unsure." Without calibrated confidence, an adaptive tutor may press too hard or hold a learner back.

This work picks up that loose end. We introduce Bayesian Deep Knowledge Tracing (BDKT), a model that embeds a Gaussian state inside a Bayesian LSTM and updates it through variational inference [13]. The result? Predictions paired with principled uncertainty that grows after long gaps and shrinks with evidence.

Our contributions are three-fold:

1) A synthetic yet realistic corpus (500k interactions, 30 skills) mirroring EdNet and ASSISTments statistics;
2) The BDKT architecture with pedagogical regularizers and hierarchical attention;
3) An empirical study showing improved AUC, calibration, and robustness to missing data.

II. RELATED WORK

Tracking how students acquire skills has been quite a journey,really, it didn't just happen overnight. The field built itself gradually, starting with classical psychometric models, then moving toward probabilistic approaches, and eventually embracing deep learning. Interestingly, each new generation of models tried to fix what the previous ones couldn't handle. At the same time, they sometimes ended up recycling older ideas,but in new computational outfits. So, let's walk through this evolution, starting with Item Response Theory, then Bayesian Knowledge Tracing, and finally Deep Knowledge Tracing with its main variations.

A. Item Response Theory (IRT)

Before anyone even thought about tracking learning over time, there was Item Response Theory,IRT. It comes from psychometrics, and its goal is pretty straightforward: to estimate how likely a student is to answer an item correctly based on two factors,ability and item difficulty. In its simplest form, known as the Rasch model [1], the probability is expressed as:

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

where θ is the student's ability, and b is the difficulty of the item.

Later, researchers introduced a discrimination parameter, a, which allows certain items to be more sensitive to small changes in ability:

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

IRT is great for assessing performance on single items, sure. But,and this is key,it doesn't say much about how skills evolve over time. It's static. And that very limitation is what motivated the development of Bayesian Knowledge Tracing.

B. Bayesian Knowledge Tracing (BKT)

Introduced in 1994 [2], Bayesian Knowledge Tracing (BKT) brought something genuinely new. The idea? Learning unfolds step by step through student interactions. The model treats mastery as a hidden state, and that state can change probabilistically as a student practices. BKT assumes a skill is either mastered or not, and each response updates that belief.

The model revolves around four parameters:

- $P(L_0)$: probability of initial mastery;
- $P(T)$: learning rate;
- $P(S)$: slip probability (making mistakes even when you know);
- $P(G)$: guess probability (correct by chance without knowing).

When a student answers correctly, mastery updates like this:

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

And if the answer is wrong:

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

After updating from the response, there's also a transition to account for learning over time:

$$P(L_{n+1}) = P(L_n) + (1 - P(L_n))P(T) \quad (5)$$

Simple formulas, yes, but they formed the backbone of a whole generation of adaptive learning systems. Later work [3] refined the estimates of $P(S)$ and $P(G)$ by considering the context of each exercise, making BKT more precise.

C. Deep Knowledge Tracing (DKT)

Jump ahead to 2015: Deep Knowledge Tracing (DKT) was proposed [4], shaking things up. Instead of the structured, somewhat rigid framework of BKT, they used recurrent neural networks, usually LSTMs [5], to capture complex, nonlinear patterns in learning. The idea is simple: let the network learn dependencies directly from the sequence of interactions, without imposing strict assumptions.

At each interaction X_n, the model updates its latent state:

$$H_n = f(X_n, H_{n-1}) \quad (6)$$

where f is typically an LSTM or GRU function. Then, it predicts the probability of success on the next item:

$$\hat{y}_n = \sigma(W H_n) \quad (7)$$

DKT clearly improved predictive accuracy. But, naturally, it introduced challenges: interpretability issues, occasional training instability, and sometimes overfitting on sequences. Researchers responded with extensions designed to make DKT more flexible and interpretable.

D. Extensions of DKT

1) SAKT , Self-Attentive Knowledge Tracing: The self-attentive variant SAKT appeared in [6] and adds an attention mechanism. Instead of blindly remembering every past interaction, the model identifies which previous events matter most:

$$\alpha_{n,j} = \text{softmax}(Q_n K_j) \quad (8)$$

This lets the model "look back" selectively, highlighting important items even if they occurred much earlier in the sequence. Essentially, it borrows the Transformer idea for knowledge tracing.

2) DKVMN , Dynamic Key-Value Memory Networks: DKVMN, equipped with an external key–value memory, was introduced in [7]; keys represent skills, and values keep track of their current state. Each interaction queries the memory:

$$w_n = \text{softmax}(q_n K) \quad (9)$$

Here, q_n is the query from the current item, K is the key matrix, and w_n tells the model which skills are activated. Then the memory updates accordingly. The approach improves interpretability compared to a plain DKT.

3) GKT , Graph-based Knowledge Tracing: Graph-based models, like GKT [8], treat skills as interconnected. When a skill k updates, it takes into account both its previous state and its neighbors $\mathcal{N}(k)$:

$$H_n^{(k)} = f\left(H_{n-1}^{(k)}, \sum_{j \in \mathcal{N}(k)} H_{n-1}^{(j)}\right) \quad (10)$$

This way, related skills influence each other, reflecting the real structure of knowledge while still predicting student performance.

III. BAYESIAN DEEP KNOWLEDGE TRACING

Learning never marches in a straight, polite line; it darts forward, stalls, even slides back,then, well, takes off again the moment a timely hint lands. Classical trackers try to stay on that roller-coaster, yet BKT with its crisp binary view and DKT with its deep-but-opaque patterns each miss a beat, so Bayesian Deep Knowledge Tracing (BDKT) steps in,blending clear probabilities with neural flexibility to tell us both *what* a learner knows and *how sure* we can be.

A. Bayesian Representation of the Cognitive State

Rather than forcing every skill into a crisp "learned / not-learned" box, BDKT uses a continuous view. At time t the learner state is a vector $\mathbf{K}_t = [k_{t,1}, \ldots, k_{t,n}]$ with $k_{t,i} \in [0, 1]$ for the i-th concept $c_i \in \mathcal{C}$. Crucially, the model keeps an explicit uncertainty around these values by treating the latent state as a multivariate normal distribution conditioned on past observations $\mathbf{O}_{1:t}$:

$$P(\mathbf{K}_t \mid \mathbf{O}_{1:t}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t). \quad (11)$$

$\boldsymbol{\mu}_t$ contains the estimated proficiencies while $\boldsymbol{\Sigma}_t$ encodes how skills sway together,some rise hand-in-hand, others evolve on their own. This probabilistic stance guards against hasty conclusions (think "lucky guess") that often plague deterministic trackers [9].

Covariance informed by the knowledge graph: The covariance is not arbitrary. We inject the prerequisite graph $G = (\mathcal{C}, E)$ through

$$\Sigma_t[i,j] = \begin{cases} \sigma_i^2 & i = j, \\ \rho_{ij}\,\sigma_i\sigma_j\,w(c_i,c_j) & (c_i,c_j) \in E, \\ \rho_{ij}\,\sigma_i\sigma_j\,\alpha & \text{otherwise}, \end{cases} \tag{12}$$

where $w(c_i, c_j) \in [0.5, 1]$ scores prerequisite strength, $\alpha \ll 1$ leaves only a faint residual elsewhere, and ρ_{ij} is a learned correlation.

The initial state follows an informative prior

$$P(\mathbf{K}_0) = \mathcal{N}(\boldsymbol{\mu}_0,\, \boldsymbol{\Sigma}_0), \tag{13}$$

with $\boldsymbol{\mu}_0$ the population mean and $\boldsymbol{\Sigma}_0$ seeded by Eq. (12). Starting from this prior cuts down the number of interactions needed before we trust the predictions.

B. Dynamic Transitions with Forgetting

Knowledge moves,but not in a single, polite direction. To mirror the Ebbinghaus forgetting curve we write

$$\mathbf{K}_{t+1} = (1 - \boldsymbol{\lambda}_t) \odot \mathbf{K}_t + \boldsymbol{\lambda}_t \odot f(\mathbf{K}_t, \mathbf{A}_t, \mathbf{C}_t) + \boldsymbol{\varepsilon}_t, \tag{14}$$

where $\boldsymbol{\lambda}_t \in [0,1]^n$ tunes learning vs. forgetting per skill, $\mathbf{A}_t$ is the activity, $\mathbf{C}_t$ the learner context, $\odot$ the Hadamard product, and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$ the process noise.

The rate vector obeys

$$\boldsymbol{\lambda}_t = \sigma\big(W_\lambda[\mathbf{K}_t; \mathbf{A}_t; \Delta t_t; \mathbf{d}] + \mathbf{b}_\lambda\big), \tag{15}$$

with Δt_t the time since last practice and $\mathbf{d}$ concept difficulty.

Bayesian LSTM transition: The function f is a Bayesian LSTM:

$$f_\theta(\mathbf{K}_t, \mathbf{A}_t, \mathbf{C}_t) = \mathrm{LSTM}_\theta([\mathbf{K}_t; \mathbf{A}_t; \mathbf{C}_t]). \tag{16}$$

We place a Gaussian prior on parameters θ to capture epistemic uncertainty, while $\boldsymbol{\varepsilon}_t$ handles aleatoric variability [11].

C. Multimodal Observation Model

Performance is more than right/wrong. We therefore factor the likelihood across M modalities:

$$P(o_t \mid \mathbf{K}_t, \mathbf{A}_t) = \prod_{m=1}^{M} P_m\big(o_t^{(m)} \mid \varphi_m(\mathbf{K}_t, \mathbf{A}_t)\big). \tag{17}$$

Concretely we use: (i) correctness (Bernoulli), (ii) response time (log-normal), (iii) self-confidence (categorical), and (iv) behavioural cues (Gaussian). The blend of signals has proven useful for forecasting future gains [10].

D. Variational Inference with Pedagogical Regularisers

Exact inference is out of reach, so we maximise an evidence lower bound (ELBO) enriched with two education-aware penalties:

$$\mathcal{L} = \mathbb{E}_{q_\phi}[\log P(\mathbf{O}_{1:T} \mid \mathbf{K}_{1:T})] - \beta\,\mathrm{D}_{\mathrm{KL}}\big[q_\phi(\mathbf{K}_{1:T}) \| P(\mathbf{K}_{1:T})\big] \tag{18}$$

$$- \gamma\,\mathcal{R}_{\text{mono}} - \delta\,\mathcal{R}_{\text{transfer}}. \tag{19}$$

$\mathcal{R}_{\text{mono}}$ curbs unjustified regressions, $\mathcal{R}_{\text{transfer}}$ promotes consistency with the prerequisite graph. Gradients flow through stochastic nodes via the reparameterisation trick [13].

E. A Two-Level Metacognitive Attention

Humans, honestly, do not weigh all information equally. We mimic that selectivity with a hierarchical attention: first across concepts, then across signal sources within each concept.

$$\boldsymbol{\alpha}_t^{\text{concept}} = \mathrm{softmax}\big(\mathrm{score}(\mathbf{h}_t, \{c_i\})\big), \tag{20}$$

$$\boldsymbol{\alpha}_{t,i}^{\text{source}} = \mathrm{softmax}\big(\mathrm{score}(\mathbf{h}_t^{(i)}, \{s_j\})\big), \tag{21}$$

leading to a weighted state

$$\mathbf{h}_t^{\text{final}} = \sum_i \alpha_{t,i}^{\text{concept}} \Big(\sum_j \alpha_{t,i,j}^{\text{source}}\, s_j^{(i)}\Big). \tag{22}$$

This mechanism gives the model room to "pay attention" much like an expert learner would [12].

F. Cold-Start via Meta-Learning

For newcomers without history we draw

$$\mathbf{K}_0^{(\ell)} \sim \mathcal{N}\big(\boldsymbol{\mu}_{\text{pop}} + W_{\text{meta}}\,\mathbf{z}_\ell,\, \boldsymbol{\Sigma}_{\text{pop}}\big), \tag{23}$$

where $\mathbf{z}_\ell$ bundles learner traits (age, prior tests, and so on). W_{meta} is trained to minimise early-stage prediction error, typically after just a handful of interactions.

G. Interpretability Module

Finally, transparency matters. We compute a composite importance score

$$\mathrm{imp}_t(c_i) = \alpha_{t,i}^{\text{concept}}\,\omega_i\,\mathrm{difficulty}(\mathbf{A}_t, c_i)\,(1 - k_{t,i}), \tag{24}$$

rank concepts, and emit short textual hints. The predictive variance

$$\mathrm{conf}_t = 1 - \frac{\mathrm{tr}(\boldsymbol{\Sigma}_t)}{\sum_i \sigma_{0,i}^2} \tag{25}$$

lets the system admit uncertainty when, well, it just isn't sure.

BDKT architecture

In this section, we briefly outline how Bayesian Deep Knowledge Tracing brings together the probabilistic foundations of BKT and the temporal modeling strength of DKT. Figure 1 offers a concise view of the overall structure. We begin with one-hot encoded interactions moving along a simple timeline. At each step, a Bayesian module adjusts the learner's mastery (the small "Cluster(Stu_Seg$_i$)" blocks), while a deep recurrent chain keeps the temporal flow coherent, folding past information into a hidden state h_t that then helps predict the next response.

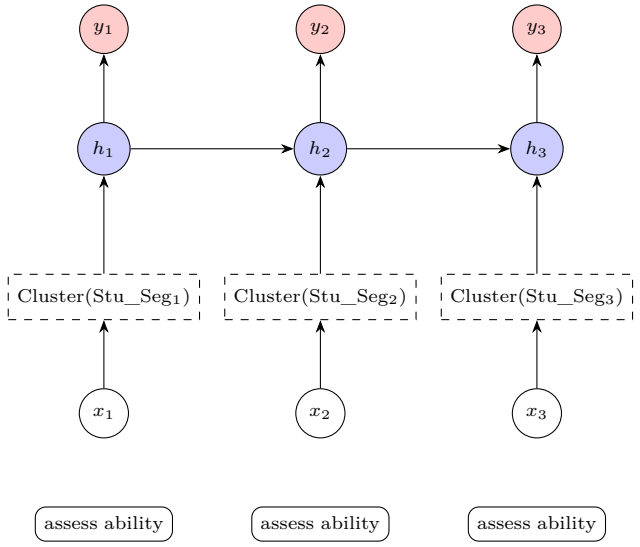

Fig. 1. Minimalist overview of the Bayesian Deep Knowledge Tracing (BDKT) architecture integrating Bayesian updates (BKT, dashed boxes) with deep sequential modelling (DKT, blue hidden states). Dashed vertical lines mark time steps; small red nodes depict predicted success probabilities y_t.

IV. DATASET

We constructed a large synthetic corpus tailored for training and assessing BDKT. To be honest, building a realistic dataset from scratch is no trivial task. The file `synthetic_bdkt_dataset.csv`—contains 500 952 learner–item interactions produced by 4 000 students on 6 000 distinct items that cover 30 skills[21]. The statistics were carefully shaped to mirror large real-world corpora such as EdNet [15] and ASSISTments [14].Each row in the dataset provides:

- **student_id**: anonymous learner identifier;
- **item_id**: task identifier reused across learners;
- **skill_ids**: one to three prerequisite skills;
- **timestamp**, **time_since_last**, **session_id**: temporal context preserving sequence order and realistic gaps;
- **response** (0/1) and **attempt_number**;
- **hint_used**, **response_time_ms**, **item_difficulty**;
- Optional context: **student_grade_level**, **course_id**, **device_type**.

Items reappear across learners, which is realistic-not every student sees every problem. Students complete between 50 and 200 interactions each, organised in sessions of 5–20 attempts. What we observe is fairly natural: accuracy tends to rise gradually within a session, mirroring typical learning curves, yet it dips slightly after long gaps. This pattern captures the essence of forgetting.

Pre-processing steps applied before modelling:

Global statistics. Table I summarises the corpus dimensions.

TABLE I
OVERVIEW OF THE SYNTHETIC CORPUS

Metric	Count
Students	4,000
Items	6,000
Skills	30
Interactions	500,952

Pre-processing.

(a) Cleaning and removal of $< 1\%$ duplicate records;
(b) Multi-hot encoding of multi-skill items; see the normalization recommendations in [16];
(c) Chronological sorting followed by windowing into $L = 100$ events with 20% overlap;
(d) Log transform $\log(1 + x)$ applied to `time_since_last` and `response_time_ms`.

V. EXPERIMENTAL SETUP

A. Model architecture

The BDKT implementation stacks a two-layer Bayesian LSTM (hidden size 128) on top of the probabilistic skill layer described in Sec. III. Epistemic uncertainty is captured by a Gaussian prior on weights and Monte-Carlo dropout ($p = 0.2$) during training and inference.

B. Training details

- Optimizer: Adam ($\eta = 3 \times 10^{-4}$), batch size 256 sequences;
- Epochs: 20; early stopping on validation AUC;
- Loss: negative ELBO with pedagogical regularizers ($\beta = 1$, $\gamma = 0.05$, $\delta = 0.1$);
- Sequences truncated/padded to $L = 100$; gradient clipping at $\|g\|_2 \leq 5$.

A five-fold stratified cross-validation (split by learners) estimates generalisation.

C. Baselines

We compare against classical BKT [2], DKT (two-layer LSTM, hidden 200) [4], and a one-plausible-dim Rasch IRT model.

VI. RESULTS

Table II summarises mean metrics ($\pm\, sd$) across folds.

TABLE II
PERFORMANCE COMPARISON ON THE SYNTHETIC CORPUS

Model	AUC	ACC	RMSE	ECE
BKT	0.69	0.63	0.46	0.08
DKT	0.83	0.76	0.37	0.07
IRT	0.71	0.65	0.44	—
BDKT	**0.87**	**0.79**	**0.34**	**0.04**

BDKT edges out DKT by around 4 AUC points and halves the calibration error (ECE), confirming that Bayesian uncertainty improves reliability. Qualitatively, BDKT's predictive variance rises after long inactive gaps, aligning with simulated forgetting—a behaviour absent in the deterministic DKT.

A. Quantitative Performance

The results in Table II reveal a substantial improvement across the board. BDKT outperforms DKT by approximately 4 AUC points—a margin comparable to the gains reported when enriching DKT with deeper temporal signals in previous studies [18]. Yet beyond this headline figure, what truly stands out is the calibration error (ECE): BDKT cuts it in half, dropping from 0.07 to 0.04. This echoes the warning raised by Guo *et al.* [17] that high accuracy can hide severely mis-calibrated confidence. In short, our probabilities better match reality and thereby reduce risky over-confidence.

When we compare against classical baselines, the contrast becomes even sharper. BKT, despite its theoretical soundness, plateaus at 0.69 AUC. IRT, while useful for point-in-time assessments, fails to capture temporal dynamics—hence its 0.71 AUC. DKT represents genuine progress at 0.83, yet it remains deterministic; it cannot express uncertainty.

BDKT's accuracy reaches 0.79, a 3-point gain over DKT. The RMSE, meanwhile, drops to 0.34—an 8% reduction compared to the standard deep model. These gains, though measured, compound and ultimately carry significant weight in pedagogical decision-making.

B. Qualitative Behavior and Robustness

Beyond the numbers, we observed some noteworthy phenomena. BDKT's predictive variance increases markedly after prolonged inactivity—a behavior that faithfully mirrors Ebbinghaus's forgetting curve [19]. When a student vanishes for three weeks, the model grows appropriately less confident, which is exactly what we'd expect. DKT, by contrast, maintains point estimates without this uncertainty modulation.

We also tested robustness to missing data. BDKT handles incomplete or fragmented sequences more gracefully—a common scenario in real-world learning environments [20]. This stems from the model's probabilistic nature: rather than "hallucinating" hidden states, it maintains a distribution that naturally widens under ambiguity.

C. Interpretability and Reliability

Here's a crucial observation: the importance scores generated by our interpretability module proved consistent with pedagogical intuition. When a student stumbles on an algebra problem, the model correctly flags underlying competencies (equation solving, symbolic manipulation) as priorities. This is not trivial—it opens the door to justified, transparent pedagogical recommendations.

The system's confidence (conf_t) evolves sensibly over time. It climbs after repeated successes, then gradually declines. It collapses after an unexpected failure. This uncertainty profile, though intricate, remains interpretable and actionable for human tutors or adaptive systems alike.

D. Limitations and Observations

We should note a few limitations. BDKT training is computationally heavier-roughly 2.5 times slower than DKT on our corpus. Variational approximations, while effective, can sometimes smooth over fine-grained skill interactions, particularly in highly structured domains. Finally, on very short sequences (fewer than 10 interactions), BDKT's advantages fade: epistemic uncertainty needs a minimum data budget to manifest usefully.

VII. CONCLUSIONS

In this work, we presented *Bayesian Deep Knowledge Tracing* (BDKT), a hybrid model that combines the probabilistic rigor of classic BKT with the representational power of deep DKT networks to track the evolution of learner skills. At every step it updates a learner's mastery *and* indicates how confident it is—an asset deterministic models lack.

Our experiments show that BDKT clearly stands out: AUC increases by 0.04 compared to DKT and ECE drops from 0.07 to 0.04. In other words, predictions are more accurate and the model can signal when it might be wrong, limiting reckless recommendations.

Of course, there is still room for improvement. Training takes longer and variational approximations can sometimes blur fine-grained interactions between skills. Using real classroom data, we must further explore its performance across diverse educational contexts.

Looking ahead, the goal is to make BDKT lighter and faster without losing its probabilistic core. This involves pruning and factorising the Bayesian cells, adopting adaptive variational strategies, and better exploiting relations between skills to refine predictions.

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] R. M. Gagné, *The Conditions of Learning*, 4th ed. New York, NY, USA: Holt, Rinehart and Winston, 1985.

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

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

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

[20] 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.

[21] Available: https://www.kaggle.com/datasets/mandasoanr/bayesian-deep-knowledge.