Data-driven Item Selection and Generation in Assessments and Learning

Andrew Lan

UMassAmherst | Manning College of Information & Computer Sciences

New Work: Open-ended Knowledge Tracing

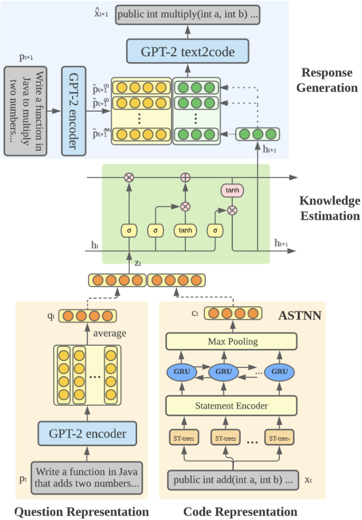
Old

Student responses are **binary-valued**, correct/incorrect. Items are characterized by a few parameters: difficulty, scale

New

Use large language models (LMs) to encode item statements and generate **knowledge-guided, open-ended** response predictions

Proof-of-concept: programming exercises dataset



Liu, Wang, Baraniuk, and Lan, EMNLP 2022

Student Code Prediction

<table>
<tr><th>predicted code</th><th>actual student code</th></tr>
<tr><td>

```
public String zipZap(String str)
{
  for (int i = 0; i < str.length()- 2; i++)
    {
    if (str.charAt(i) == 'z' &&
        str.charAt(i + 2) == 'p')
    {
     str.replace("", str.substring(i + 1));
    }
  }
 return str;
}
```

</td><td>

```
public String zipZap(String str)
{
  for (int i = 0; i < str.length() - 2; i++)
  {
    if (str.charAt(i) == 'z' &&
        str.charAt(i + 2) == 'p')
    {
      str.replace("0", str.substring(i + 1));
      return str;
    }
  }
  return str;
}
```

</td></tr>
<tr><td>

```
public boolean evenlySpaced(int a, int b, int c)
{
    int diffone = b - a;
    int difftwo = c - b;
    if (diffone = difftwo) {
        return true;
    }
    else {
        return false;
    }
}
```

</td><td>

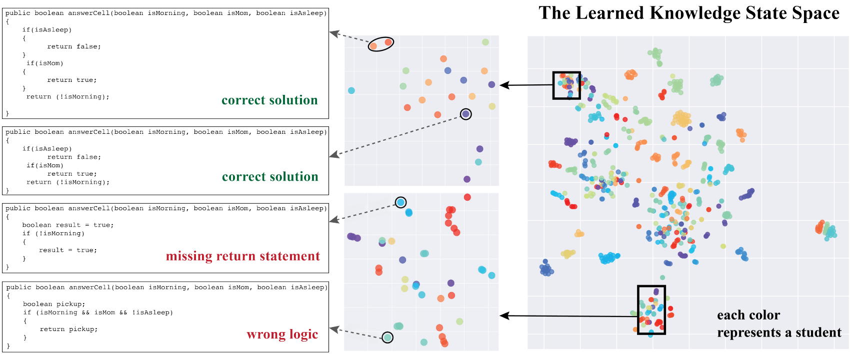
```
public boolean evenlySpaced(int a, int b, int c)
{
    int diffone = b - a;
    int difftwo = c - b;
    boolean question = false;
    if (diffone == difftwo) {
        question = true;
        return question;
    }
    else {
        return question;
    }
}
```

</td></tr>
</table>

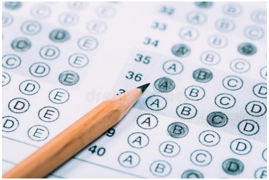
- OKT can make **personalized** predictions of **structure and approach** in actual student code

Interpreting Knowledge States

Write a function in Java that implements the following logic:
Your cell phone rings. Return true if you should answer it. Normally you answer, except in the morning you only answer if it is your mom calling. In all cases, if you are asleep, you do not answer.



Back to the Talk: Assessment and Learning



Assessment and learning can now happen simultaneously

Digital Learning Platforms

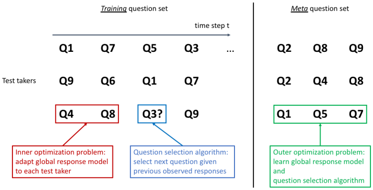
Opportunity

Digital learning platforms (DLPs) produce large-scale learning data and enables personalization

Data-driven Item Selection and Generation

- Item selection
  - **Bilevel optimization** based computerized adaptive testing

Ghosh and Lan, IJCAI 2021



- Item generation
  - **Controlled generation** of math word problems using language models

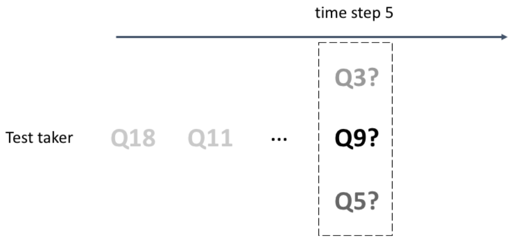
Equation	`x = num1 - num2`
Context	`cousin game playing points scored video`
Gen. MWP	Zach scored num1 points in the football game. Ben scored num2 points. How many more points did Zach score than Ben?

Computerized Adaptive Testing (CAT)

Goal

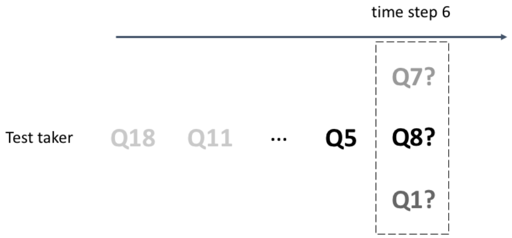
Reduce test length needed to accurately measure learner ability

How CAT Works



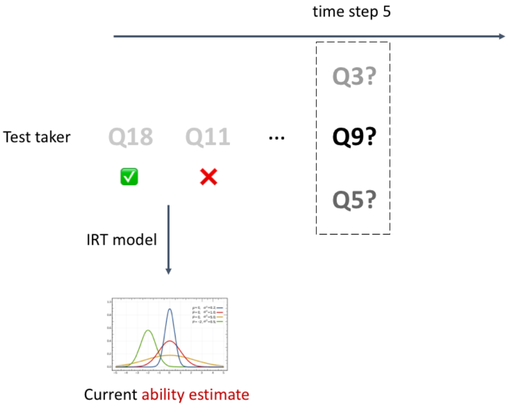
Select **personalized** next question for each test taker
from a **question bank**

How CAT Works



Select **personalized** next question for each test taker from a **question bank**

How CAT Works



- Estimate test taker **ability** using item response theory (IRT)
- Select next item that provides most **information** on ability

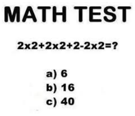
IRT

- Simple yet powerful

$$p(Y_{i,j} = 1) = \sigma(\theta_i - b_j),$$

The **1PL** IRT model

$\theta_u \in \mathbb{R}$:
test taker u's
ability



$b_i \in \mathbb{R}$:
question (item)
i's **difficulty**

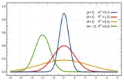
- $Y_{i,j} \in \{0, 1\}$: binary-valued question response correctness
- $\sigma(\cdot) : \mathbb{R} \to [0, 1]$: the sigmoid function

There are many variants; we use 1PL as a running example

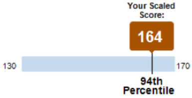
How CAT Works

time step t

Test taker Q18 Q11 ... **Q1**



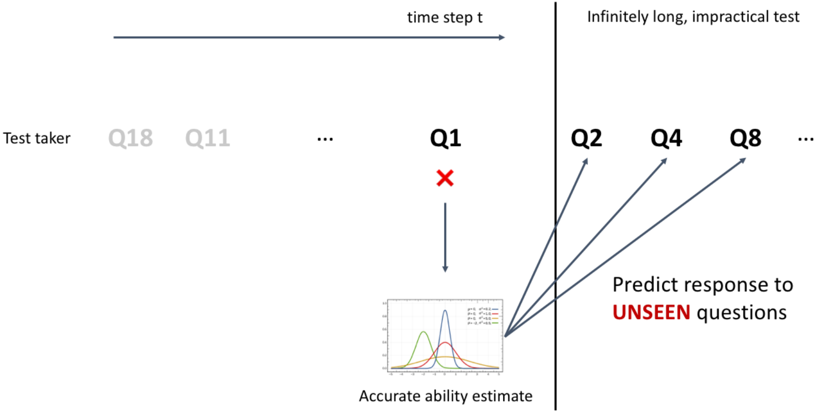
Accurate ability estimate



Final score

- At the end of the test, score according to the **ability estimate**
- Fixed-length and varied-length CAT

Nature of CAT



- If the test contains every question in the question bank, that score would be highly accurate
- But that is not practical
- Ability is a **proxy** for score on that long test

Some problems with CAT

- IRT is **not** the most flexible and predictive model
- The informativeness metric is **static**

Opportunity

Models and item selection algorithms that can truly exploit large-scale learner response data

BOBCAT: Bilevel Optimization-Based CAT

Goal: learn a **data-driven** question selection algorithm and support **flexible response models**

BOBCAT

We solve the following bilevel optimization problem:

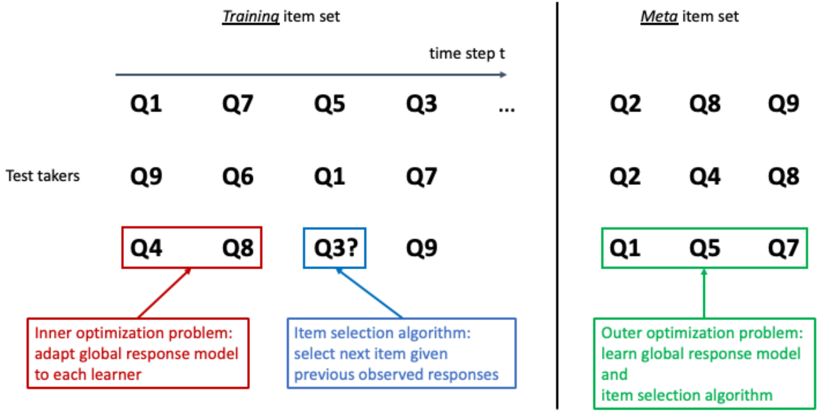
$$\underset{\boldsymbol{\gamma},\boldsymbol{\phi}}{\text{minimize}} \ \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\Gamma_i}\ell\Big(Y_{i,j}, g(j;\boldsymbol{\theta}_i^*)\Big) := \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\boldsymbol{\theta}_i^*, \Gamma_i) \tag{1}$$

$$\text{s.t.} \ \boldsymbol{\theta}_i^* = \arg\min_{\boldsymbol{\theta}_i}\sum_{t=1}^{n}\ell\Big(Y_{i,j_i^{(t)}}, g(j_i^{(t)};\boldsymbol{\theta}_i)\Big) + \mathcal{R}(\boldsymbol{\gamma},\boldsymbol{\theta}_i) := \mathcal{L}'(\boldsymbol{\theta}_i) \tag{2}$$

$$\text{where} \ j_i^{(t)} \sim \Pi(Y_{i,j_i^{(1)}}, \ldots, Y_{i,j_i^{(t-1)}}; \boldsymbol{\phi}) \in \Omega_i^{(t)} \tag{3}$$

- (1): outer optimization problem, learn **global response model** parameters $\boldsymbol{\gamma}$ and **item selection algorithm** parameters $\boldsymbol{\phi}$
- (2): inner optimization problem, learn **local response model** parameters for each test taker $\boldsymbol{\theta}_i^*$
- (3): item selection algorithm

BOBCAT Illustration



Meta-learning setup that splits data into training and meta sets

BOBCAT Details

- **Agnostic** to the response model: IRT/neural network
- **Global-local** parameter split in the response model - flexible, one example:
  - Global: difficulty/weights&biases
  - Local: ability/input vector
- **Gradient** calculation for global variables $\boldsymbol{\gamma}, \boldsymbol{\phi}$
  - Unbiased using REINFORCE
  - Biased using influence function scores (works better)
- Computational efficiency: **more** efficient than CAT
  - The item selection algorithm is a neural network with raw responses as input
  - Only need a forward pass
  - No need to update the ability estimate after each response

Experiments: Data and Metrics

Dataset	EdNet	Junyi	Eedi-1	Eedi-2	ASSISTments
Students	312K	52K	119K	5K	2.3K
Questions	13K	25.8K	27.6K	1K	26.7K
Interactions	76M	13M	15M	1.4M	325K

- Data: five **large-scale**, **real-world** learner response datasets
- Metrics: accuracy and AUC on test set

Experiments: Methods (selected)

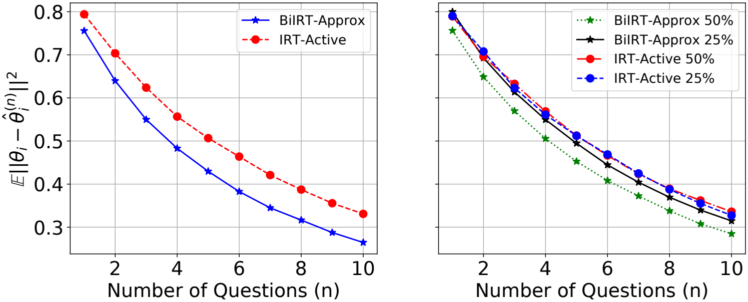
- **IRT-active**: IRT as response model without BOBCAT, uncertainty sampling as item selection algorithm
- **BiIRT-active**: IRT as response model in BOBCAT, uncertainty sampling as item selection algorithm
- **BiNN-active**: neural network as response model in BOBCAT, uncertainty item as question selection algorithm
- **BiNN-unbiased**: neural network as response model in BOBCAT, learned item selection algorithm with unbiased gradient estimate
- **BiNN-approx**: neural network as response model in BOBCAT, learned item selection algorithm with biased, approximate gradient estimate

Experiments: Results

Dataset	n	IRT-Active	BiIRT-Active	BiIRT-Unbiased	BiIRT-Approx	BiNN-Approx
EdNet	1	70.08	70.92	71.12	**71.22**	**71.22**
	3	70.63	71.16	71.3	71.72	**71.82**
	5	71.03	71.37	71.45	71.95	**72.17**
	10	71.62	71.75	71.79	72.33	**72.55**
Junyi	1	74.52	74.93	74.97	**75.11**	75.1
	3	75.19	75.48	75.53	75.76	**75.83**
	5	75.64	75.79	75.75	76.11	**76.19**
	10	76.27	76.28	76.19	76.49	**76.62**
Eedi-1	1	66.92	68.22	68.61	**68.82**	68.78
	3	68.79	69.45	69.81	70.3	**70.45**
	5	70.15	70.28	70.47	70.93	**71.37**
	10	71.72	71.45	71.57	72.0	**72.33**
Eedi-2	1	63.75	64.83	65.22	65.3	**65.65**
	3	65.25	66.42	67.09	67.23	**67.79**
	5	66.41	67.35	67.91	68.23	**68.82**
	10	68.04	68.99	68.84	69.47	**70.04**
ASSISTments	1	66.19	68.69	69.03	**69.17**	68.0
	3	68.75	69.54	69.78	**70.21**	68.73
	5	69.87	69.79	70.3	**70.41**	69.03
	10	71.04	70.66	**71.17**	71.14	69.75

- Data-driven item selection algorithms are better than informativeness-driven ones
- Biased approximate gradient works much better than unbiased gradient
- Neural network works better than IRT on larger datasets

Experiment: Ability Estimation



- Even if we still score test takers using **IRT ability estimates**, the learned item selection algorithm is much more effective than typical informativeness metrics
- It gets better with **more training data**

Test Security Problems

Method	Exposure (median)	Exposure ($>$20%)	Overlap (mean)
IRT-Active	0.51%	0.25%	6.03%
BiNN-Approx	0%	1.54%	28.64%

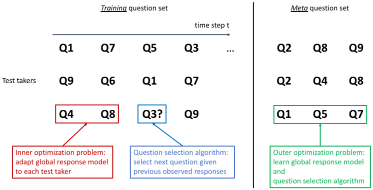
- Learned question selection algorithms lead to higher **question exposure rates** and much higher **test overlap rates**
- Forthcoming work adds corresponding **constraints** into our optimization objective to address this problem

BOBCAT

- BOBCAT improves CAT by **learning** item selection algorithms from data: the more data, the better the algorithm
- Supports any response model

Data-driven Item Selection and Generation

- Item selection
  - **Bilevel optimization** based computerized adaptive testing



- Item generation
  - **Controlled generation** of math word problems using language models

Equation	`x = num1 - num2`
Context	`cousin game playing points scored video`
Gen. MWP	Zach scored num1 points in the football game. Ben scored num2 points. How many more points did Zach score than Ben?

Wang, Baraniuk, and Lan, EMNLP 2021

Personalizing Items

- Even if we know how to select items perfectly, they still come from a **finite item pool**

Educational Measurement
ISSUES AND PRACTICE
NCME national council on measurement in education

Educational Measurement: Issues and Practice
Fall 2020, Vol. 39, No. 3, pp. 100–105

Standardization and *UNDERSTAND*ardization in Educational Assessment

Stephen G. Sireci, *University of Massachusetts Amherst*

Problem

Items may not match learner interests or be **culturally relevant**

- Manually generating personalized items is helpful but not a scalable approach

Math Word Problems (MWPs)

MWP: Joan found 70 seashells on the beach. She gave Sam some of her seashells. She has 27 seashells. How many seashells did she give to Sam?
Equation: `x = (70 - 27)`

- Textual statement, underlying equation, relatively simple
- Can use **language models** (LMs), e.g., GPT-3, PaLM, to automatically generate
- But these black-box models are not controllable

Goal

Controllable MWP generation by specifying **equation** & **context**

Controllable MWP Generation

X = (6 - 2)

Joanna, balloons

Joanna has six balloons. Two popped.
How many does she have left ?

Controllable MWP Generation

`x = ( 6 * 2 ) + 3`

Joanna, balloons

Joanna bought **two sets of six** balloons. She then bought **three more**. How many balloons does she have now?

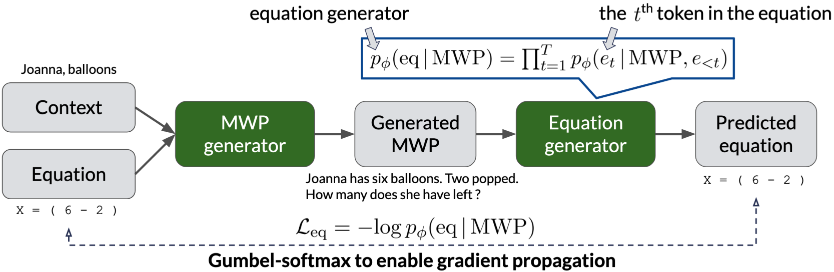
Controllable MWP Generation

$X = (6 * 2) + 3$

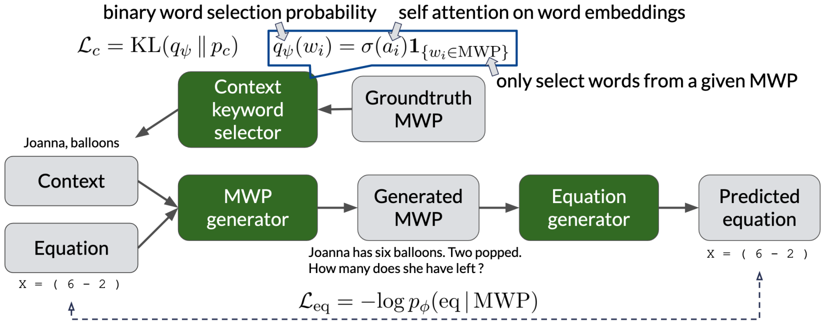
Mike, apple, basket

Mike has six **apples** in each of the two **baskets** and three **apples** on the table. How many **apples** does he have in total?

Key Technique: Equation Consistency Control



Key Technique: Context Keyword Selection



Quantitative Results

	Arithmetic				MAWPS				Math23K			
	BLEU-4	METEOR	ROUGE-L	ACC-eq (0.769)	BLEU-4	METEOR	ROUGE-L	ACC-eq (0.755)	BLEU-4	METEOR	ROUGE-L	ACC-eq (0.672)
seq2seq-rnn	0.075	0.152	0.311	0.413	0.153	0.175	0.362	0.472	0.196	0.234	0.444	0.390
+ GloVe	**0.351**	0.310	0.555	0.399	0.592	0.412	0.705	0.585	0.275	0.277	0.507	0.438
seq2seq-tf	0.339	0.298	0.524	0.405	0.554	0.387	0.663	**0.588**	0.301	0.294	0.524	0.509
GPT	0.237	0.248	0.455	0.401	0.368	0.294	0.538	0.532	0.282	0.297	0.512	0.477
GPT-pre	0.316	**0.322**	0.554	0.403	0.504	0.391	0.664	0.512	0.325	**0.333**	0.548	0.498
ours	0.338	**0.322**	**0.567**	**0.453**	**0.596**	**0.427**	**0.715**	0.557	**0.332**	0.330	**0.549**	**0.513**

- Three MWP datasets
- Baselines: sequence-to-sequence, fine-tuning GPT
- Metrics: BLEU-4, METEOR, ROUGE-L (language quality), ACC-eq (equation consistency)
- Baselines tend to not really "learn" how to generate MWPs

	Arithmetic		MAWPS		Math23K	
	BLEU-4	ACC-eq	BLEU-4	ACC-eq	BLEU-4	ACC-eq
$\mathcal{L}_{eq}$ (softmax)	0.110	0.417	0.308	**0.555**	0.284	0.466
$\mathcal{L}_{eq}$ (Gumbel-softmax)	**0.303**	**0.455**	**0.522**	0.527	**0.306**	**0.495**
keyword, TF-IDF	0.313	**0.424**	0.518	0.536	0.310	0.498
keyword, noun+pronoun	0.316	0.413	0.504	0.512	**0.325**	0.498
context selection	**0.320**	0.412	**0.533**	**0.542**	0.324	**0.501**
full model w/o $\mathcal{L}_c$	0.303	**0.455**	0.522	0.527	0.306	0.495
full model w/o $\mathcal{L}_{eq}$	0.320	0.412	0.491	0.500	0.324	0.501
full model w/o both	0.316	0.403	0.504	0.512	0.325	0.498
full model	**0.338**	0.453	**0.596**	**0.557**	**0.332**	**0.513**

Qualitative Results: Fixed Context, Varying Equation

Context: `candies`

Equation #1: `x = num1 + num2`

seq2seq-tf: ethan has num1 presents . alissa has num2 more than ethan . how many presents does alissa have ? (in training data)

GPT-pre: There are num1 scissors in the drawer. Keith placed num2 scissors in the drawer. How many scissors are now there in total? (irrelevant to context)

ours: Mildred collects num1 candies. Mildred's father gives Mildred num2 more. How many candies does Mildred have? (✓)

Equation #2: `x = num1 - num2`

seq2seq-tf: mildred weighs num1 pounds . carol weighs num2 pounds . how much heavier is mildred than carol ? (in training data)

GPT-pre: Joan has num1 blue balloons but lost num2 of them. How many blue balloons does Joan have now? (irrelevant to context)

ours: There are num1 candies in the jar. num2 are eaten by a hippopotamus. How many candies are in the jar? (✓)

Equation #3: `x = num1 * num2`

seq2seq-tf: each banana costs $ num1 . how much do num2 bananas cost ? (in training data)

GPT-pre: Joan has saved num1 quarters from washing cars. How many cents does Joan have? (inconsistent with equation)

ours: Each child has num1 candies. If there are num2 children, how many candies are there in all? (✓)

Equation #4: `x = num1 / num2`

seq2seq-tf: there are num1 bananas in diane ' s banana collection . if the bananas are organized into num2 groups , how big is each group ? (in training data)

GPT-pre: Joan has num1 blue marbles. Sandy has num2 times more blue marbles than Melanie. How many blue marbles does Joan have? (inconsistent with equation)

ours: There are num1 candies in the candy collection. If the candies are organized into num2 groups, how big is each group? (✓)

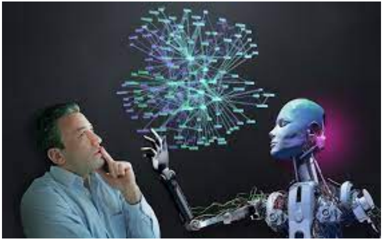
Qualitative Results: Fixed Equation, Varying Context

Equation: `x = num1 + num2 + num3`	
Context #1: `violin piano acoustic guitar`	**Context #2:** `beets eggplant`
seq2seq-tf: sara grew num1 onions , sally grew num2 onions , and fred grew num3 onions . how many onions did they grow in all ? (in training data)	**seq2seq-tf**: sara grew num1 onions , sally grew num2 onions , and fred grew num3 onions . how many onions did they grow in all ? (in training data)
GPT-pre: There are num1 dogwood trees currently in the park. Park workers will plant num2 dogwood trees today and num3 dogwood trees tomorrow. How many dogwood trees will the park have when the workers are finished? (irrelevant to context)	**GPT-pre**: There are num1 orchid bushes currently in the park. Park workers will plant num2 orchid bushes today and num3 orchid bushes tomorrow. How many orchid bushes will the park have when the workers are finished? (irrelevant to context)
ours: Mike joined his school's band. He bought a clarinet for $ num1, a music stand for $ num2, and a song book for $ num3. How much did Mike spend at the music store? (✓)	**ours**: Sara grew num1 beets, Sally grew num2 beets, and Fred grew num3 beets. How many beets did they grow in total? (✓)

MWP Generation

- We provide a method for the context/equation-controllable MWP generation method based on LMs
- LMs have the ability to adapt to many topics due to their intrinsic knowledge
- However, LMs are not very good at **mathematical reasoning**
- The **safety** of the generated text also needs to be monitored

The Future



- Contrary to popular belief, big data in education is still in its early stages
- Need to develop ways for human and artificial intelligence to **work together**