Artificial Intelligence in Education: Contrastive Learning to LLM-like Autoregressive Modeling of Knowledge Tracing

Ph.D. Qualification Examination

Submitted by: Indronil Bhattacharjee

Committee Members:

Dr. Christabel Wayllace
Dr. Huiping Cao
Dr. Tuan Le

BE BOLD. Shape the Future.®
New Mexico State University

Outline

- Knowledge Tracing
- Related Works
- Contrastive Learning
- Autoregressive Design
- Analysis of CL4KT Paper
- Analysis of LLM-KT Paper
- Possible Future Works
- Conclusion

Knowledge Tracing

- Transferring knowledge through teaching is one of the essential aspects for human intelligence.
- A human teacher can track the knowledge of students
- Rise of online education platforms
- Similar need for machines to track the knowledge of students and improve their learning experience.
- This is known as the **Knowledge Tracing (KT)** problem

Knowledge Tracing (Continued)

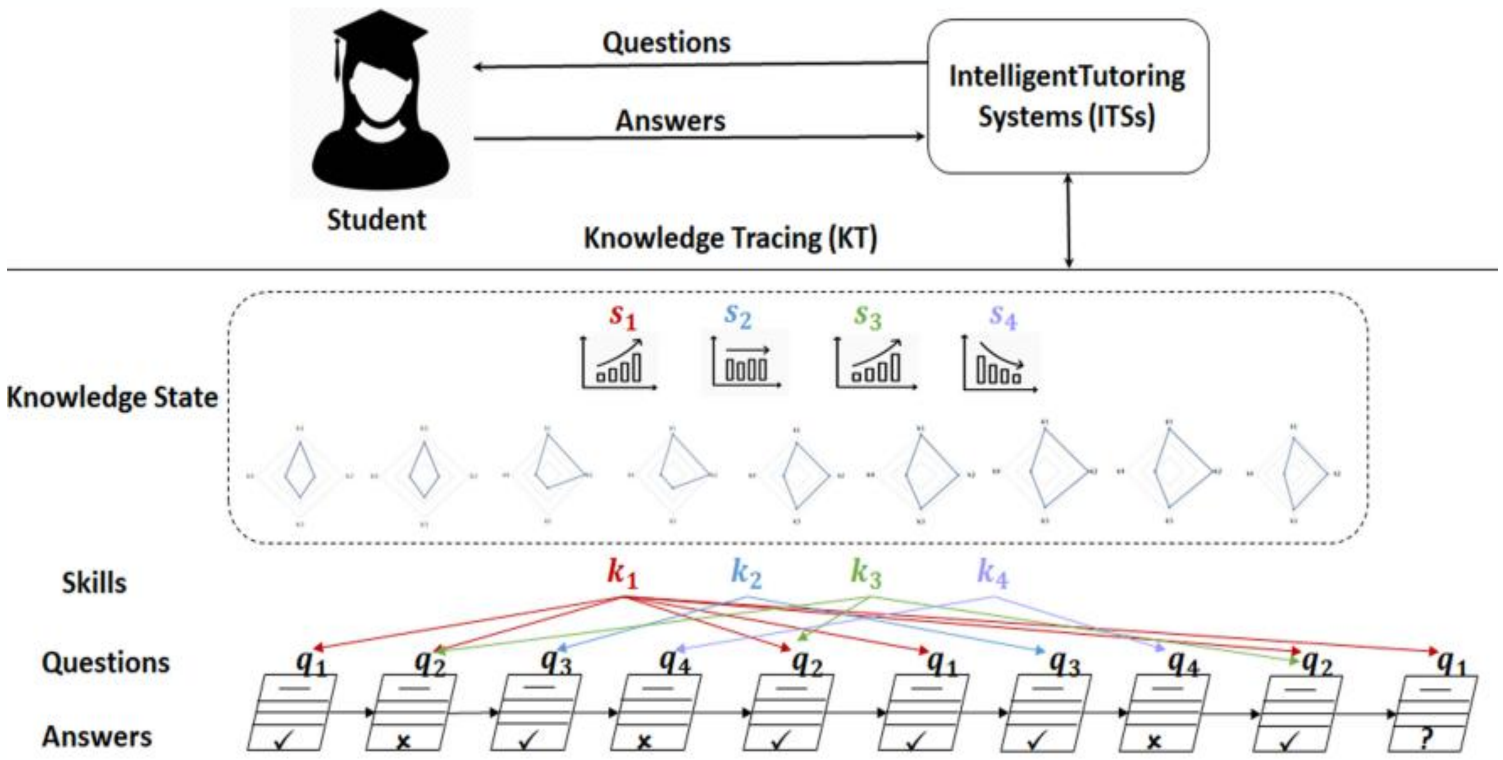


Figure.

An example scenario for knowledge tracing in an Intelligent Tutoring System

Reference: Knowledge Tracing: A Survey, Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes, ACM Computing Surveys, Vol. 55, No. 11, Article 22, 2023

Knowledge Tracing (Continued)

- Given sequences of learning interactions in online learning systems, knowledge tracing aims to:
  - **Monitor students' evolving knowledge states**
  - **Predict their performance on future exercises**

- The measured knowledge states can be further applied to:
  - **Individualize students' learning schemes**
  - **Maximize their learning efficiency**

Related Works (Deep Sequential Models)

Deep Sequential KT Models that utilize an autoregressive framework to dynamically monitor the knowledge states of students:

- **DKT:** utilizes an LSTM layer to simulate the learning processes of students.
- **DKT+:** enhanced DKT model by tackling the issues of inconsistency.
- **DKT-F:** enhanced original DKT by considering students' forgetting behaviors.
- **qDKT:** variant of DKT that models a student's response to individual questions over a period of time.
- **IEKT:** concurrently evaluates the knowledge states of students through the modules of cognitive understanding and knowledge acquisition.
- **QIKT:** enhances the interpretability of students' knowledge modeling by collectively learning its representation through a question-centric knowledge acquisition module and a problem-solving module.
- **AT-DKT:** introduces two auxiliary learning tasks, namely the question tagging prediction task and the individualized prior knowledge prediction task, to enhance the prediction performance of the DKT model.

Related Works (Attention-Based Models)

Attention-Based Models that capture dependencies between interactions via the attention mechanism:

- **SAKT:** employs self-attention to find the relevance between historical interactions and KCs.
- **SAINT:** Transformer-based model that encodes questions and responses in the encoder and decoder respectively.
- **AKT:** utilizes three self-attention modules to estimate the relevance between questions and historical interactions, and explicitly models a student's forgetting behavior through a monotonic attention mechanism.
- **simpleKT:** investigates the ordinary dot-product attention-based KT models by capturing the individual differences among questions that cover the same set of KCs.
- **sparseKT:** incorporates a k-selection module to only pick items with the highest attention scores to improve the attention-based DLKT approaches.

Related Works (Memory-based and Other Models)

Memory Augmented KT Models that capture latent relations between KCs and student knowledge states via memory networks:

- **DKVMN:** integrates a static matrix to preserve the relationships among KCs, and a dynamic matrix to monitor the evolving state of a student's knowledge.
- **DeepIRT:** incorporates DKVMN and item response theory to enhance the interpretability of the prediction output of DKVMN.

Other KT Models that do not belong to the above categories:

- **ATKT:** introduces adversarial perturbations into the student interaction sequences to improve the generalization capability.
- **GKT:** casts the knowledge structure as a graph and reformulates the KT task as a time series node-level classification problem in GNN.
- **HawkesKT:** utilizes the Hawkes process to model temporal cross-effects in student historical interactions

Contrastive Learning

- A self-supervised learning technique that learns by comparing samples.
- It teaches models to pull together similar representations
- And push apart dissimilar ones.
- No need for explicit labels
- The model learns from augmented views of the data.

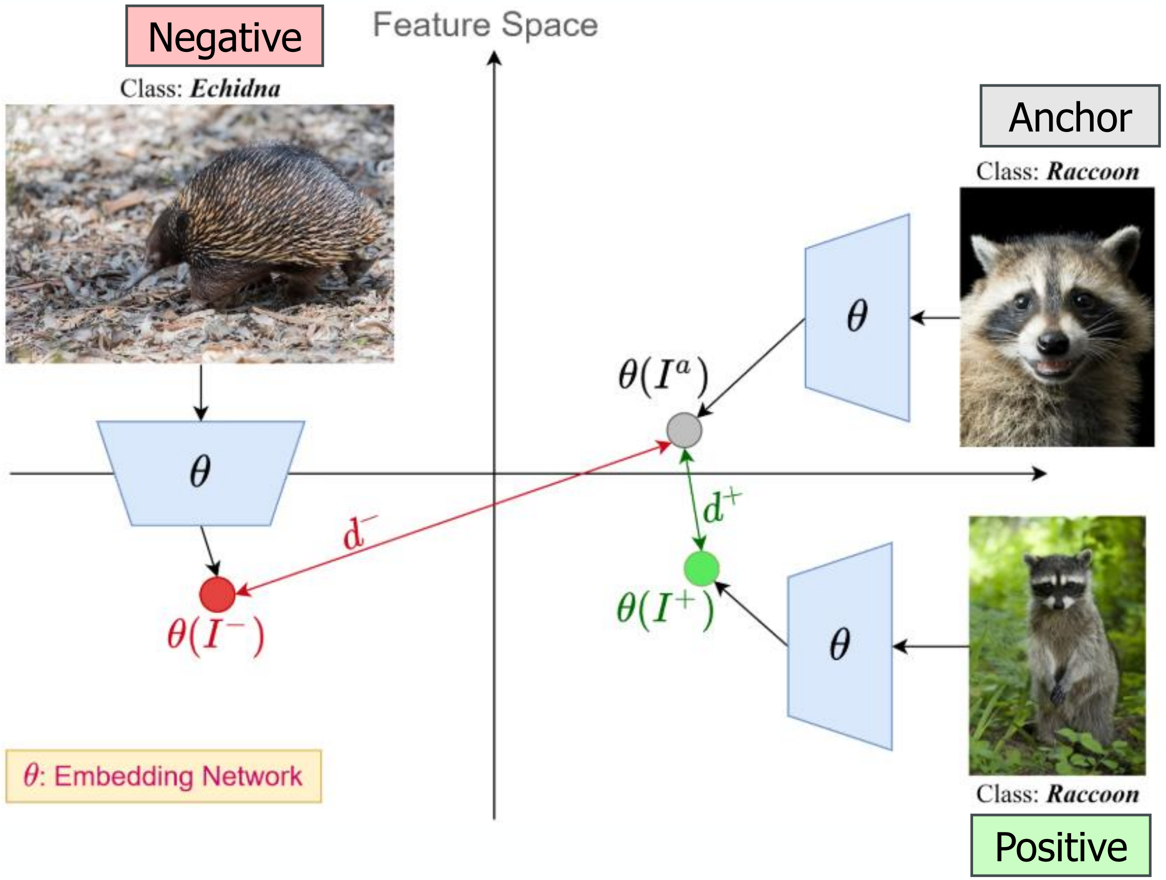


Figure. **An example scenario for Contrastive Learning**

Reference: Rohit Kundu, The Beginner's Guide to Contrastive Learning, 2022. https://www.v7labs.com/blog/contrastive-learning-guide

Contrastive Learning in KT

- Student modeling by contrasting similar vs. dissimilar learning patterns.
- How It Works:
  - Input sequence of question-response
  - Augmentations (Positive Pairs)
  - Hard Negatives (Negative Pairs)
- Objective is pull representations of similar histories together. Push apart those of dissimilar ones.
- Leads to more discriminative and generalizable representation.

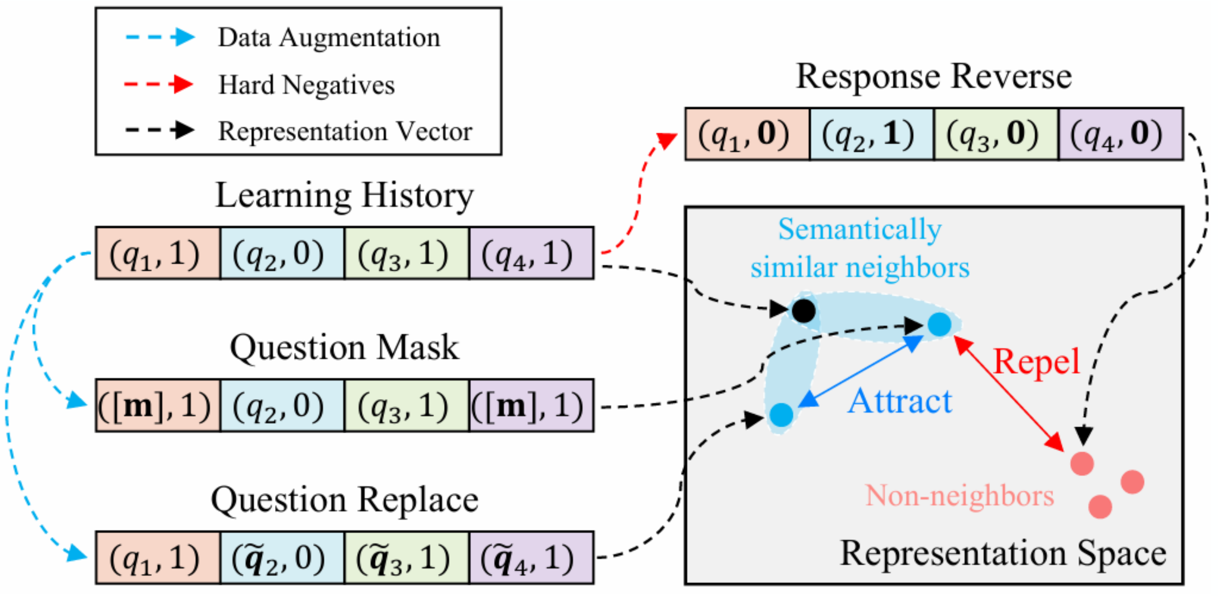


Figure. **An overview for Contrastive Learning in KT**

Reference: Rohit Kundu, The Beginner's Guide to Contrastive Learning, 2022. https://www.v7labs.com/blog/contrastive-learning-guide

Contrastive Learning for Knowledge Tracing

Wonsung Lee	Upstage, Republic of Korea
Kyoungsoo Park	I-Scream Edu Co. Ltd, Republic Of Korea
Jaeyoon Chun	I-Scream Edu Co. Ltd, Republic Of Korea
Sungrae Park	KAIST, Republic Of Korea
Youngmin Lee	Upstage, Republic of Korea

THE WEB CONFERENCE ACM

ACM Web Conference 2022 (WWW '22)

CL4KT Architecture

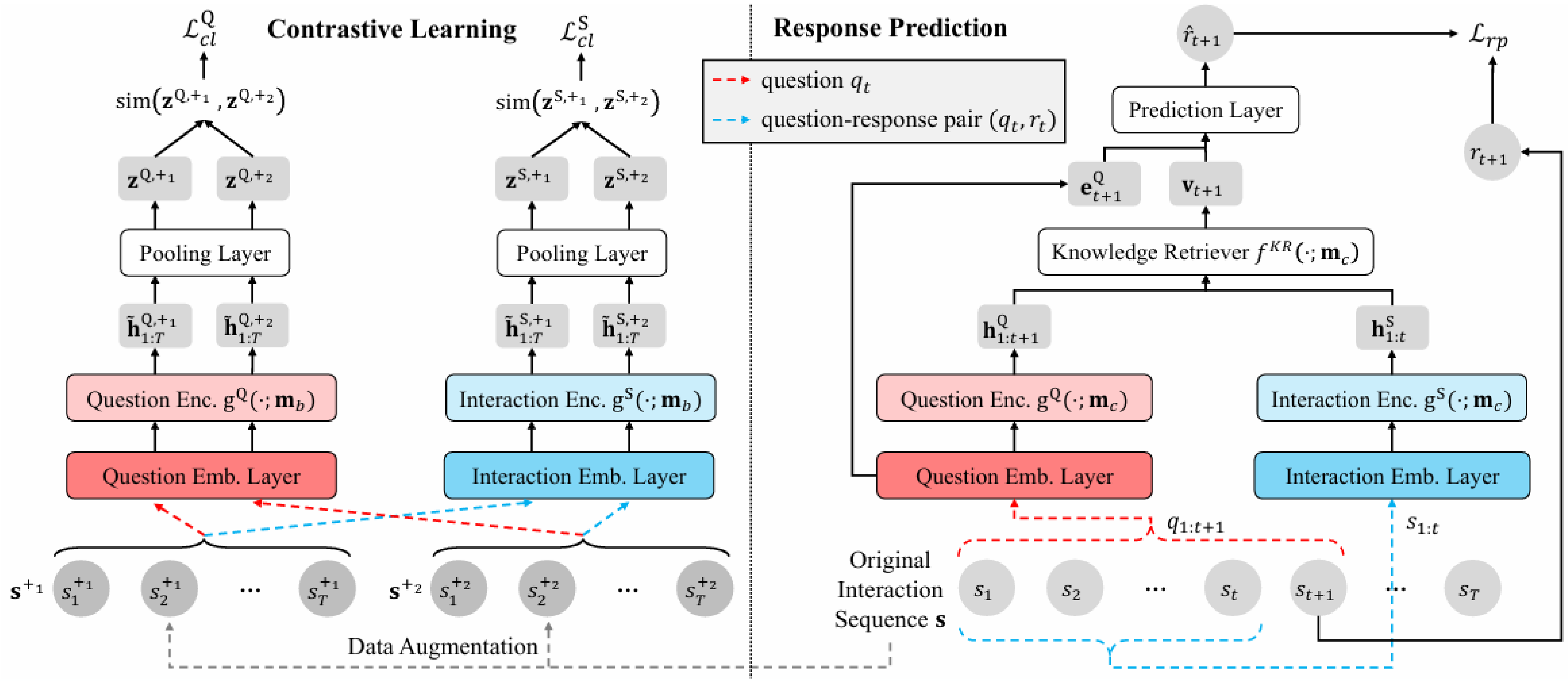


Figure.

Overall architecture of CL4KT.

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

CL4KT Architecture (Continued)

- **Input:** Two augmented sequences (s^{+1}, s^{+2}) from original history using masking, cropping, permuting, or replacing.
- **Embeddings:**
  - Question Emb. Layer → **Question Enc. g^Q**
  - Interaction Emb. Layer → **Interaction Enc. g^s**
- **Output Representations:**
  - z^{Q+1}, z^{Q+2} (question-based)
  - z^{s+1}, z^{s+2} (interaction-based)
- **Loss:**
  Contrastive loss L_{cl} pulls together similar views (augmented) and pushes apart unrelated ones:

$$\mathcal{L}_{cl} = \mathcal{L}_{cl}^{Q} + \mathcal{L}_{cl}^{S}$$

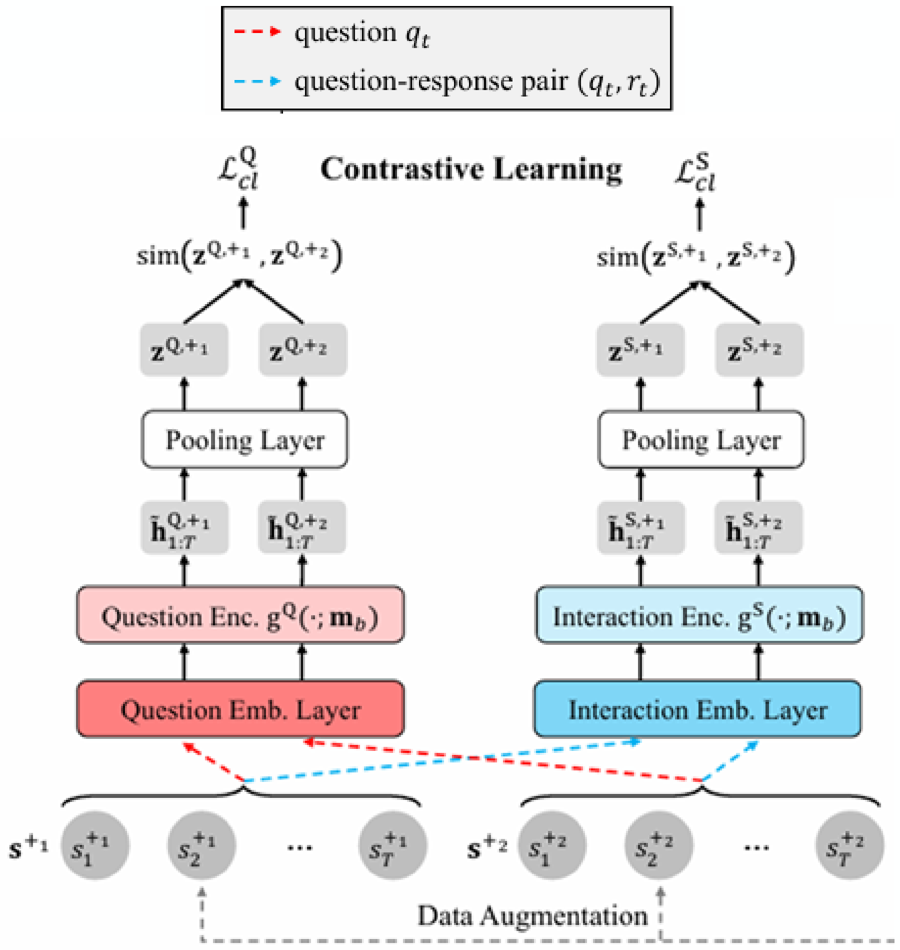
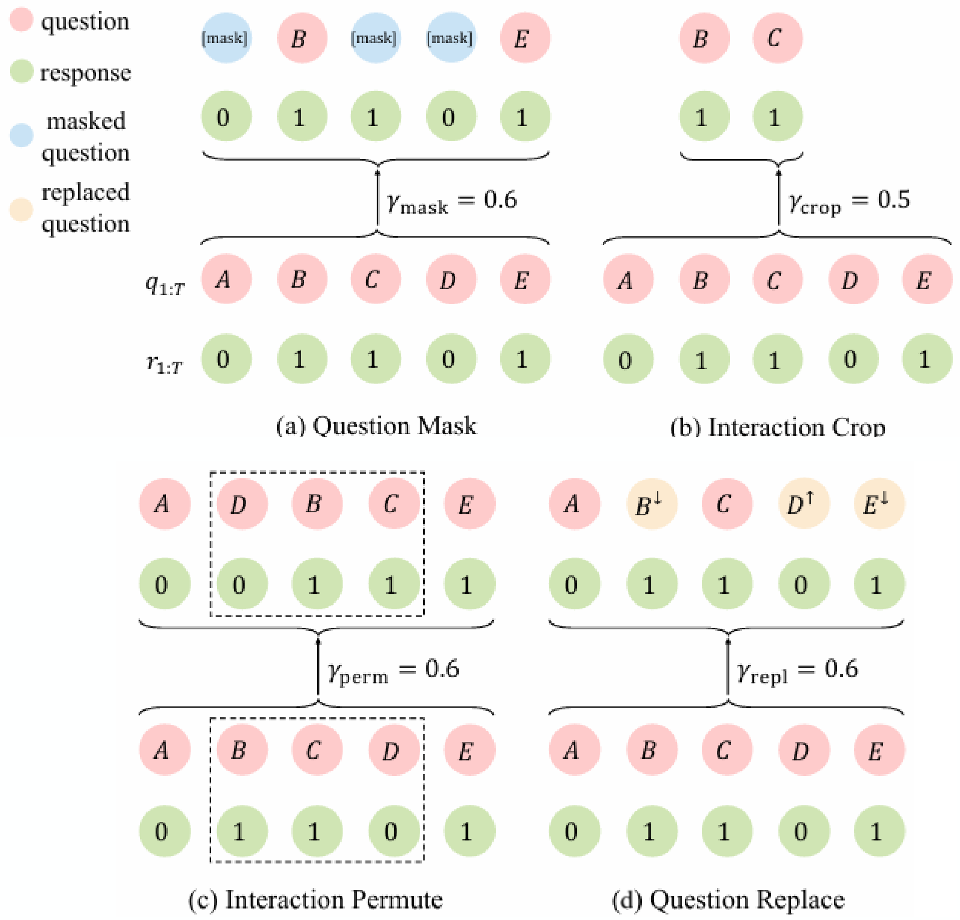


Figure. Contrastive Learning Network

Reference: *Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA*

Learning History Augmentation

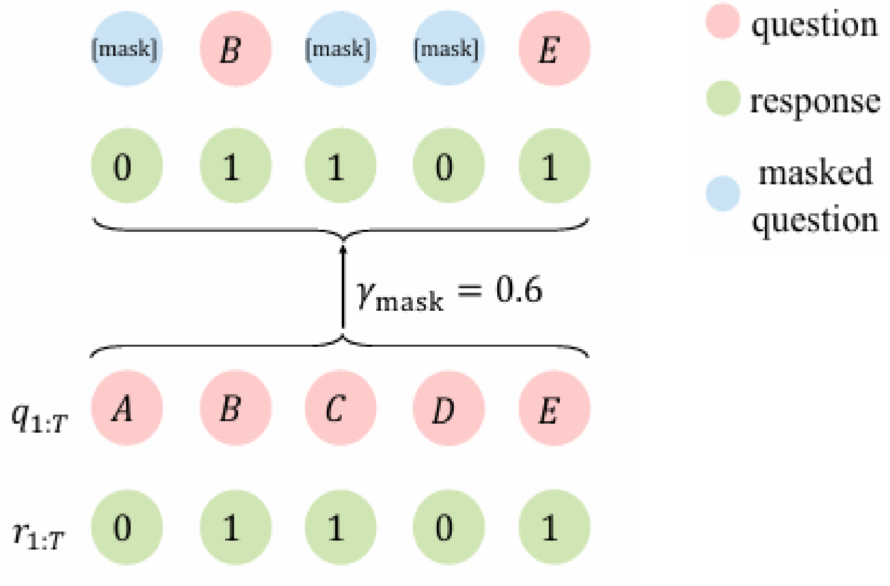


- Question Mask
- Interaction Crop
- Interaction Permute
- Question Replace

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

Augmentation: Question Mask

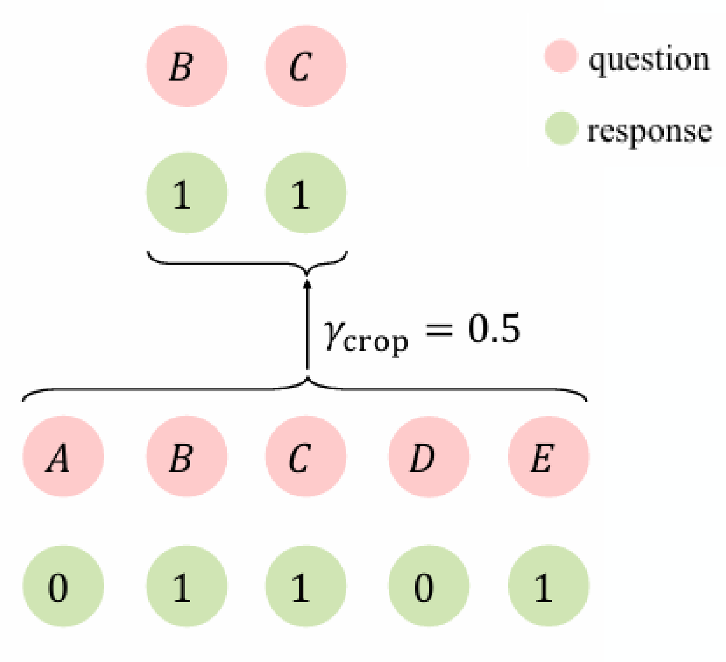
- Inspired by masked language modeling (like BERT).
- Randomly replaces some question IDs in the sequence with a [MASK] token.
- Responses remain unchanged to maintain learning semantics.
- Encourages the model to infer missing context, making it robust to sparse or incomplete data.



(a) Question Mask

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

Augmentation: Interaction Crop



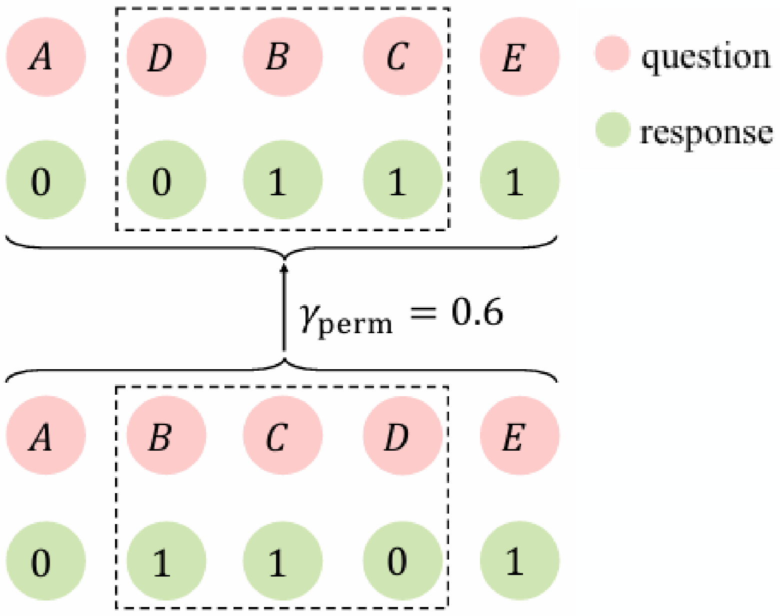
(b) Interaction Crop

- Extracts a random subsequence of student interactions.
- Provides a localized view of the learning history.
- Helps the model focus on short-term learning dynamics.
- Simulates real-world settings where only part of a student's data is available.

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

Augmentation: Interaction Permute

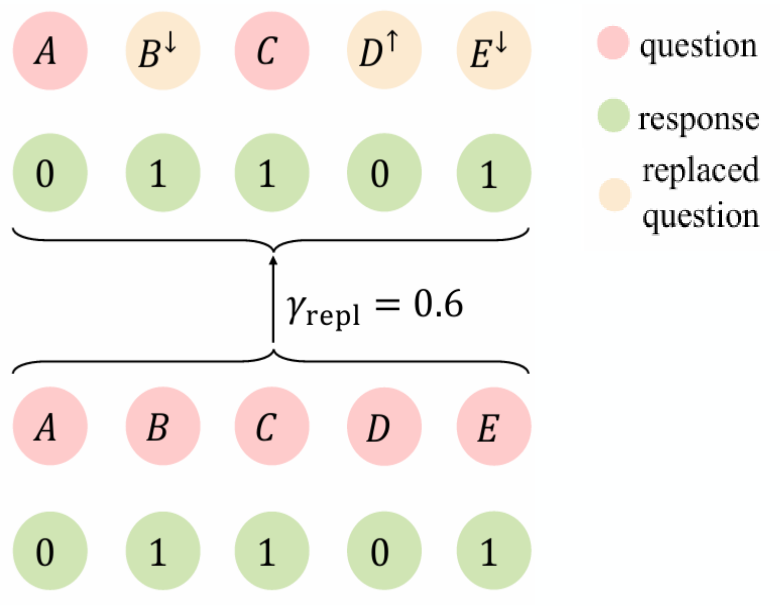
- Shuffles the order of questions in a continuous subsequence.
- Assumes order doesn't significantly change student proficiency.
- Tests the model's order invariance within short interaction windows.



(c) Interaction Permute

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

Augmentation: Question Replace

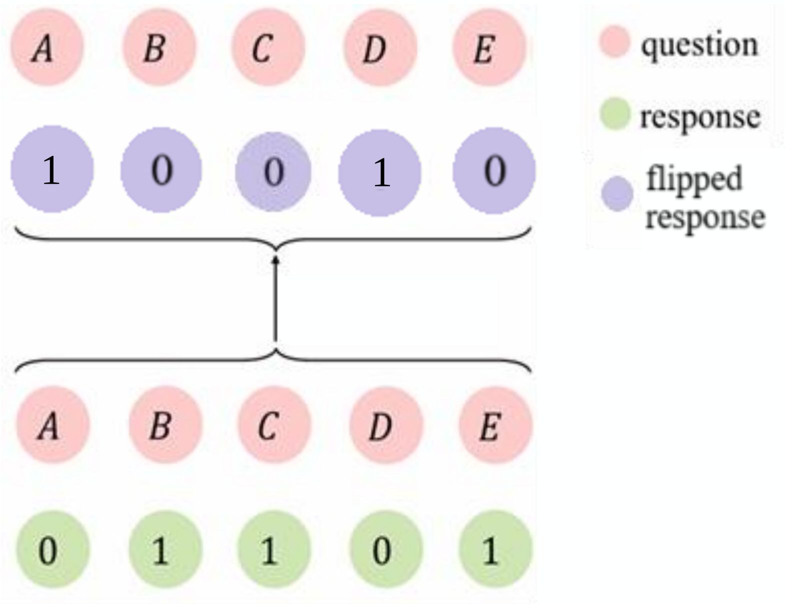


(d) Question replace

- Substitutes questions with easier or harder ones based on correctness.
- Guided by a difficulty-ranked question graph.
- Reflects pedagogical theory: mastery of hard topics implies mastery of easier ones.
- Simulates realistic variations in assessment difficulty while maintaining knowledge state.

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

Hard Negative Sampling



Hard Negative Sampling

- Introducing challenging negative samples that are semantically close to positive ones.
- Encouraging the model to learn finer distinctions between similar knowledge states.
- Created by flipping the response in one view
- Generate a realistic but misleading scenario.

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

Datasets

- **Algebra05 and algebra06:** Algebra questions
- **assist09:** ASSISTment intelligent tutoring system.
- **slepemapy:** An online system, slepemapy.cz, geography facts.
- **spanish:** Middle school students practicing spanish exercises.
- **statics:** A college level engineering statics course.

Datasets	#students	#questions	#skills	#interactions
algebra05	571	173,113	112	607,014
algebra06	1,138	129,263	493	1,817,450
assist09	3,695	17,728	112	282,071
slepemapy	5,000	2,723	1,391	625,523
spanish	182	409	221	578,726
statics	333	-	1,223	189,297

Research Questions Addressed

The study systematically explores 5 core questions to assess the performance and robustness of CL4KT:

- **RQ1**: How does CL4KT perform compared to state-of-the-art KT methods?
- **RQ2**: How do augmentation types and proportions affect CL4KT?
- **RQ3**: What is the impact of the contrastive learning loss hyperparameter (λ)?
- **RQ4**: How critical are the individual components (e.g., augmentation)?
- **RQ5**: Do the learned representations generalize well?

RQ1: Overall Performance

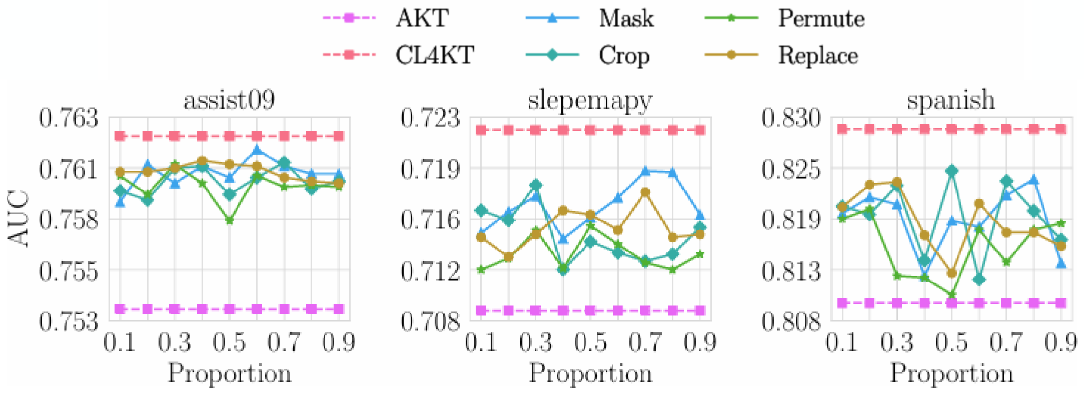
Comparison with Baselines

- CL4KT consistently outperforms all baselines (DKT, SAKT, AKT, DKVMN, etc.) across AUC and RMSE.
- AKT often yields second-best AUCs, thanks to additional question-skill relations.
- CL4KT's self-supervised signals enhance knowledge representation, even without auxiliary inputs.

Dataset	Metric	IRT	PFA	DKT	DKVMN	SAKT	AKT	CL4KT
algebra05	AUC	0.7141	0.7481	0.7636	0.7562	0.7637	0.7676	**0.7891**
	RMSE	0.4005	0.3932	0.3921	0.3907	0.3899	0.3952	**0.3815**
algebra06	AUC	0.6559	0.7460	0.7589	0.7463	0.7512	0.7474	**0.7733**
	RMSE	0.4025	0.3848	0.3820	0.3864	0.3862	0.3896	**0.3791**
assist09	AUC	0.6708	0.7284	0.7504	0.7475	0.7491	0.7532	**0.7624**
	RMSE	0.4631	0.4444	0.4371	0.4375	0.4381	0.4372	**0.4333**
slepemapy	AUC	0.6210	0.6583	0.6986	0.7064	0.6846	0.7090	**0.7218**
	RMSE	0.4068	0.4020	0.3978	0.3962	0.4062	0.3978	**0.3926**
spanish	AUC	0.6956	0.7467	0.8066	0.8027	0.8065	0.8097	**0.8289**
	RMSE	0.4596	0.4428	0.4139	0.4156	0.4179	0.4177	**0.4049**
statics	AUC	0.7404	0.7489	0.7674	0.7736	0.7492	0.7872	**0.7943**
	RMSE	0.4303	0.4096	0.4111	0.3975	0.4105	0.3967	**0.3945**

Table: Performance comparison among CL4KT and Other benchmarks

RQ2: Effect of Augmentation



Impact of Augmentation Types & Proportions

- CL4KT using any single augmentation surpasses AKT across datasets.
- Best-performing augmentation varies:
  - **Question Masking** (Assist09, Slepemapy)
  - **Interaction Cropping** (Spanish)
- Too high/low augmentation rates hurt performance and optimal rates yield peak AUCs.

Aug.	Method	assist09	slepemapy	spanish
None	DKT	0.7504	0.6986	0.8066
	SAKT	0.7491	0.6846	0.8065
Mask	DKT_{cl}	0.7574	0.7051	0.8108
	$SAKT_{cl}$	0.7505	0.6904	0.8123
	CL4KT	**0.7617**	**0.7189**	0.8234
Crop	DKT_{cl}	0.7571	0.7047	0.8135
	$SAKT_{cl}$	0.7512	0.6879	0.8126
	CL4KT	0.7610	0.7179	**0.8243**
Permute	DKT_{cl}	0.7575	0.7042	0.8122
	$SAKT_{cl}$	0.7510	0.6884	0.8141
	CL4KT	0.7609	0.7150	0.8201
Replace	DKT_{cl}	0.7572	0.7049	0.8135
	$SAKT_{cl}$	0.7507	0.6875	0.8113
	CL4KT	0.7611	0.7174	0.8231

Table: **Performance of DKT, SAKT, DKT_{cl}, $SAKT_{cl}$, and CL4KT with an individual data augmentation**

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

RQ3: Role of Contrastive Loss

- Increasing λ too high overwhelms supervised objective and harms AUC.
- Optimal λ balances RP and CL — avoids over-regularization.
- Enhanced baselines (DKT_{cl}, $SAKT_{cl}$) with CL show consistent improvements over originals.
- CL4KT with single augmentation still beats CL-augmented baselines.

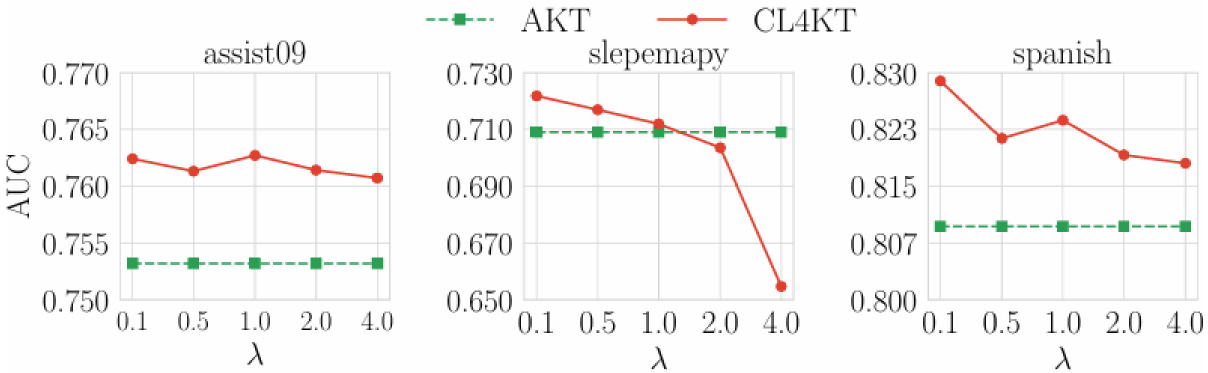


Figure: **Performance comparison with respect to λ**

Aug.	Method	assist09	slepemapy	spanish
None	DKT	0.7504	0.6986	0.8066
	SAKT	0.7491	0.6846	0.8065
Mask	DKT_{cl}	0.7574	0.7051	0.8108
	$SAKT_{cl}$	0.7505	0.6904	0.8123
	CL4KT	**0.7617**	**0.7189**	0.8234
Crop	DKT_{cl}	0.7571	0.7047	0.8135
	$SAKT_{cl}$	0.7512	0.6879	0.8126
	CL4KT	0.7610	0.7179	**0.8243**
Permute	DKT_{cl}	0.7575	0.7042	0.8122
	$SAKT_{cl}$	0.7510	0.6884	0.8141
	CL4KT	0.7609	0.7150	0.8201
Replace	DKT_{cl}	0.7572	0.7049	0.8135
	$SAKT_{cl}$	0.7507	0.6875	0.8113
	CL4KT	0.7611	0.7174	0.8231

Table: **Performance of DKT, SAKT, DKT_{cl}, $SAKT_{cl}$, and CL4KT with an individual data augmentation**

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

RQ4: Ablation Study

- Removing each augmentation reduces performance — Replace removal has largest drop.
- Without bidirectional encoding, AUC drops → shows context is vital.
- Hard negative sampling significantly improves model generalization.
- Multicomponent augmentation + bidirectional encoder + hard negatives = key to success.

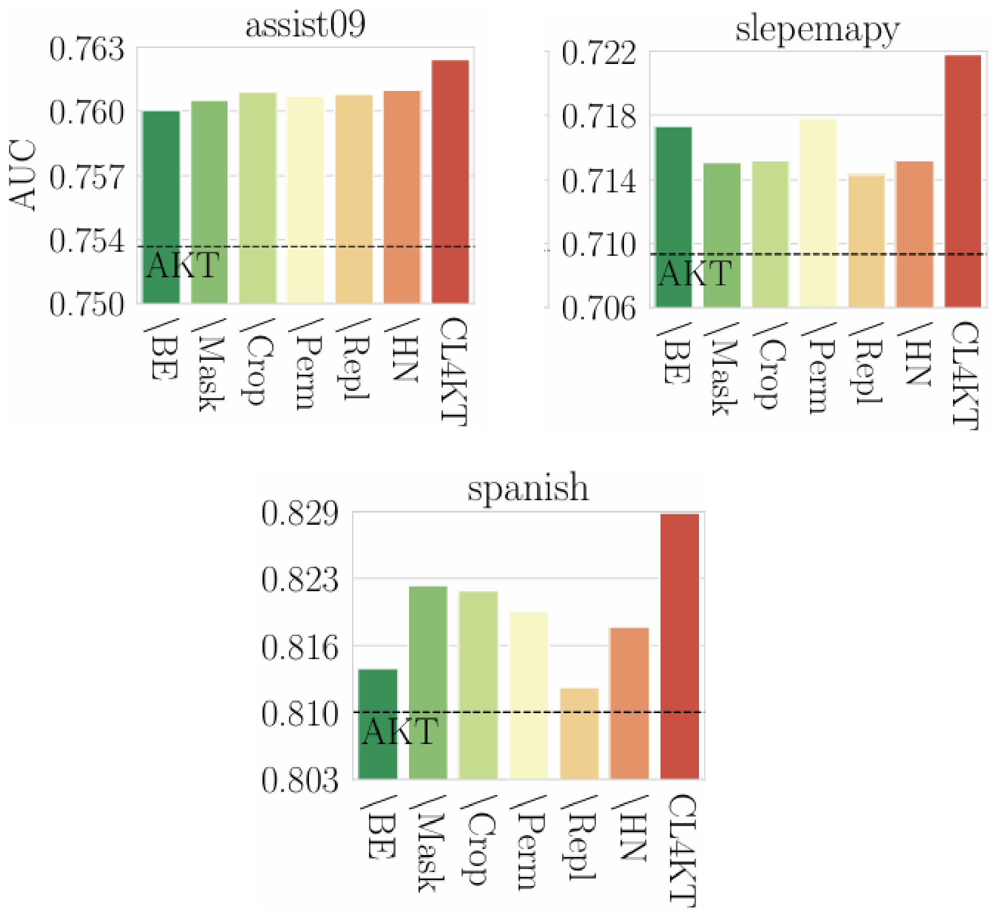
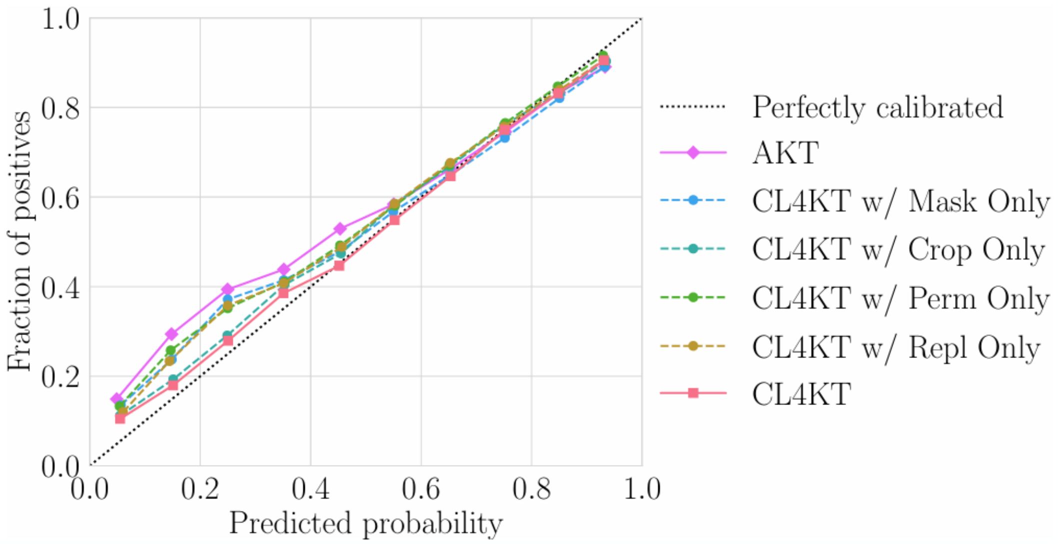


Figure: **Performance comparison between CL4KT and its variants**

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

RQ5: Quality of Representations

- CL4KT exhibits better calibration (predicted probability vs. observed accuracy) than AKT.
- Ensures more reliable prediction for downstream educational decisions.
- Lower uniformity values = better quality embeddings (more evenly distributed).
- Both question and interaction embeddings from CL4KT outperform AKT in uniformity.



Uniformity	Method	assist09	slepemapy	spanish
Question	AKT	-2.920	-3.210	-1.337
	CL4KT	-2.954	-3.226	-1.382
Interaction	AKT	-3.143	-3.444	-1.977
	CL4KT	-3.185	-3.468	-2.097

Table: Uniformity values of learned representations from AKT and CL4KT
(Lower numbers are better)

Reference: Wonsung Lee, et al.. 2022. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022 (WWW '22), New York, NY, USA

Key Takeaways

- Contrastive learning enhances KT model robustness under data sparsity.
- Pedagogically meaningful augmentations improve representation quality.
- Hard negative sampling sharpens student representation discrimination.
- CL4KT generalizes across datasets without domain-specific tuning.

Reference: Sketchbubble

Strengths and Weaknesses

✅ Contrastive approach leads to higher AUC and better calibrated predictions.

✅ Augmentation strategies are interpretable.

✅ Bi-directional Transformer encoder captures global learning context.

✅ Compatible with existing KT models like DKT and SAKT (via augmentation).

✅ Empirical gains validated via 5-fold CV across diverse datasets.

❌ Requires careful hyperparameter tuning (e.g., γ for augmentations, λ for loss weighting).

❌ Question permutation assumes order-invariance, which may not hold in all learning tasks.

❌ Limited applicability tested beyond math-based datasets like Assist09 or Algebra05..

Knowledge Tracing as Language Processing: A Large-Scale Autoregressive Paradigm

Bojun Zhan	Jinan University, China
Teng Guo	Jinan University, China
Xueyi Li	Jinan University, China
Mingliang Hou	TAL Education Group, China
Qianru Liang	Jinan University, China
Boyu Gao	Jinan University, China
Weiqi Luo	Jinan University, China
Zitao Liu	Jinan University, China

AIED RECIFE 2024 International Conference on Artificial Intelligence in Education (AIED 2024)

LLM-KT Architecture

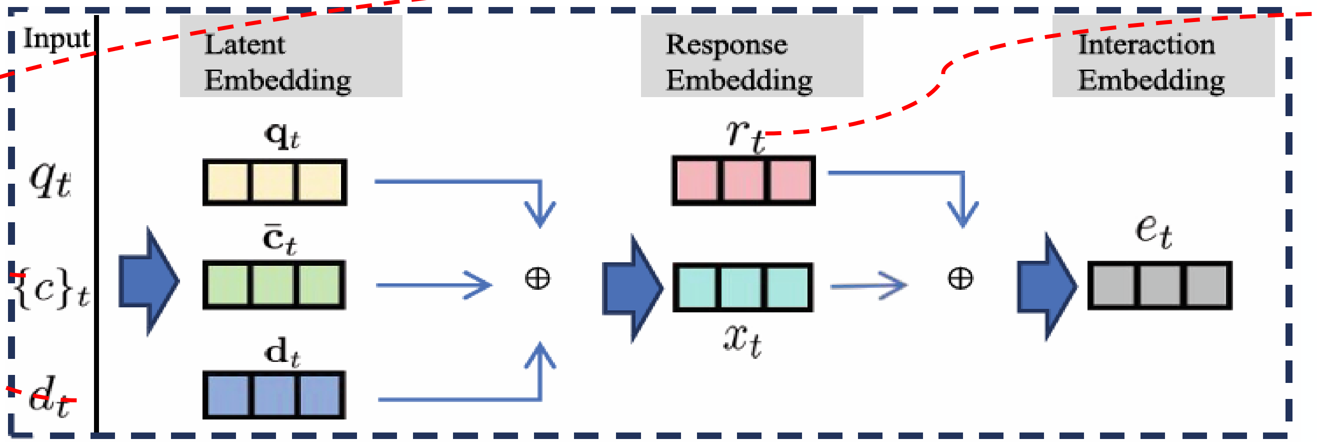


Figure 2. Interaction encoder of LLM-KT.

Reference: Bojun Zhan et al. 2024. Knowledge Tracing as Language Processing: A Large-Scale Autoregressive Paradigm. Springer Nature Switzerland, AIED 2024, Recife, Brazil.

🔢 **Input Components:**

- **q_t:** Question ID at time t
- **$\{c\}_t$:** Knowledge concept(s) associated with q_t
- **d_t:** Database ID of the question
- **r_t:** Response (correct/incorrect)

🧠 **Embedding Layers:**

Latent Embedding

- q_t → Question Embedding **($\boldsymbol{q_t}$)**
- $\{c\}_t$ → Concept Embedding **($\hat{\boldsymbol{c}}_t$)**
- d_t → Database Embedding **($\boldsymbol{d_t}$)**
- Combine → $x_t = \hat{c}_t + d_t + q_t$

Response Embedding

- r_t → Response Embedding **($\boldsymbol{r_t}$)**
- Combine → $e_t \rightarrow x_t + r_t$

Interaction embedding

- Output → e_t

LLM-KT Architecture (Continued)

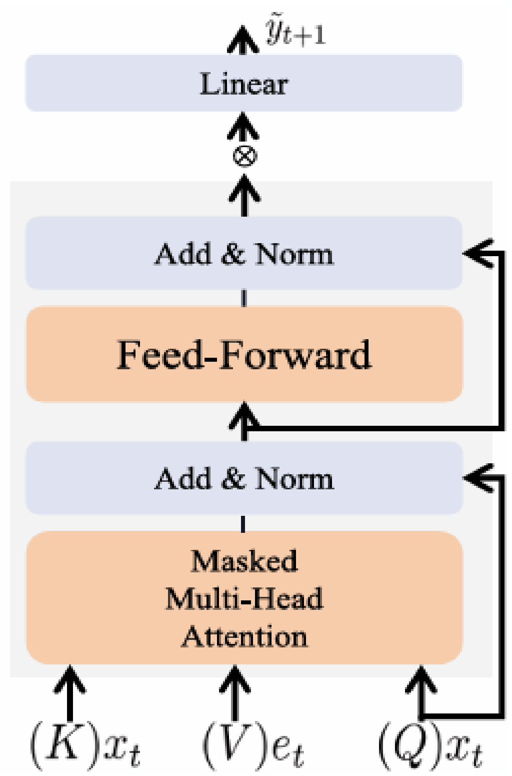


Figure 2. Forward procedure of LLM-KT.

- Input:
  - e_t → Interaction embedding
  - x_t → Position-aware token embedding
- Masked Multi-Head Attention ensures autoregressive
- Feedforward Layer applies nonlinear transformation
- Add & Norm Layers stabilize training via residual connections and normalization
- Linear Output Layer → Final prediction $\tilde{y}_{t+1}$ for the next question's response

Reference: Bojun Zhan et al. 2024. Knowledge Tracing as Language Processing: A Large-Scale Autoregressive Paradigm. Springer Nature Switzerland, AIED 2024, Recife, Brazil.

Research Questions Addressed

The study systematically explores 2 core questions to assess the performance of LLM architecture in KT:

- **RQ1:** Is it feasible to apply the LLM-like architectures in the field of KT?

  → **Can a decoder-only Transformer (like GPT) effectively model a student's knowledge over time?**

- **RQ2:** Can increasing model size improve performance in KT?

  → **Does scaling the number of parameters (334M to 7B) consistently enhance prediction accuracy?**

RQ1: Overall Performance

- LLM-KT-334M outperformed all 18 benchmark KT models in AUC
- Achieved **9.7%** average AUC improvement over attention-based deep KT models like AKT and SAKT.
- Simplified autoregressive decoder outperformed complex models (e.g., IEKT, LPKT) despite its minimal architecture

Method	Model Type	AUC	Accuracy
DKT	Sequential	0.6649	0.6775
DKT+	Sequential	0.6579	0.6756
DKT-F	Sequential	0.6768	0.6834
DKVMN	Memory	0.6518	0.6689
ATKT	Other	0.6675	0.6786
GKT	Other	0.6469	0.6676
SAKT	Attention	0.6539	0.6722
SAINT	Attention	0.7047	0.6709
AKT	Attention	0.7065	0.6788
HawkesKT	Others	0.7376	0.7113
DeepIRT	Memory	0.6485	0.6701
LPKT	Sequential	0.7758	0.7321
IEKT	Sequential	0.7787	0.7335
qDKT	Sequential	0.7689	0.7273
AT-DKT	Sequential	0.6902	0.6892
simpleKT	Attention	0.6990	0.6699
QIKT	Sequential	0.7715	0.7292
sparseKT	Attention	0.6930	0.6698
LLM-KT-334M	Attention	**0.7933**	**0.7420**
LLM-KT-794M	Attention	0.7928	0.7415
LLM-KT-1.8B	Attention	0.7900	0.7401
LLM-KT-4B	Attention	0.7761	0.7312
LLM-KT-7B	Attention	0.7762	0.7312

Table: Performance comparison among LLM-KT and other benchmarks

Reference: Bojun Zhan et al. 2024. Knowledge Tracing as Language Processing: A Large-Scale Autoregressive Paradigm. Springer Nature Switzerland, AIED 2024, Recife, Brazil.

RQ2: Impact of Model Size

- AUC plateaued beyond 1.8B.
- LLM-KT-4B and 7B even underperformed due to overfitting (loss & AUC dropped together)
- Suggests data scale must match model scale. EdNet, while large for KT, is small compared to LLM training data in NLP
- Optimal performance achieved by mid-sized models (334M-1.8B)

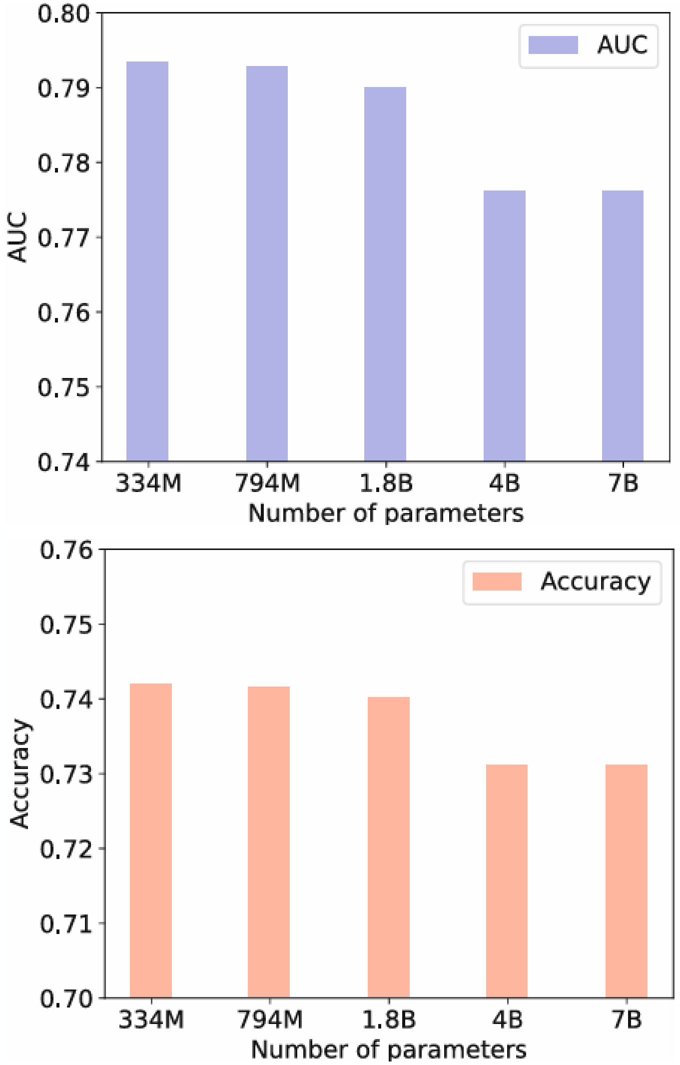


Figure: **AUC and ACC vs Number of parameters of LLM-KT Model**

Reference: Bojun Zhan et al. 2024. Knowledge Tracing as Language Processing: A Large-Scale Autoregressive Paradigm. Springer Nature Switzerland, AIED 2024, Recife, Brazil.

Key Takeaways

- Large Language Models can be effectively adapted for Knowledge task.
- Even simplified architectures (decoder-only transformers) outperform complex KT models.
- The autoregressive paradigm, combined with large-scale modeling, achieves state-of-the-art AUC performance, demonstrating the viability of LLM-based KT frameworks.

Reference: Sketchbubble

Strengths and Weaknesses

✅ Scales with model size — LLM-KT-334M to 1.8B

✅ Simple architecture with autoregressive self-attention

✅ Avoids rigid KC-question linking (unlike traditional KT)

✅ Learns implicit knowledge transitions from raw sequence data.

❌ Limited dataset diversity could affect generalizability.

❌ The approach does not yet address student behavior variance (e.g., forgetting)

❌ No ablation studies on the impact of question, KC, or dataset embeddings individually.

Possible Future Directions

- Develop real-time adaptive augmentation tuned to student behavior.
- Apply LLM-KT in few-shot and cold-start settings for new students or topics.
- Generate interpretable feedback using learned embeddings for student guidance.
- Conduct longitudinal classroom studies to assess real-world impact.
- Transfer contrastive models across subjects like math, language, and coding.
- Enhance explainability by visualizing contrastive embedding dynamics.

Reference: Sketchbubble

Conclusion

- CL4KT introduces contrastive learning with educationally meaningful augmentations, significantly boosting performance under sparse conditions.
- LLM-KT reframes KT as language modeling, showing that autoregressive transformers can effectively model student learning trajectories.
- Both works demonstrate the power of representation learning in KT.
- Future research should integrate qualitative validation, multi-modal data, and adaptive augmentation for real-world impact.

Thank you!