Quality meets Diversity: A Model-Agnostic Framework for Computerized Adaptive Testing

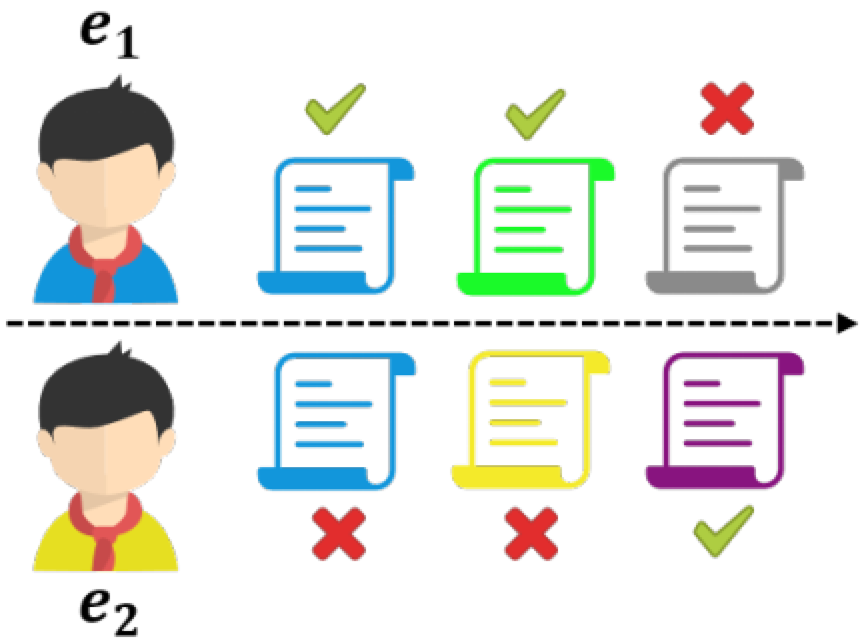
Haoyang Bi[1], Haiping Ma[2], Zhenya Huang[1,*], Yu Yin[1], Qi Liu[1], Enhong Chen[1], Yu Su[3], Shijin Wang[3]
[1]Anhui Province Key Laboratory of Big Data Analysis and Application
School of Computer Science and Technology, University of Science and Technology of China
{bhy0521, yxonic}@mail.ustc.edu.cn, {huangzhy, qiliuql, cheneh}@ustc.edu.cn
[2]Anhui University, hpma@ahu.edu.cn
[3]IFLYTEK Research, {yusu, sjwang}@iflytek.com

Reporter: Haoyang Bi

Outline

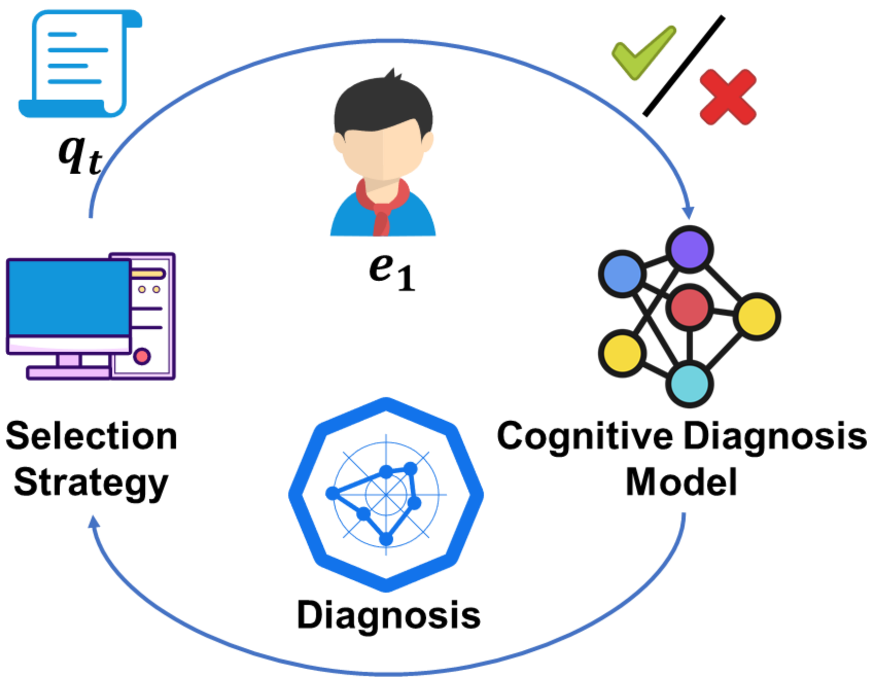
Background

- Adaptive learning makes education better
  - AI liberates manpower
- Adaptive testing makes examination better
  - Reasonable tests discover rather than compare
  - Tailored tests make more accurate, more efficient



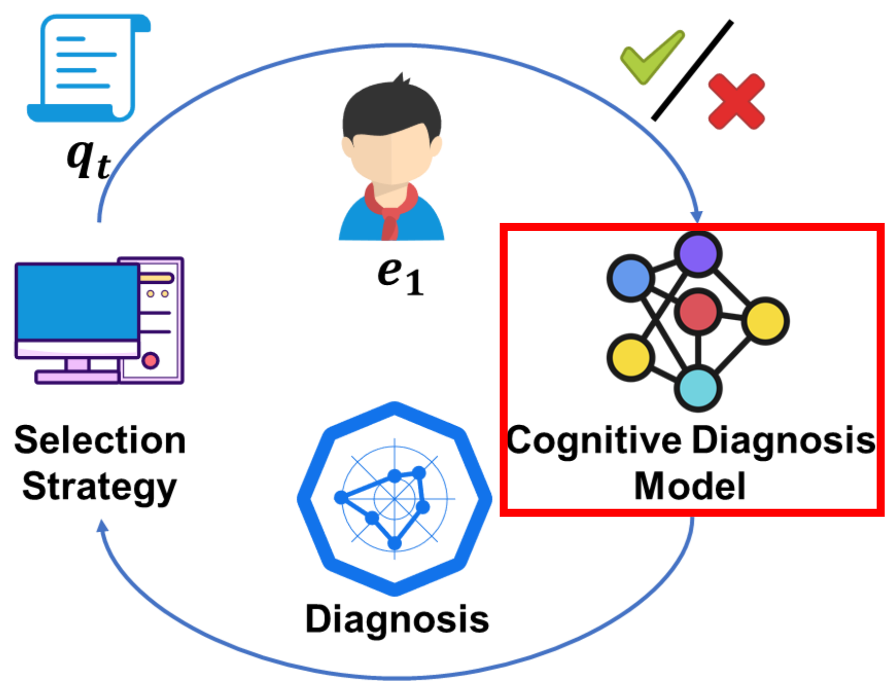
Background

- Cognitive Diagnosis Model
  - Learn about the examinee through answered questions
- Selection Strategy
  - Select questions for the examinee to answer
- Objective
  - Design an optimal strategy to make the CDM learn accurately and efficiently



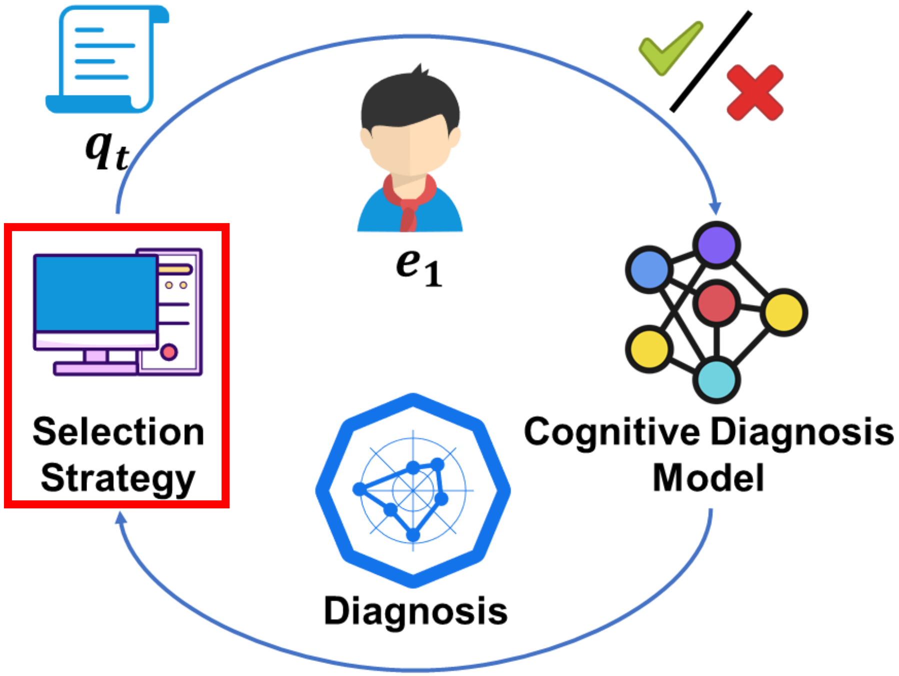
Background

- Cognitive Diagnosis Model
  - Learn about the examinee through answered questions
- Selection Strategy
  - Select questions for the examinee to answer
- Objective
  - Design an optimal strategy to make the CDM learn accurately and efficiently



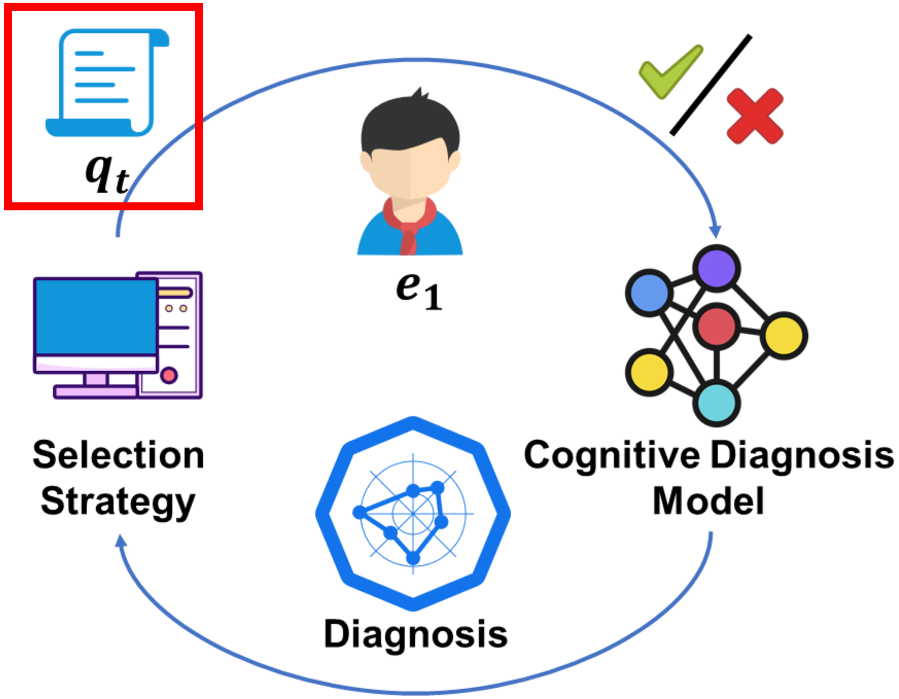
Background

- Cognitive Diagnosis Model
  - Learn about the examinee through answered questions
- Selection Strategy
  - Select questions for the examinee to answer
- Objective
  - Design an optimal strategy to make the CDM learn accurately and efficiently



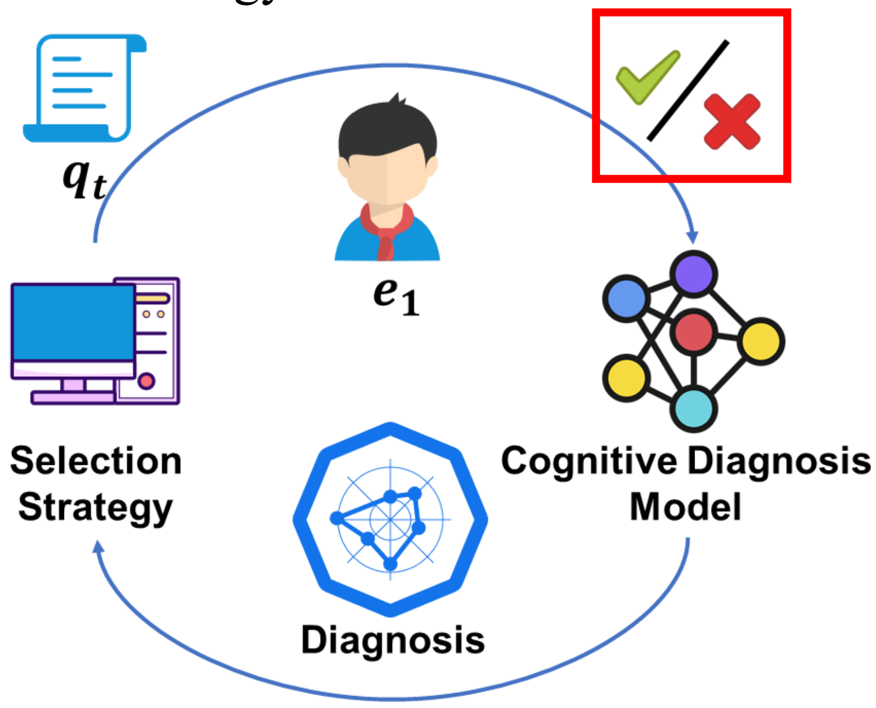
Background

- Cognitive Diagnosis Model
  - Learn about the examinee through answered questions
- Selection Strategy
  - Select questions for the examinee to answer
- Objective
  - Design an optimal strategy to make the CDM learn accurately and efficiently



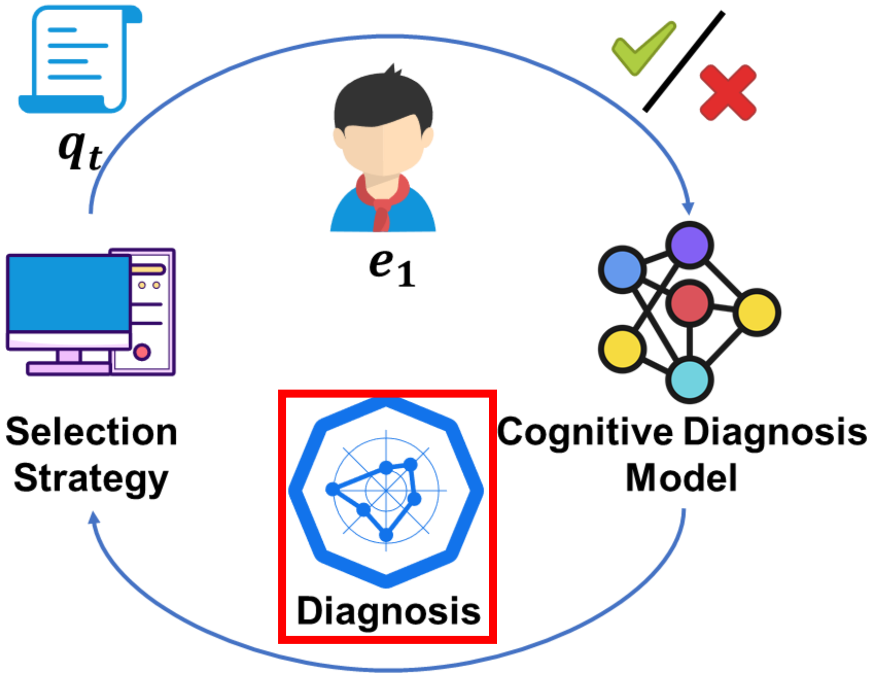
Background

- Cognitive Diagnosis Model
  - Learn about the examinee through answered questions
- Selection Strategy
  - Select questions for the examinee to answer
- Objective
  - Design an optimal strategy to make the CDM learn accurately and efficiently



Background

- Cognitive Diagnosis Model
  - Learn about the examinee through answered questions
- Selection Strategy
  - Select questions for the examinee to answer
- Objective
  - Design an optimal strategy to make the CDM learn accurately and efficiently



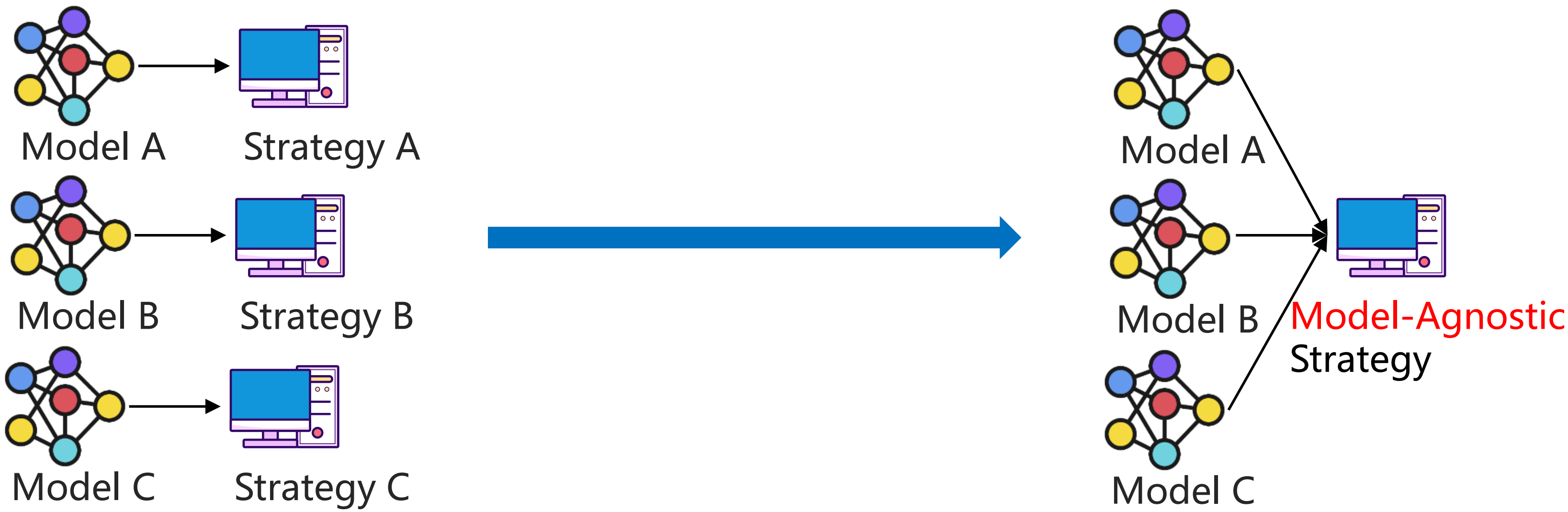
Background

- Related work
  - Cognitive Diagnosis Model
    - Item Response Theory Model (IRT)
    - Multidimensional IRT (MIRT)
    - Neural Cognitive Diagnosis Model (NCD)
  - Question Selection Strategy
    - Maximum Fisher Information (MFI) → IRT
    - Kullback-Leibler Information (KLI) → IRT
    - D-Optimality → MIRT
    - Multivariate Kullback-Leibler Information (MKLI) → MIRT
    - ? → NCD

IRT $$p(\boxed{\theta}) = \frac{1}{1 + e^{-a(\theta - b)}} \Longrightarrow I(a, b|\theta) = \frac{\left[\frac{\partial p(\theta)}{\partial \theta}\right]^2}{p(\theta)(1 - p(\theta))} \Longrightarrow \min SE(\theta)$$

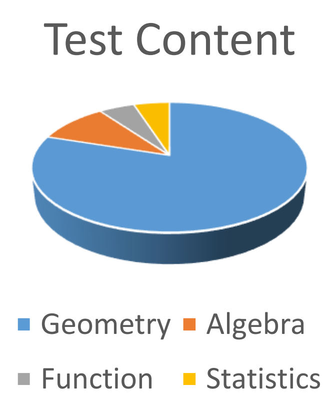
Background

- Challenge I: Model-specific vs Model-agnostic



- Challenge II: Quality vs Diversity

Nothing useful if the student will **definitely** answer the question right/wrong

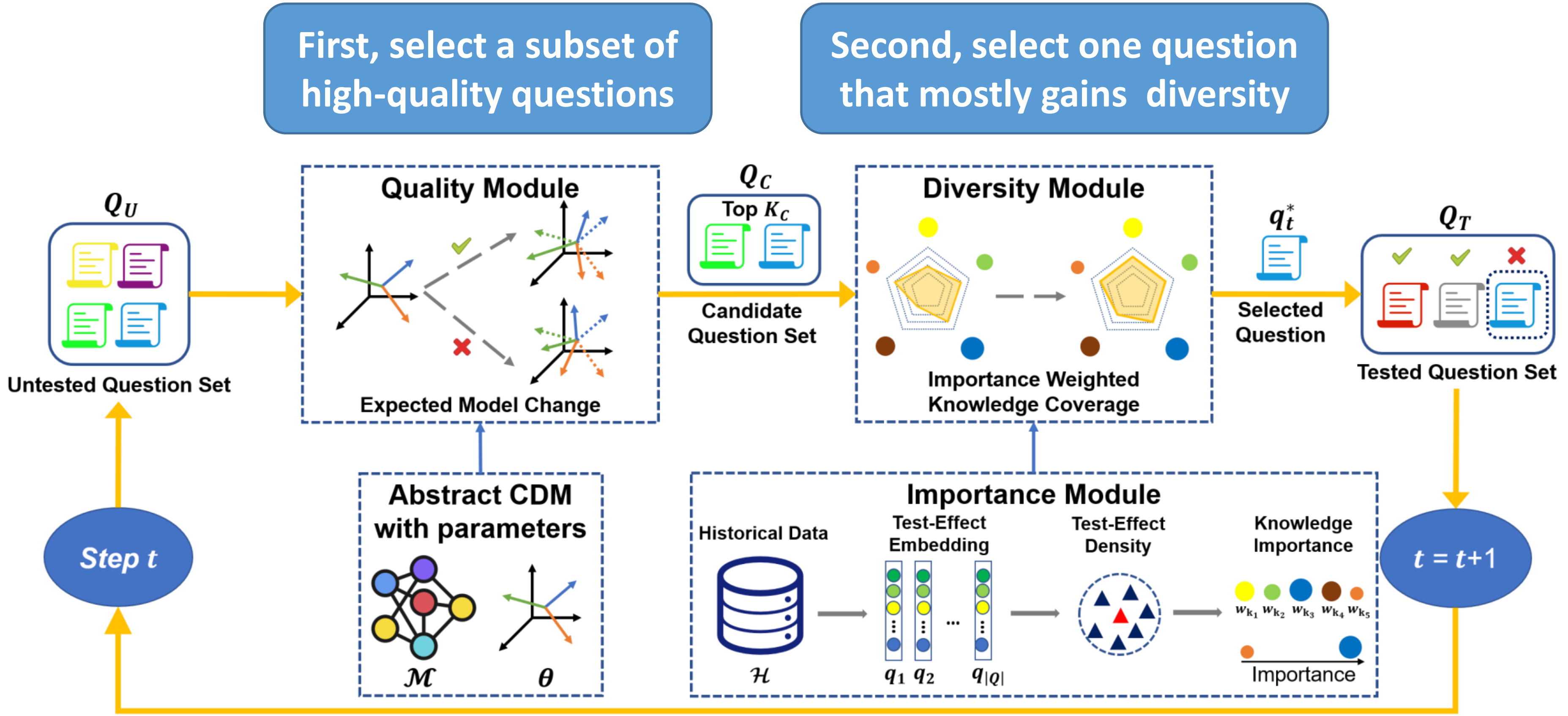


Biased results if the knowledge coverage is **imbalanced**

Outline

MAAT Framework

- Model-Agnostic Adaptive Testing (MAAT)
  - Work with an abstract model to achieve model-agnostic
  - Aim at the two high-level objective in a two-stage method

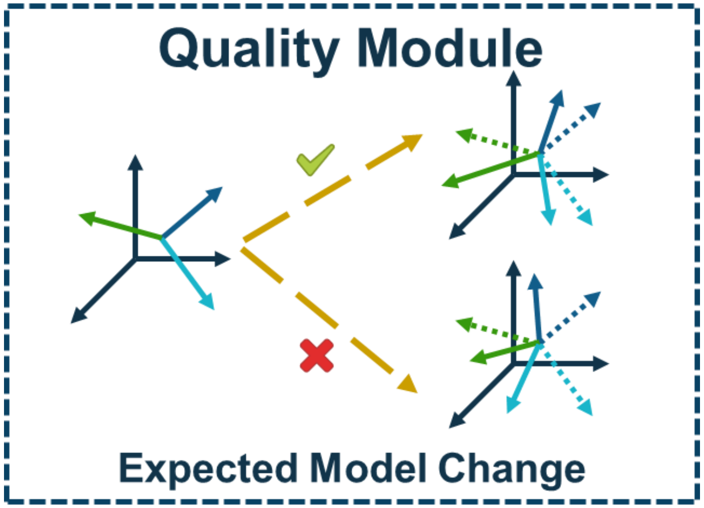


MAAT Framework

- Quantify the quality of a question
  - How we measure the information the abstract CDM obtains
- Expected Model Change (EMC)
  - Idea: the more information, the more change in model
  - Though with an abstract model, we can still observe its numerical param values
- Challenge: change is unknown before the result
  - Solution: estimate with expectation

$$\mathrm{EMC}(q_j) = \mathbb{E}_{a_{ij}\sim p}\Delta\mathcal{M}(< e_i, q_j, a_{ij} >)$$

- Example
  - If q is answered correctly, $\Delta\theta$=0.8
  - If q is answered wrongly, $\Delta\theta$=0.2
  - $p(right)$=0.7, $p(wrong$=0.3)
  - $EMC(q)$=0.7∗0.8+0.3∗0.2=0.62



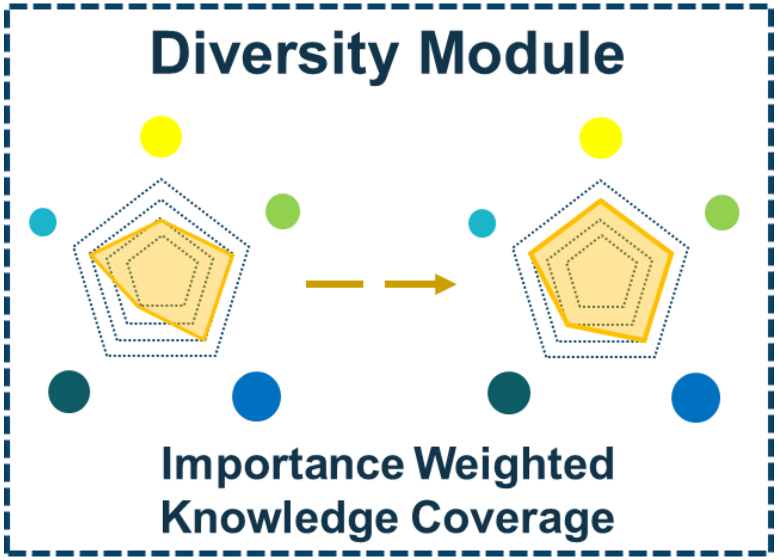
MAAT Framework

- Quantify the diversity of a set of questions
  - Intuitively done with knowledge concepts related to questions
- Importance Weighted Knowledge Coverage (IWKC)
  - Idea: the more knowledge concepts covered, the more diverse the questions are

$$\mathrm{IWKC}(Q_T) = \frac{\sum_{k \in K} w_k * IncCov(k, Q_T)}{\sum_{k \in K} w_k},$$

$$IncCov(k, Q_T) = \frac{cnt(k, Q_T)}{cnt(k, Q_T) + 1},$$

- Challenge: effective and efficient optimization
  - Difficulty: finding the question set with maximum IWKC is proved to be NP-hard
- Solution: fortunately with the submodular property of IWKC, the simple greedy algorithm can achieve a suboptimal solution with optimal ratio $1-1/e$

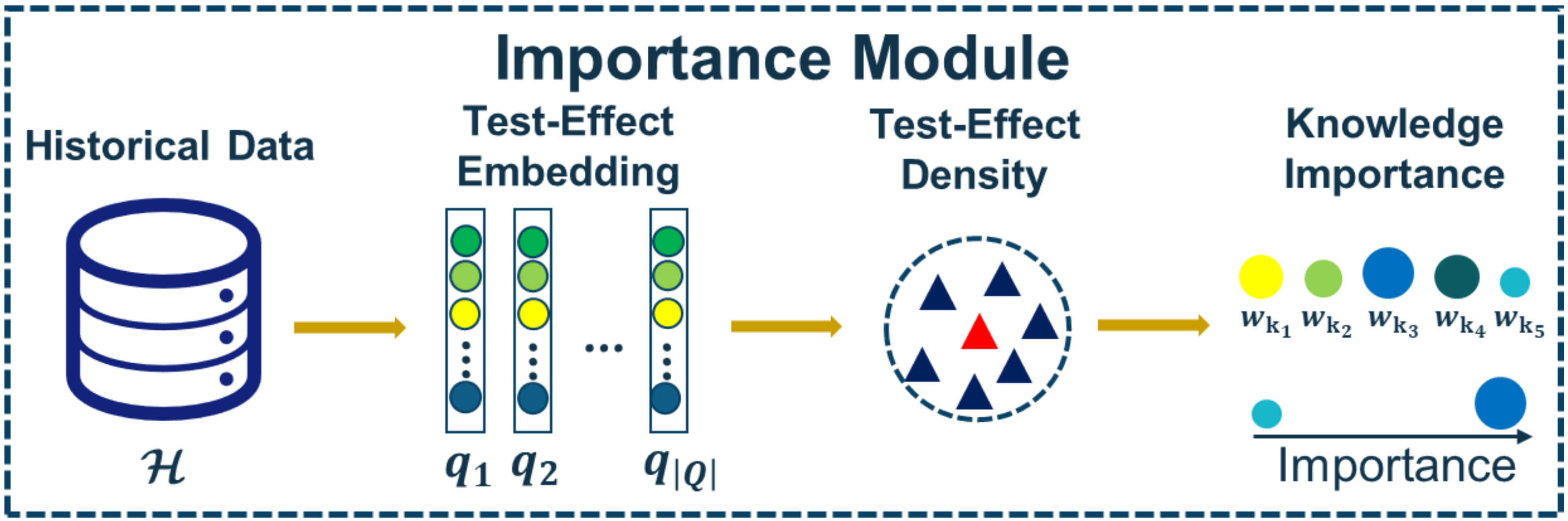


MAAT Framework

- Why distinguish the importance
  - A test may have different emphasis on knowledge
- Importance weight in IWKC
  - A constant for each knowledge concept
  - Pre-computed by the importance module

$$\mathrm{IWKC}(Q_T) = \frac{\sum_{k\in K} \boxed{w_k} * IncCov(k, Q_T)}{\sum_{k\in K} w_k},$$

$$IncCov(k, Q_T) = \frac{cnt(k, Q_T)}{cnt(k, Q_T)+1},$$



Outline

Experiment

- Two real datasets on education
  - EXAM: student examination records
  - ASSIT: student exercise records
- Data preprocess
  - Filter knowledge concepts & questions
- Baseline methods
  - IRT: RAND, MFI, KLI
  - MIRT: RAND, D-Opt, MKLI
  - NCDM: RAND
- Evaluation metrics
  - Quality metric: $Inf(\mathcal{S}) = AUC(\{M(e_i, q_j|\boldsymbol{\theta})|e_i \in E, q_j \in Q\}).$
  - Diversity metric: $Cov(\mathcal{S}) = \frac{1}{|K|}\sum_{k \in K} \mathbb{1}[k \in Q_T].$

TABLE II
STATISTICS OF THE DATASETS

Dataset	EXAM	ASSIST
Num. students	4,307	1,505
Num. questions	527	932
Num. concepts	31	22
Num. records	105,586	59,500
Avg. records per student	24.5	39.5
Avg. records per question	200.4	63.8
Avg. questions per concept	17.0	44.38

Experiment

- Quality comparison
  - AUC on performance prediction
  - Adapt to all the models (model-agnostic)
- Diversity comparison
  - Coverage on knowledge
  - Grow rapidly because of the intrinsic aim

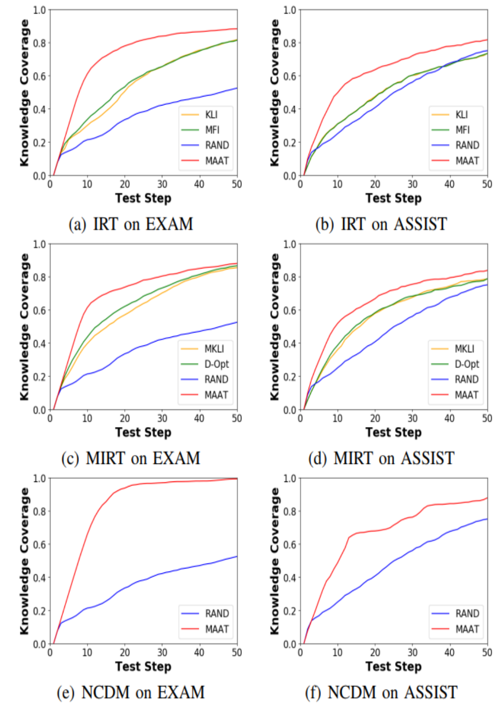


Fig. 3. Diversity Comparison with Coverage Metric

QUALITY COMPARISON WITH AUC METRIC

(a) EXAM

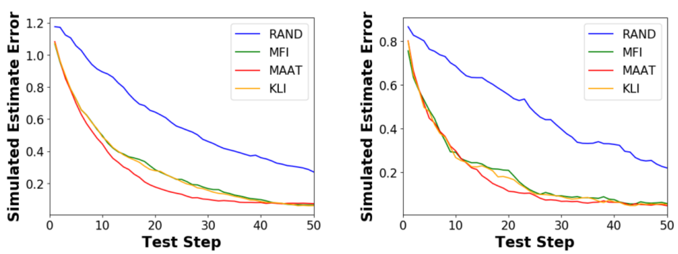
Methods	IRT		MIRT		NCDM	
	@25	@50	@25	@50	@25	@50
RAND	0.6435	0.7076	0.7426	0.7767	0.7081	0.7566
MFI	0.7092	0.7207	-	-	-	-
KLI	0.7081	0.7257	-	-	-	-
D-Opt	-	-	0.7515	0.7710	-	-
MKLI	-	-	0.7502	0.7747	-	-
MAAT	**0.7192**	**0.7319**	**0.7600**	**0.7861**	**0.7614**	**0.7868**

(b) ASSIST

Methods	IRT		MIRT		NCDM	
	@25	@50	@25	@50	@25	@50
RAND	0.6619	0.6664	0.6734	0.6902	0.6832	0.7217
MFI	0.6659	0.6691	-	-	-	-
KLI	0.6658	0.6692	-	-	-	-
D-Opt	-	-	0.6832	0.7004	-	-
MKLI	-	-	0.6781	0.6877	-	-
MAAT	**0.6674**	**0.6703**	**0.6903**	**0.7063**	**0.7084**	**0.7334**

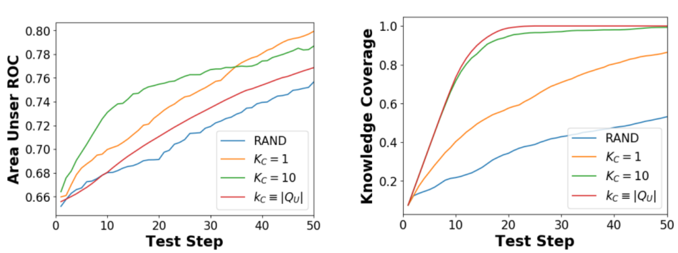
Experiment

- Ablation study
  - How MAAT keeps a balance when quality meets diversity
  - A relatively small K_C keeps the best balance of quality and diversity
- Case study
  - The first 10 steps of a typical examinee in EXAM
  - can select diverse questions while keeping high quality as well



(a) EXAM (b) ASSIST

Fig. 4. Simulated Estimate Error comparison



(a) Quality (b) Diversity

Fig. 5. Change in quality and diversity with different K_C for ablation study

TABLE IV
RESULTS ON A TYPICAL EXAMINEE FOR CASE STUDY

MAAT		D-Opt		MKLI	
Concept	Inf	Concept	Inf	Concept	Inf
Function	0.6666	Function	0.6652	Triangle	0.6645
Set	0.6710	Equation	0.6686	Algebra	0.6689
Equation	0.6763	Equation	0.6717	Equation	0.6732
Triangle	0.6841	Triangle	0.6756	Function	0.6774
Algebra	0.6905	Geometry	0.6801	Algebra	0.6810
Triangle	0.6961	Function	0.6857	Function	0.6843
Coordinates	0.7022	Geometry	0.6914	Function	0.6887
Geometry	0.7087	Triangle	0.6956	Triangle	0.6929
Real Number	0.7136	Algebra	0.6963	Inequality	0.7001
Equation	0.7188	Function	0.6998	Geometry	0.7057

Outline

Conclusion & Future work

- Overall results
  - Propose a novel Model-Agnostic Adaptive Testing framework (MAAT)
  - Address the problem of selecting both high-quality and diverse questions in the adaptive testing procedure
  - Design three sophisticated modules that worked cooperatively and iteratively
- Future work
  - Fewer hand-crafted heuristics
  - As a general framework, each component may be further improved

Thanks for your listening!

bhy0521@mail.ustc.edu.cn