

Compression-First Knowledge Tracing with a Small Language Model

Max Norris^{1,*}, Daniel Huencho² and Sahan Bulathwela^{2,3,*}

¹Department of Informatics, University of Edinburgh, The United Kingdom

²Department of Computer Science, University College London, The United Kingdom

³Centre for Artificial Intelligence, University College London, The United Kingdom

Abstract

Intelligent tutoring systems depend on accurate models of student mastery over individual skills or knowledge components. Knowledge tracing (KT) is the standard evaluation task: given a student's prior interactions, predict the correctness of their response to a new question. Most KT models represent each interaction as opaque identifiers, discarding the pedagogical signal in the question text; recent text-aware methods restore this signal, but at a steep cost, since raw tokens fill the context window so quickly that encoders fit only a handful of interactions, and decoders compensate by scaling to a multi-billion-parameter backbones. We introduce *ModernBERT-KT*, which compresses each interaction to two embedding tokens via a frozen sentence encoder and small MLP adapters, allowing a ~150M-parameter ModernBERT-base backbone to process up to ~4,000 interactions per student in its native 8,192-token window, around 50× smaller than recent LLM-based KT methods. ModernBERT-KT matches established ID-based and text-aware baselines on XES3G5M and FOUNDATIONALASSIST, and generalises to questions unseen at training time through content embeddings.

Keywords

Knowledge Tracing, Small Language Models, Educational Data Mining

1. Introduction

Intelligent Tutoring Systems (ITS) aim to deliver personalised learning at scale, which depends on accurate models of each student's mastery over individual skills, or knowledge components (KCs) [1, 2]. These mastery estimates drive the diagnosis of weak areas and the planning of personalised interventions and curricula. Knowledge Tracing (KT) has emerged as the de facto task for evaluating such models [1, 3]: given a student's prior interactions, predict the correctness of their response to a new question. Accurate prediction requires both extracting pedagogically relevant features from the items students engage with and modelling the temporal patterns of knowledge acquisition and forgetting.

Despite this, most KT models represent each interaction as a tuple of opaque identifiers (a question ID, a KC ID, and a binary correctness label), including DKT [3], DKVMN [4], SAKT [5], and AKT [6]. This ID-based design has two core limitations. First, question IDs carry no semantic content, discarding the pedagogical signal contained in question text such as linguistic complexity [7], topic, and conceptual structure. Second, the model cannot generalise to questions unseen during training, known as the cold-start problem [8].

Recent text-aware KT models address these gaps. LKT [8] fine-tunes encoder models such as BERT, RoBERTa, and DeBERTa on the full text of prior questions, and Next-Token Knowledge Tracing (NTKT; 9) reformulates KT as next-token prediction over a serialised student history. Both improve over ID-based baselines and generalise to unseen questions. Two constraints, however, block deployment at scale. First, both rely on *context dumping*: placing every prior question's tokens into the model's context window. With student histories routinely spanning hundreds to thousands of interactions, this rapidly exhausts the 512-token windows of smaller encoders and strains even the longer windows

SLM4ED'26: The 1st Workshop of Small Language Models for Education (SLM4ED), June 28, 2026, Seoul, Republic of Korea

*Corresponding author.

✉ s2583684@ed.ac.uk (M. Norris); daniel.mardini.25@ucl.ac.uk (D. Huencho); m.bulathwela@ucl.ac.uk (S. Bulathwela)

id 0009-0004-0949-573X (M. Norris); 0000-0002-5878-2143 (S. Bulathwela)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

of decoder-only models. Second, leading decoder-only KT systems use billion-parameter backbones whose inference cost is substantially higher than purpose-built KT models for marginal accuracy gains [10], rendering them infeasible in many production settings. We hypothesise that compressing each interaction to a small, fixed number of adapter-aligned tokens decouples context length from text length, allowing a small language model to attend over thousands of interactions where raw-text methods accommodate fewer than twenty, and that this compression-first design enables a $\sim 150\text{M}$ -parameter encoder to match text-aware systems that rely on substantially larger backbones.

This paper makes three contributions. First, we introduce a *compression-first* KT architecture in which each interaction is reduced to two tokens produced by a frozen MPNet sentence encoder [11, 12] and aligned to the backbone via small learnable MLP adapters, in the spirit of the LLaVA visual-instruction recipe [13, 14]. A single $\sim 150\text{M}$ -parameter ModernBERT-base backbone [15] can therefore attend up to $\sim 4,000$ interactions per student in one forward pass, enough to cover a full course of activity. Second, we show that this design matches established ID-based and text-aware KT baselines on the standard XES3G5M and FOUNDATIONALASSIST benchmarks while preserving the deployment-cost profile that makes purpose-built KT models attractive at scale [10]. As a third finding, the use of content embeddings rather than question identifiers enables generalization to questions absent from the training set, addressing the cold-start limitation of ID-based KT.

2. Methodology

In this section, we formulate the KT task as a masked token prediction problem within a sequence modeling framework. We then detail our proposed deep learning architecture.

2.1. Research questions

The two research questions that drive our approach are the following:

- **RQ1:** How well does compressed content injection perform on standard KT benchmarks?
- **RQ2:** Does the long-history capacity enabled by compression carry useful signal?
- **RQ3:** Does the model generalise to questions never seen at training time through question *content* rather than question identity?

2.2. Problem Setup

For a given student, we observe a sequence of N interactions. Each interaction at time step $t \in \{1, \dots, N\}$ is represented as a tuple (q_t, c_t, r_t) , where q_t denotes the natural-language text of a question or dialogue turn, c_t is the text of the associated knowledge concept, and $r_t \in \{0, 1\}$ is the observed binary response. The objective at each interaction t is to accurately predict the student’s performance, r_t , utilizing the information available at that step. While standard Knowledge Tracing (KT) models predict r_t autoregressively from the historical prefix [3, 1], we instead adopt a bidirectional masked-response formulation.

2.3. ModernBERT-KT Architecture

This architecture can be viewed as a three-level system, consisting of an initial layer of frozen, pre-processed embeddings, followed by a layer of adapters that generate a projection of these embeddings, and finally, a Small Language Model (SLM) serving as the primary learner for the knowledge tracing task (Figure 1).

The first component is a frozen sentence encoder that captures the semantic content of each question and knowledge concept. We use all-mpnet-base-v2, a sentence transformer trained on over one billion sentence pairs for semantic similarity [11]. With this model we map each question text q_t and knowledge-concept text c_t to a 768-dimensional vector. These embeddings are pre-computed once per

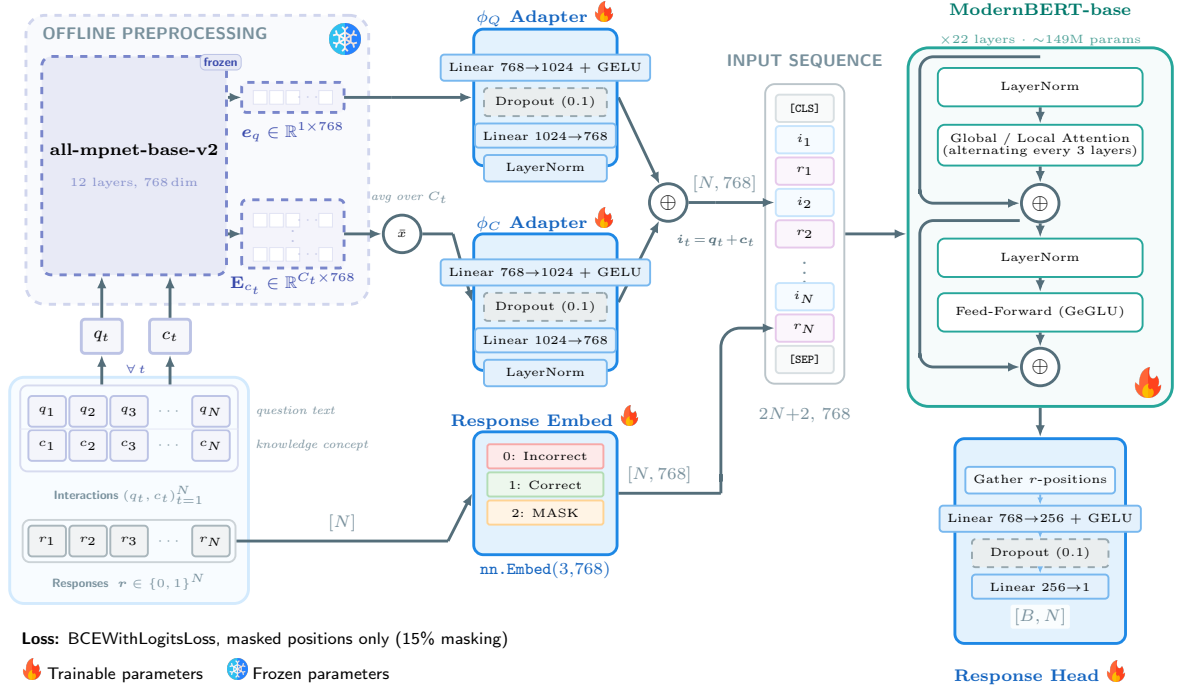


Figure 1: ModernBERT-KT architecture. Each interaction is compressed to two tokens: an interaction token (adapted question + KC embeddings) and a response token. A ModernBERT-base backbone (~ 150 M parameters) processes up to 512 such interaction pairs in a single 1,026-token sequence.

dataset and stored as a lookup table, adding no trainable parameters and zero runtime cost during training or inference.

When a single interaction is tagged with several knowledge concepts we average the raw MPNet embeddings of all active KCs before applying ϕ_C at injection time, preserving the two-token-per-interaction layout, following the compensatory aggregation scheme of Scarlatos et al. [16]

Two lightweight MLP adapters are used to learn a dimension-preserving $\mathbb{R}^{768} \rightarrow \mathbb{R}^{768}$ projection mapping into a learning-focused geometric space, isolating pedagogical features so the ModernBERT backbone can dedicate its full capacity to estimating student performance. This design follows the spirit visual-instruction-tuning recipe introduced by LLaVA [13], where a small learned projection aligns a frozen external encoder to a pretrained transformer. Each adapter is a two-layer MLP with hidden size 1024 and dropout 0.1.

The projected question and knowledge-concept vectors are summed into a single interaction token $\mathbf{i}_t = \mathbf{q}_t + \mathbf{c}_t$. Paired with a response token drawn from a learnable embedding table over {correct, incorrect, [MASK]}, each interaction occupies exactly two tokens in the interleaved sequence:

$$[\text{CLS}], \mathbf{i}_1, r_1, \mathbf{i}_2, r_2, \dots, \mathbf{i}_N, r_N, [\text{SEP}],$$

yielding $2N + 2$ tokens for N interactions, or 1,026 tokens for a full history of 512 interactions. In contrast, raw-text methods such as LKT [8] and NTKT [9] encode each interaction as raw subword tokens. Scaling these methods to 512 interactions at typical question lengths would demand tens of thousands of tokens, far exceeding practical context limits for models of comparable size. This compression strategy enables the backbone to attend to a student’s complete interaction history within a single forward pass.

The compressed sequence fits within the native 8,192-token context of ModernBERT-base, a ~ 150 M-parameter bidirectional encoder optimised for long-sequence efficiency [15]. Its transformer stack alternates local sliding-window attention (128 tokens) with global full-sequence attention every third

layer; local layers model fine-grained temporal dependencies among neighbouring interactions, while periodic global layers propagate information across the full student history without incurring quadratic cost over the entire sequence. At this scale the complete system is approximately 50 times smaller than the $\sim 8\text{B}$ -parameter Llama-3.1 backbone used by LLMKT [16].

During training, 15% of response tokens are replaced with [MASK] uniformly at random, and a per-position prediction head is trained with binary cross-entropy to recover the true responses. Because the encoder is bidirectional, each masked position draws on both past and future context for this prediction, mirroring BERT-style masked language modelling. At evaluation time, the model instead predicts each response r_t from the causal prefix $\{i_{1..t}, r_{1..t-1}\}$, simulating real-time deployment. All AUC and accuracy numbers reported in this paper use this causal protocol, so the results reflect real-time prediction constraints rather than the richer training-time context.

3. Experimental Setup

In this section, we describe the baseline models, datasets, and the series of experiments employed to test the proposed ModernBERT-KT model against the outlined RQs.

Baselines We compare against three baselines. DKT [3] encodes each interaction as a single token combining the knowledge concept ID and the binary response, processes the sequence with an LSTM, and predicts the next response via a per-concept sigmoid head. AKT [6] replaces the recurrence with causal self-attention over Rasch-style embeddings that separate question difficulty from concept identity, and applies a monotonic exponential distance decay to attention weights to encode recency. Both DKT and AKT operate on numerical IDs, discard all textual content, and are trained with a causal next-step objective where every position is supervised. LKT [8] takes the opposite approach: it concatenates the raw question and concept text into the input sequence of a pre-trained encoder, inserts special response tokens ([CORRECT], [INCORRECT], [MASK]) between interactions, and fine-tunes the full model with a masked response prediction objective identical in structure to ours. We reimplement LKT with DeBERTa-v3-base, their best-performing backbone on XES3G5M, and train it under the same evaluation protocol and hyperparameter budget as our model to ensure a controlled comparison. The DeBERTa-v3-base backbone capped at 512 input tokens (fits 5–10 interactions), whereas ModernBERT-KT fits 512 interactions via a 2-token-per-interaction injection layout (1026-token sequence).

Dataset We evaluate our approach on two benchmark datasets exhibiting diverse characteristics (Table 1). XES3G5M [17] is a large-scale Chinese primary mathematics benchmark; we utilize the English-translated split provided by Lee et al. [8]. The second dataset is FOUNDATIONALASSIST [18] provides full question text and student responses aligned to Common Core K–12 standards.

Table 1

Summary statistics on the two datasets used in the experiments.

Dataset	Items	KCs	Interactions	Avg. seq. length
FOUNDATIONALASSIST	3,359	209	$\sim 1.66\text{M}$	332
XES3G5M	$\sim 7,650$	~ 800	$\sim 5.5\text{M}$	306

Evaluation Metrics Our experiments primarily measure predictive performance of the models in the knowledge tracing task. The task is a binary classification problem. We employ AUC-ROC, accuracy, and the F1-score (the harmonic mean of precision and recall) to compare the models.

3.1. Experiments

Experiment 1: Overall predictive performance (RQ1) To assess the proposed model’s ability to perform well against current state-of-the-art baselines, we compare the model’s performance with the competitive baselines identified in this work.

Experiment 2: Effect of Sequence Length on Performance (RQ2) To assess the model’s ability to leverage long-range dependencies, we perform a context-window ablation on 500 test students. Sampling eight late-sequence targets per student (indices >100) yields 4,000 predictions per history length $k \in \{2, 4, 8, 16, 32, 64, 128, \text{full}\}$. By varying k , we isolate the effect of historical context on predictive accuracy. Uncertainty is quantified via student-level bootstrapping ($B=1,000$) to control for intra-student correlation, providing 95% confidence intervals for the AUC at each k .

Experiment 3: Cold Start (RQ3) Generalising to questions absent from training is a core challenge for knowledge tracing. Id-based models learn each question’s difficulty parameter from many student interactions and have no signal for novel material beyond the student’s running latent state. Text-aware models can in principle close this gap by grounding predictions in a question’s semantic content rather than its identity. To test this, we construct a held-out-questions variant of each dataset by uniformly sampling 5% of the question vocabulary (383 questions on XES3G5M, 168 on FoundationalASSIST) and removing every interaction on those questions from the training and validation splits; students whose remaining sequence falls below five interactions are dropped. The test split is left intact, so each student’s stream contains both seen and held-out questions interleaved over an identical prior history. We retrain every model on this cold-start dataset and stratify test AUC by whether the predicted question is in the held-out set, reporting the held-out bucket as the headline. ModernBERT-KT with MPNet embeddings is compared against DKT and AKT, as well as against an ablation that replaces ModernBERT-KT’s question and KC embedding tensors with distribution-matched Gaussian noise, isolating the contribution of question content from any stable per-id signature.

Implementation Details For the implementation, we use a backbone learning rate of 3×10^{-5} on XES3G5M and FOUNDATIONALASSIST; the prediction head learning rate is 1×10^{-4} . Other settings are warm-up of 200–500 steps, weight decay of 0.01, a cosine learning-rate schedule, gradient clipping at 1.0, and per-device batch sizes of 64 or 128. We use early stopping on validation AUC.

4. Results and Discussion

In this section, we present the overall results observed in the two experiments addressing the two RQs.

4.1. Performance of ModernBERT-KT (RQ1)

Table 2 summarises the performance metrics observed in the main experiment evaluating the predictive performance of the proposed ModernBERT-KT method.

Results on XES3G5M. ModernBERT-KT reaches 0.816 AUC at the causal aggregate level. This is below the headline LKT numbers, which are reported under a bidirectional protocol that is not deployable as a next-step predictor. The aggregate number, however, understates the architectural advantage: short-context text-aware methods cannot reach the positions where ModernBERT-KT is strongest.

Results on FoundationalASSIST. On the FoundationalASSIST dataset, ModernBERT-KT yields the strongest results across all evaluated metrics. Our model reaches an AUC of 0.8095, an Accuracy of 0.7520, and an F1 score of 0.8127. It outperforms the DKT and LKT baselines by a margin of over 0.02 in AUC, while demonstrating a narrower, yet consistent, improvement over AKT.

Table 2

Test results across two KT benchmarks; bold marks the best result per metric. All numbers are our own runs on identical splits. XES3G5M inter-seed AUC spread < 0.001 . The sample sizes are 3,613 students for XES3G5M and 250 students for FoundationalASSIST. (†) symbolises using question IDs only, while (‡) symbolises models that use the question text.

Method	XES3G5M			FoundationalASSIST		
	AUC	Acc	F1	AUC	Acc	F1
DKT †	0.7898	0.8191	0.8938	0.7796	0.7328	0.8023
AKT †	0.8131	0.8256	0.8965	0.8065	0.7495	0.8106
LKT-DeBERTa-v3 ‡	0.7899	0.8196	0.8940	0.7603	0.7157	0.7903
ModernBERT-KT ‡ (Ours)	0.8159	0.8271	0.8988	0.8095	0.7520	0.8127

4.2. Effect of Sequence Length on Performance (RQ2)

Figure 3 characterises the context horizon window effect over the model’s performance on XES3G5M dataset. For this objective, we measure the same target positions under systematic history truncation to the last $k \in 2, 4, 8, \dots, 128$ interactions. AUC rises steeply from 0.534 at $k=2$ to 0.698 at $k=16$, which already exceeds 95% of the full-context AUC (0.723). By $k=32$ the model recovers 99.6% of the full-context performance (AUC = 0.720), and further context yields marginal gains. The analyses indicate an asymptotic learning behaviour where the bulk of the predictive utility is resolved within the first 16–32 interactions, yet the model continues to successfully leverage extended horizons.

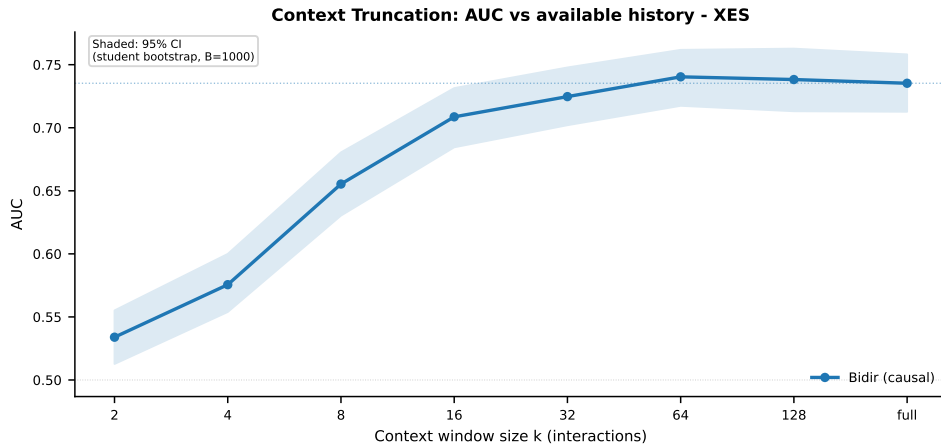


Figure 2: Test predictive performance (AUC) as a function of context window size k (number of past interactions) for ModernBERT-KT on XES3G5M. Shaded: 95% bootstrap CI.

4.3. Performance During Cold-Start (RQ3)

Figure 4.3 illustrates results from our cold start experiments. Text-aware ModernBERT-KT achieves the highest held-out-question AUC on both datasets. On FoundationalASSIST, it beats DKT, AKT, and the scrambled-embedding ablation by +0.043, +0.079, and +0.063 AUC respectively, with the smallest seen-to-cold drop of any model (0.048). On XES3G5M the same ordering holds with smaller margins: +0.018, +0.031, and +0.024 AUC. The four models are tied within ± 0.006 AUC on seen questions and separate only on held-out ones, indicating that text embeddings contribute specifically in the cold-start regime.

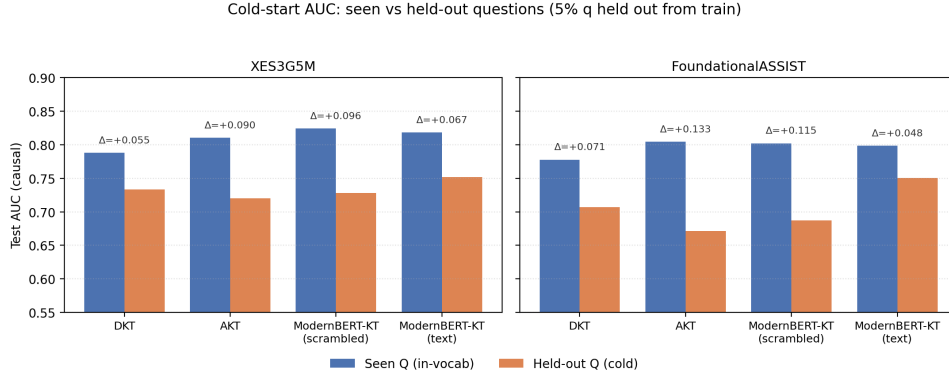


Figure 3: Test AUC comparison of seen (in-vocab) versus held-out (cold) questions on XES3G5M and FoundationalASSIST datasets. Δ indicates the performance drop in the cold-start regime, where text-aware ModernBERT-KT achieves the highest held-out AUC and smallest degradation.

5. Conclusion

We introduced ModernBERT-KT, a compression-first knowledge tracing architecture that reduces each interaction to two adapter-aligned tokens, allowing a 150M-parameter backbone to attend over thousands of interactions in a single forward pass. This is roughly $50\times$ smaller than recent LLM-based KT systems and $50\times$ longer in effective history than raw-text encoder methods. Three experiments support this design. First, the KT benchmark shows that ModernBERT-KT matches ID-based and text-aware baselines on XES3G5M and FoundationalASSIST under a next-response prediction task, demonstrating that pedagogical signal survives aggressive compression. Experiment 2 establishes the robustness of semantic representations in the cold-start setting, where the model achieves the highest held-out-question AUC on both datasets with the smallest seen-to-cold-start gap. This advantage collapses when embeddings are replaced with noise, isolating semantic content as the source of generalisation. Experiment 3 highlights the value of long-context modelling, with position-wise AUC continuing to improve well past the 16–32 interactions accessible to short-context methods.

As intelligent tutoring systems accumulate ever-longer interaction histories per student, this capacity will only grow in importance. While the proposed approaches allow achieving state-of-the-art predictive performance using sLMs (nearing the performance of billion-parameter-scale LLMs (e.g., [9]), the results also open opportunities to extend LLM-centered KT models for lifelong learning scenarios (e.g. [19, 20]). The architectural framework can also inform similar personalisation algorithms (that may incorporate knowledge [21] and interest [22] states of users) In future work we plan to investigate the applicability of this approach to text-heavy tutoring environments such as dialogue tracing, and to extend it to multi-modal settings with image-based artefacts such as geometry problems.

Acknowledgments

This work is co-funded by the University of Edinburgh UKRI AI Centre for Doctoral Training and the EC projects “Teacher-AI Complementarity (TaiCo)” (Project ID: 101177268), “Humane AI” (Grant No. 820437) and “X5GON” (Grant No. 761758) and the UCL Computer Science Strategic Research Fund.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: assist with code implementation, grammar and spelling checks, and Figure 1 development. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] G. Abdelrahman, Q. Wang, B. Nunes, Knowledge tracing: A survey, *ACM Computing Surveys* 55 (2023) 1–37.
- [2] Bulathwela, Sahan and Pérez-Ortiz, María and Yilmaz, Emine and Shawe-Taylor, John, Power to the Learner: Towards Human-Intuitive and Integrative Recommendations with Open Educational Resources, *Sustainability* 14 (2022).
- [3] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, J. Sohl-Dickstein, Deep knowledge tracing, in: *Advances in Neural Information Processing Systems*, 2015.
- [4] J. Zhang, X. Shi, I. King, D.-Y. Yeung, Dynamic key-value memory networks for knowledge tracing, in: *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 765–774.
- [5] S. Pandey, G. Karypis, A self-attentive model for knowledge tracing, in: *Proceedings of the 12th International Conference on Educational Data Mining*, 2019.
- [6] A. Ghosh, N. Heffernan, A. S. Lan, Context-aware attentive knowledge tracing, in: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.
- [7] O. C. Barbu, C. R. Beal, Effects of linguistic complexity and math difficulty on word problem solving by English learners, *International Journal of Education* 2 (2010) 1.
- [8] U. Lee, J. Bae, D. Kim, S. Lee, J. Park, T. Ahn, G. Lee, D. Stratton, H. Kim, Language model can do knowledge tracing: Simple but effective method to integrate language model and knowledge tracing task, in: *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.
- [9] M. Norris, K. Gal, S. Bulathwela, Next token knowledge tracing: Exploiting pretrained LLM representations to decode student behaviour, in: *arXiv preprint arXiv:2511.02599*, 2025.
- [10] P. Bhattacharyya, J. Mitton, R. Abboud, S. Woodhead, Faster, cheaper, more accurate: Specialised knowledge tracing models outperform LLMs, *arXiv preprint arXiv:2603.02830* (2026).
- [11] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MPNet: Masked and permuted pre-training for language understanding, in: *Advances in Neural Information Processing Systems*, 2020.
- [12] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.
- [13] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, in: *Advances in Neural Information Processing Systems*, 2023.
- [14] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [15] B. Warner, A. Chaffin, B. Clavié, et al., ModernBERT: Smarter, better, faster, longer, *arXiv preprint arXiv:2412.13663* (2024).
- [16] A. Scarlatos, R. S. Baker, A. Lan, Exploring knowledge tracing in tutor-student dialogues using LLMs, in: *Proc. of the 15th Int. Learning Analytics and Knowledge Conference*, 2025.
- [17] Z. Liu, Q. Liu, J. Chen, S. Huang, B. Gao, W. Luo, J. Tang, XES3G5M: A knowledge tracing benchmark dataset with auxiliary information, *arXiv preprint arXiv:2306.03101* (2024).
- [18] E. Worden, C. Heffernan, N. Heffernan, S. Sonkar, FoundationalASSIST: An educational dataset for foundational knowledge tracing and pedagogical grounding of LLMs, *arXiv preprint arXiv:2602.00070* (2026). URL: <https://arxiv.org/abs/2602.00070>.
- [19] S. Bulathwela, D. Van Niekerk, J. Shipton, M. Perez-Ortiz, B. Rosman, J. Shawe-Taylor, Truereason: An exemplar personalised learning system integrating reasoning with foundational models, *arXiv preprint arXiv:2502.10411* (2025).
- [20] M. Perez-Ortiz, E. Novak, S. Bulathwela, J. Shawe-Taylor, An ai-based learning companion promoting lifelong learning opportunities for all, *arXiv preprint arXiv:2112.01242* (2021).
- [21] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, J. Shawe-Taylor, Truelearn: A family of bayesian algorithms to match lifelong learners to open educational resources, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020, pp. 565–573.
- [22] S. Bulathwela, M. Pérez-Ortiz, E. Yilmaz, J. Shawe-Taylor, Power to the learner: Towards human-intuitive and integrative recommendations with open educational resources, *Sustainability* 14 (2022) 11682.