Generalizing Expert Misconception Diagnoses Through Common Wrong Answer Embedding

John Kolb, Scott Farrar, (*presenter*) Zachary A. Pardos

UC Berkeley

Computational Approaches to Human Learning (CAHL) research lab

Graduate School of Education

UC Berkeley School of Information

Berkeley EECS
Electrical Engineering and Computer Sciences

Our Goal

- Automatically generate the student misconceptions underlying a commonly occurring wrong answer

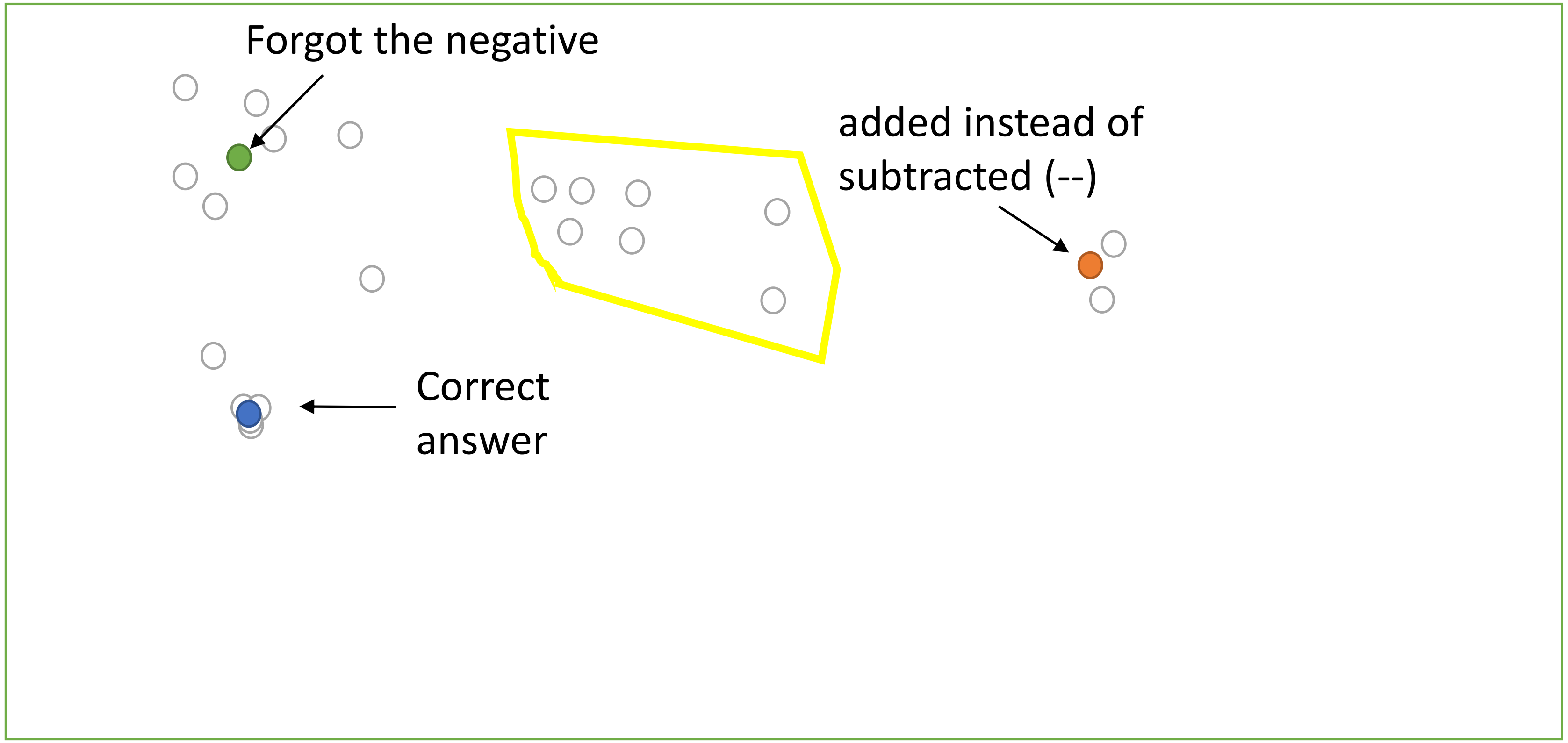
Our Goal

- Automatically generate the student misconceptions underlying a commonly occurring wrong answer

- **Misconception**: An incorrect understanding of a topic leading to false conclusions
    - Rich treatment in literature
    - Can be difficult to work backwards from answer to its origins in the student's misunderstanding

Potential Impact on teaching

Similar to a "Map"



Teacher dashboard depicting which misconception their students are close to

Observations:

(1) There are a lot of student problem solving data **(2)** There are not a lot of semantic diagnostic data

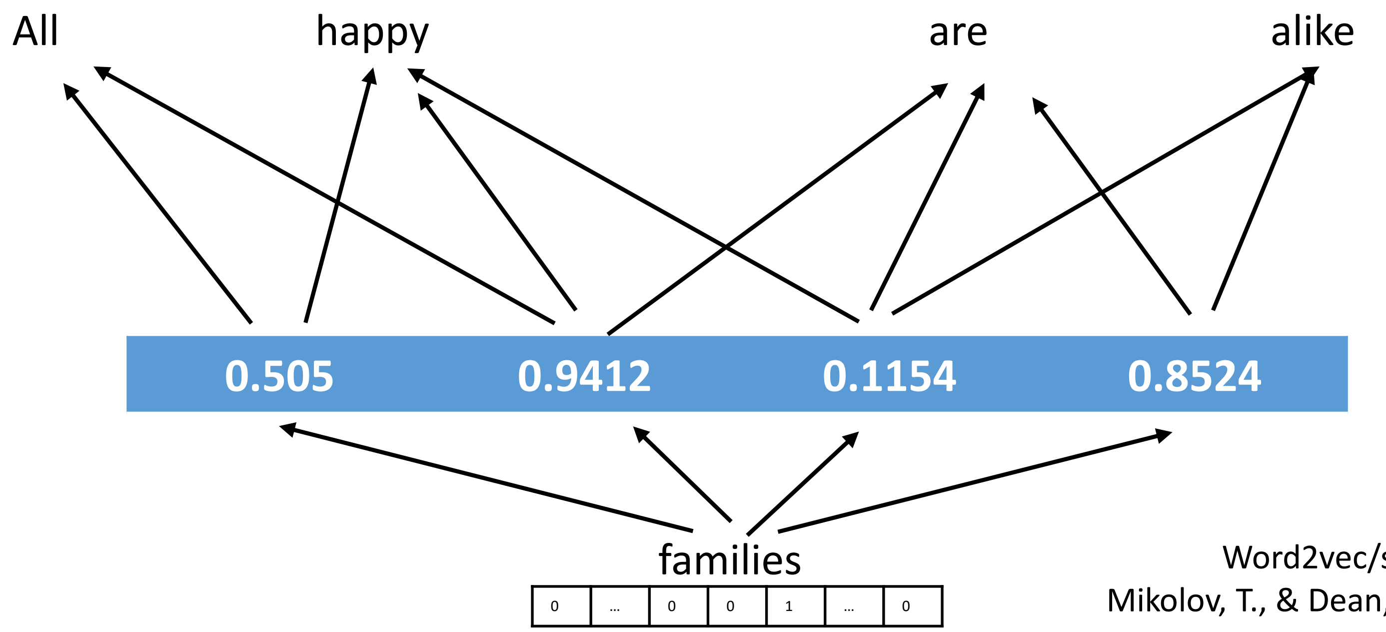
An Embedding is like a Map

$\underbrace{\text{“All happy families are alike;}}_{}$ context

“All happy families are alike; each unhappy family is unhappy in its own way.”

(e.g., Google News archive)

This training process, using SGD, is run on a corpus of 1b words
to learn vector representations of each word in the vocabulary



Word2vec/skip-gram
Mikolov, T., & Dean, J. (2013)

Related work

1. Problem2vec used to predict skill of item (Dadu & Pardos, 2017)
2. Course2vec used to find university course similarity (Pardos, Jiang, & Fan, 2019)
3. Initial pilot study of misconceptions using w2v (Pardos et al., 2018)
4. Tree or graph-based representations of student problem-solving “maps” (Eagle & Barnes, 2014;Muehling, 2017)
5. Inferring misconceptions from open-text (Michalenko, Lan, & Baraniuk, 2017)

Data Set (Khan Academy)

- Each student answer log contained:
  - *Exercise*: Overall topic/skill
  - *Problem Type*: Generic question template
  - *Seed*: Unique template instantiation
  - *Timestamp*

Data Set (Khan Academy)

- Each student answer log contained:
  - ***Exercise*: Overall topic/skill**
  - *Problem Type*: Generic question template
  - *Seed*: Unique template instantiation
  - *Timestamp*

selected for likelihood of misconceptions:

"Surface areas“

"Slope from an equation in slope-intercept form“

"Area of quadrilaterals and polygons“

"Adding and subtracting fractions"

Data Set (Khan Academy)

	Surface Areas	**Slope-Intercept**	**Area of Quads.**	**Add/Sub Fractions**
Problem Types	6			
Seeds	38			
Unique Users	105,659			
Unique Incorrect Answers	55,126			
Total Incorrect Answers	619,045			

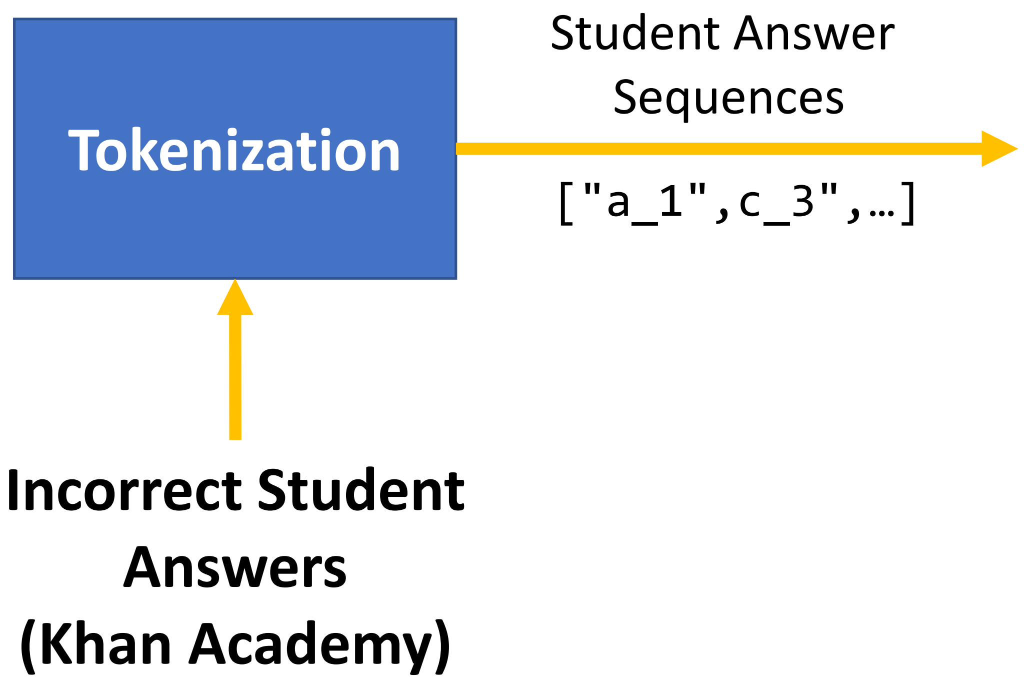
Data Set (Khan Academy)

	Surface Areas	Slope-Intercept	Area of Quads.	Add/Sub Fractions
Problem Types	6	2	2	7
Seeds	38	20	50	40
Unique Users	105,659	33,603	58,239	179,263
Unique Incorrect Answers	55,126	6,912	17,998	46,516
Total Incorrect Answers	619,045	112,390	298,356	873,916

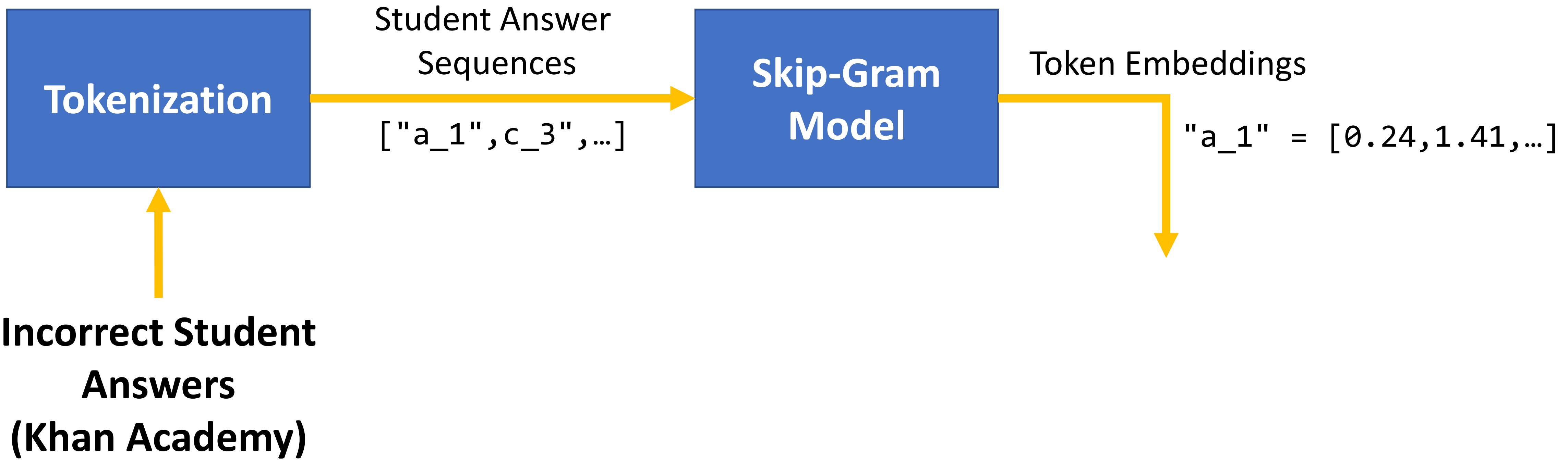
Approach Summary

Incorrect Student Answers (Khan Academy)

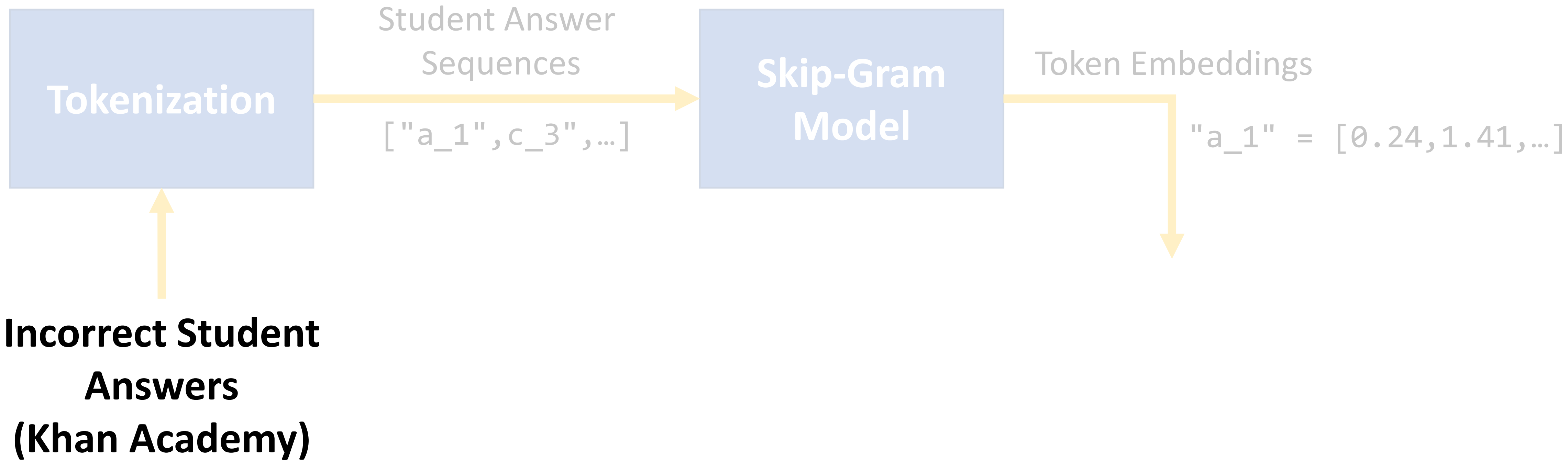
Approach Summary



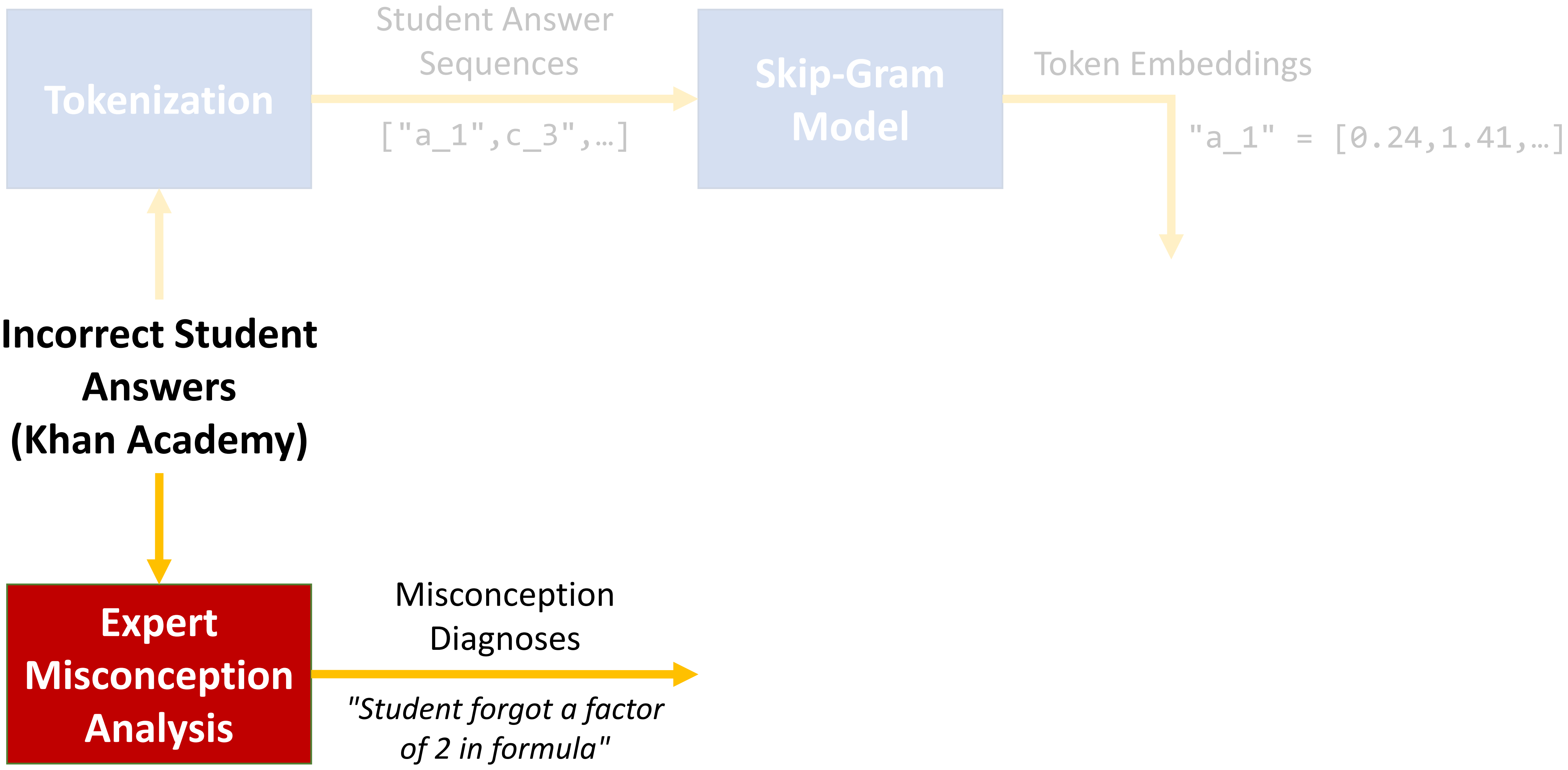
Approach Summary



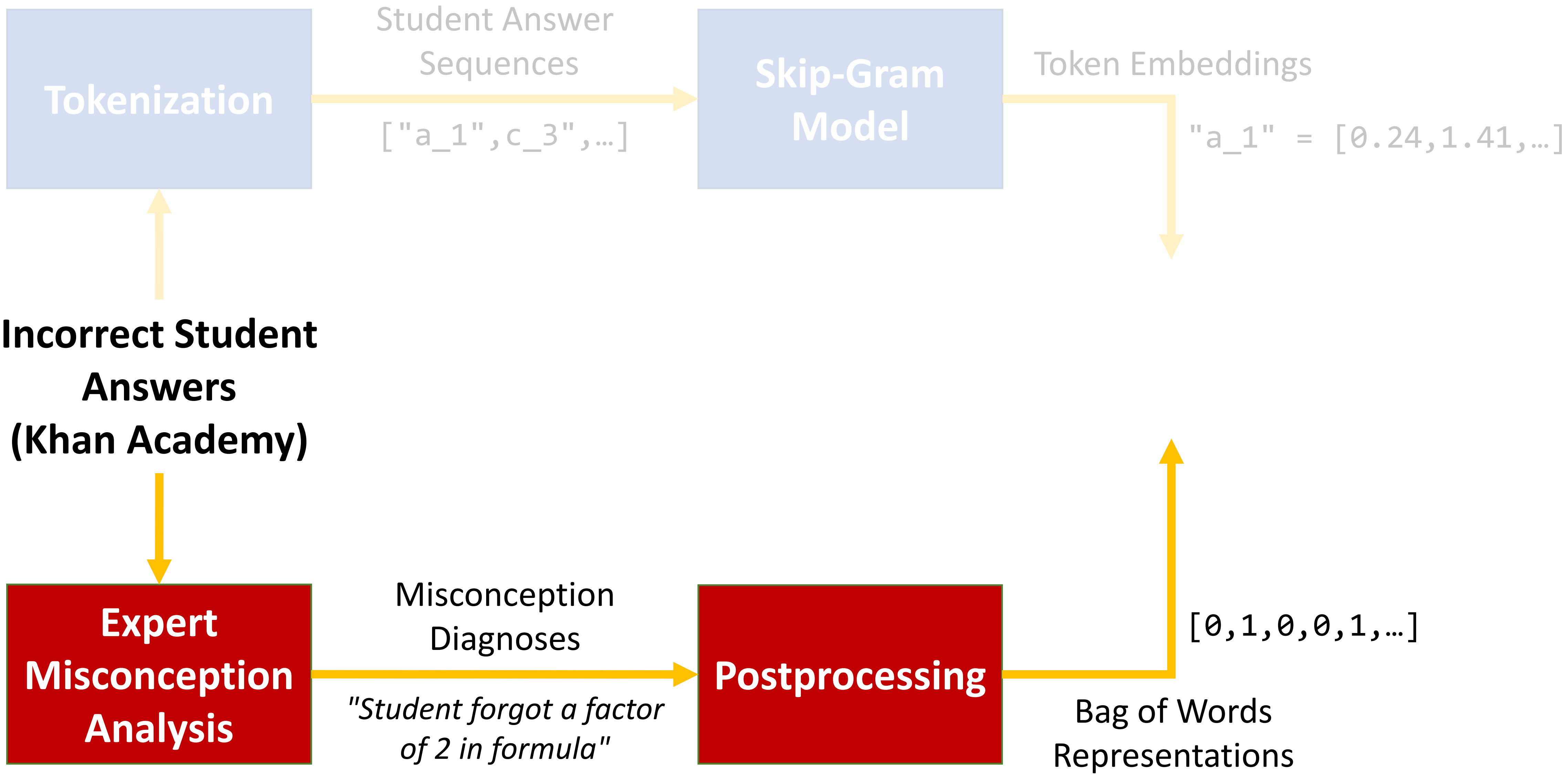
Approach Summary



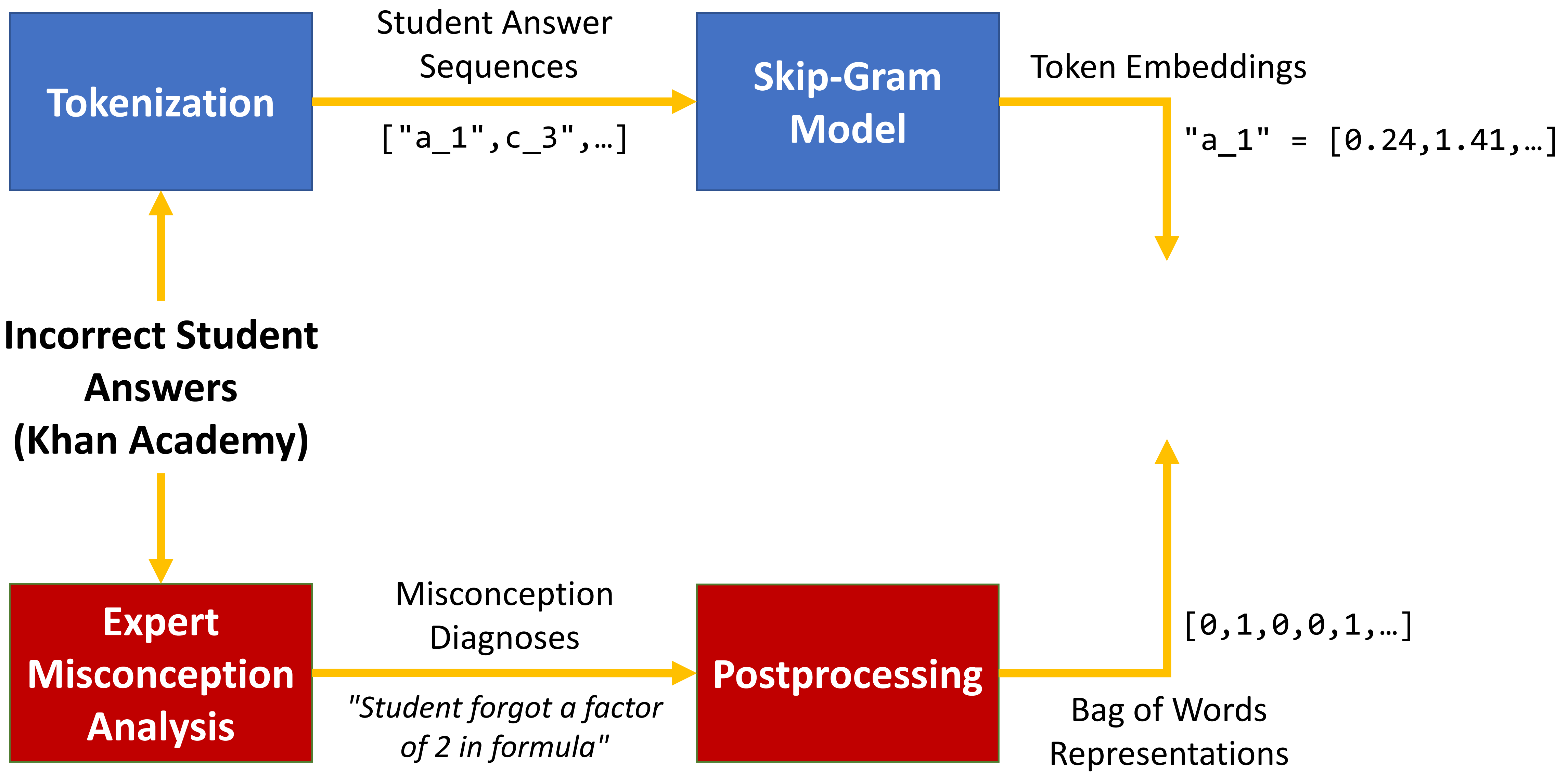
Approach Summary



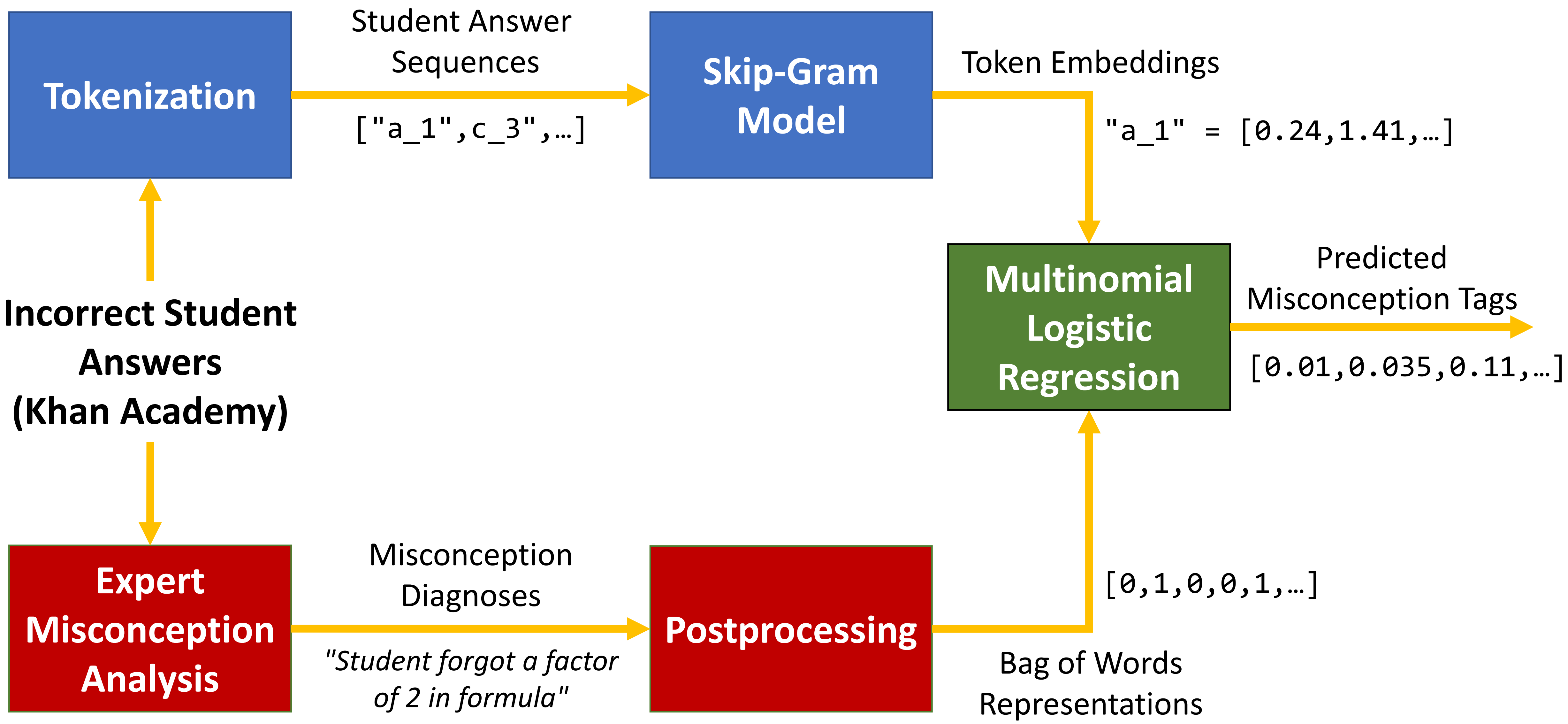
Approach Summary



Approach Summary



Approach Summary

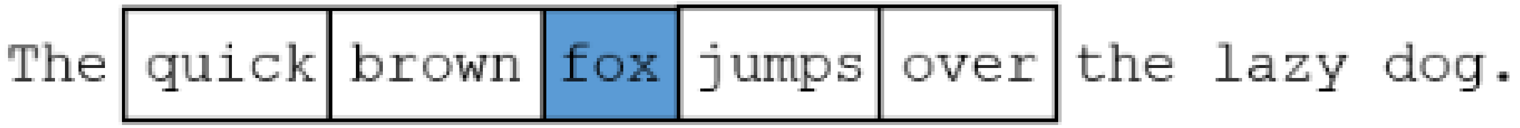


Computing Embeddings – Skip-Gram Model

- Why skip-gram?
  - Surfaces insights about words based on the *context* in which they appear (surrounding words in a sentence)

Computing Embeddings – Skip-Gram Model

- Why skip-gram?
  - Surfaces insights about words based on the *context* in which they appear (surrounding words in a sentence)



graphic http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Computing Embeddings – Skip-Gram Model

- Why skip-gram?
  - Surfaces insights about words based on the *context* in which they appear (surrounding words in a sentence)

    The | quick | brown | fox | jumps | over | the lazy dog.

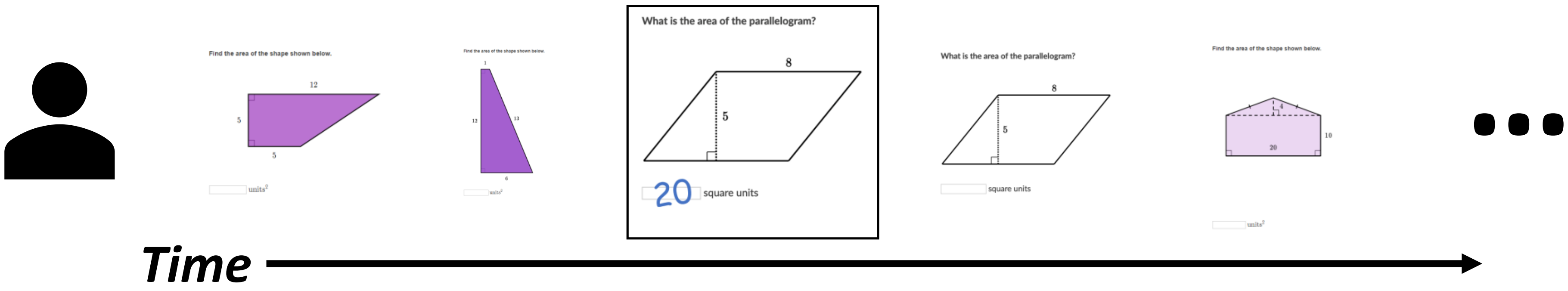
  - We seek insights about incorrect answers based on a similar context (chronologically preceding and following answers)

Computing Embeddings – Skip-Gram Model

- Why skip-gram?
  - Surfaces insights about words based on the *context* in which they appear (surrounding words in a sentence)

    The quick brown fox jumps over the lazy dog.

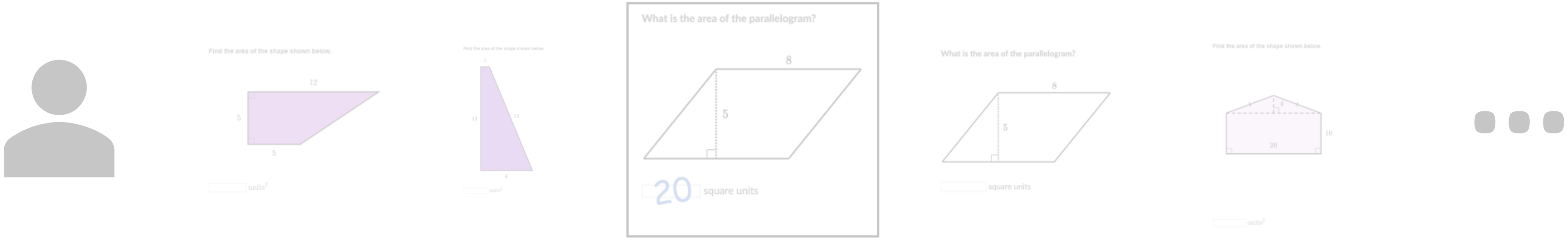
  - We seek insights about incorrect answers based on a similar context (chronologically preceding and following answers)



Computing Embeddings – Skip-Gram Model

- Why skip-gram?
    - Surfaces insights about words based on the *context* in which they appear (surrounding words in a sentence)
    - We seek insights about incorrect answers based on a similar context (chronologically preceding and following answers)

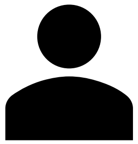
The quick brown fox jumps over the lazy dog.

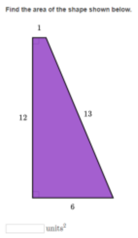


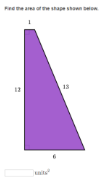
Goal: Learn a semantic space of answer vector embeddings

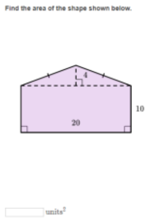
Computing Embeddings – Skip-Gram Model

Time







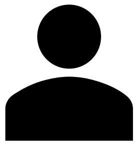


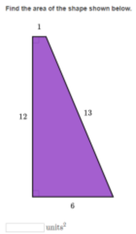
...

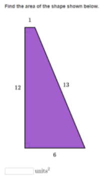
Student response sequence

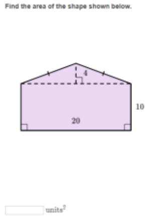
Computing Embeddings – Skip-Gram Model

Time









...

Student response sequence

Seed	x01b
Response	“55"

Seed	x01b
Response	“70"

Seed	x03e
Response	“40"

...

Computing Embeddings – Skip-Gram Model

Time

Student response sequence

Seed	x01b
Response	“55"

Seed	x01b
Response	“70"

Seed	x03e
Response	“40"

…

```
[ "x01b_2", "x01b_6", "x0e3_c", … ]
```

Skip-Gram Token Sequence

Computing Embeddings – Skip-Gram Model

Time

   …

Student response sequence

Seed	x01b
Response	“55"

Seed	x01b
Response	“70"

Seed	x03e
Response	“40"

…

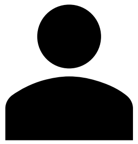
```
[ "x01b_2", "x01b_6", "x0e3_c", … ]
```

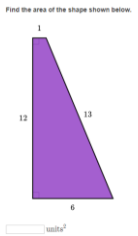
Skip-Gram Token Sequence

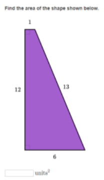
Question's Unique Seed | *frequency rank of incorrect answer* | *“correct”*

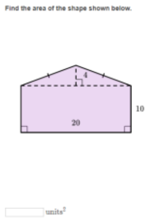
Computing Embeddings – Skip-Gram Model

Time →









...

Student response sequence

Seed	x01b
Response	“55"

Seed	x01b
Response	“70"

Seed	x03e
Response	“40"

...

["x01b_2", "x01b_6", "x0e3_c", …]

Skip-Gram Token Sequence

Collecting Expert Misconception Diagnoses

- Designed and ran a survey on the Qualtrics platform
  - Participants recruited and compensated on our behalf

- Subject population:
  - Mathematics educators teaching grades 5-12 or undergraduates
  - Minimum 2 years prior experience

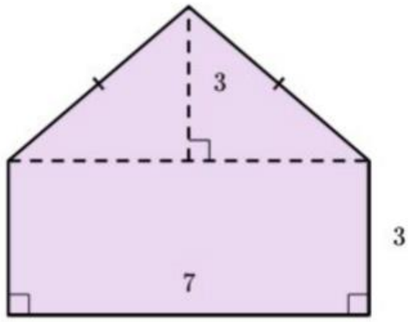
Collection Design

Instructions

- *"Respond with a general label-phrase that describes the most likely error or misconception related to the incorrect answer."*
- *"You may duplicate labels and phrases as you see appropriate."*
- *"You will see three math questions and five incorrect student answers to label for each question."*

Please explain the misconception behind the given student answers for the problem below.

Find the area of the shape shown below.



units^2

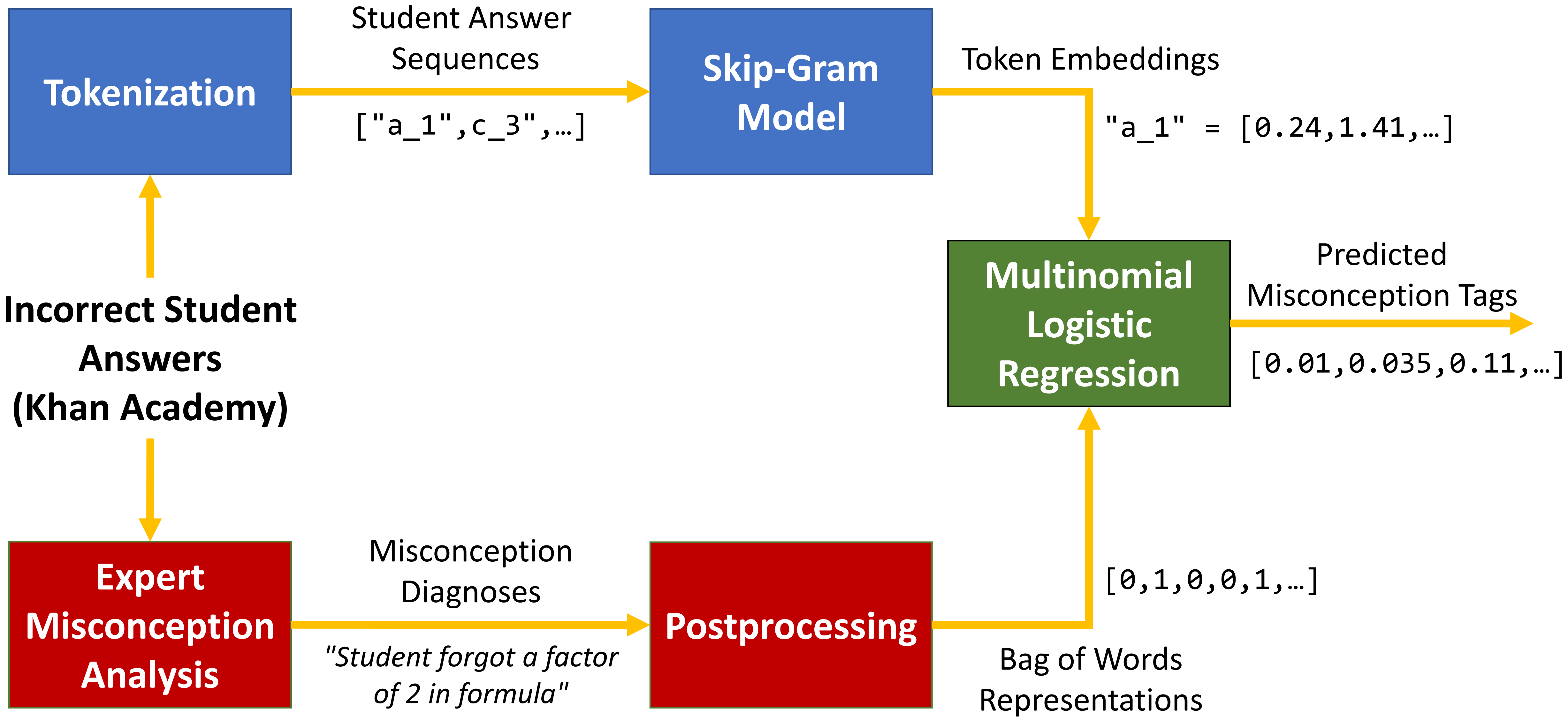
Student Response: **63**

Label:

Postprocessing of Diagnoses

- Filtered for highest quality diagnoses
  - Included work only from experts with average label character count > 20
- NLP cleaning: remove punctuation, stemming, stopword removal
  - Excluded trivially predictable words: `student`, `tried`, `used`, etc.
- Convert diagnosis to vector representation: Bag of Words
- Ended with 19 experts providing 570 unique diagnoses covering 14 of the 15 problem types and 64 of the 89 seeds

Approach Summary



Predicting/Interpolating diagnoses

- Training input: Student answer embedding (*n*-vectors) paired with an expert diagnosis (*m*-vector, bag of words) of that answer
  - *n* = Dimensionality of embedding space, skip-gram hyperparameter
  - *m* = Size of expert diagnosis vocabulary

- Prediction Task: Given an answer embedding, generate its misconception diagnosis
  - *m*-vector: probability distribution over diagnosis vocabulary

Evaluation: Leave-One-Out CV

	Evaluator	**Problem Type**	**Seed**
Folds	19	14	64
Avg. Training Data Points	302	296	314
Avg. Test Data Points	17	24	5

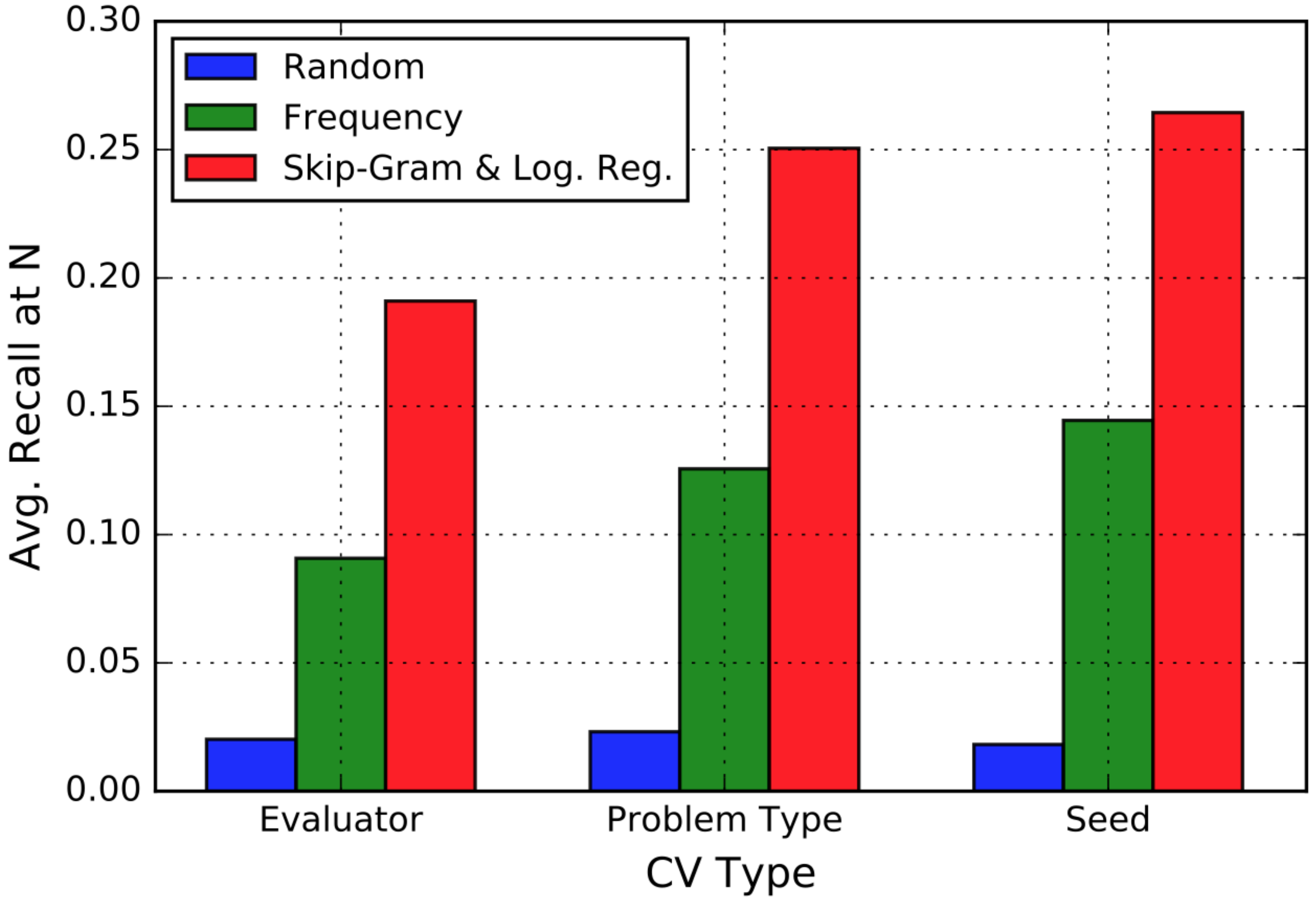
Metric: Recall @ *N*

- Given:
  - An original expert diagnosis consisting of *N* terms
  - Predicted probability distribution over the diagnosis vocabulary
- Select *N* largest elements from probability distribution and their corresponding vocab terms
- How many of these selected terms appear in the original diagnosis?
- Formally: $R = \frac{|\hat{T}_N \cap T|}{|T|}$

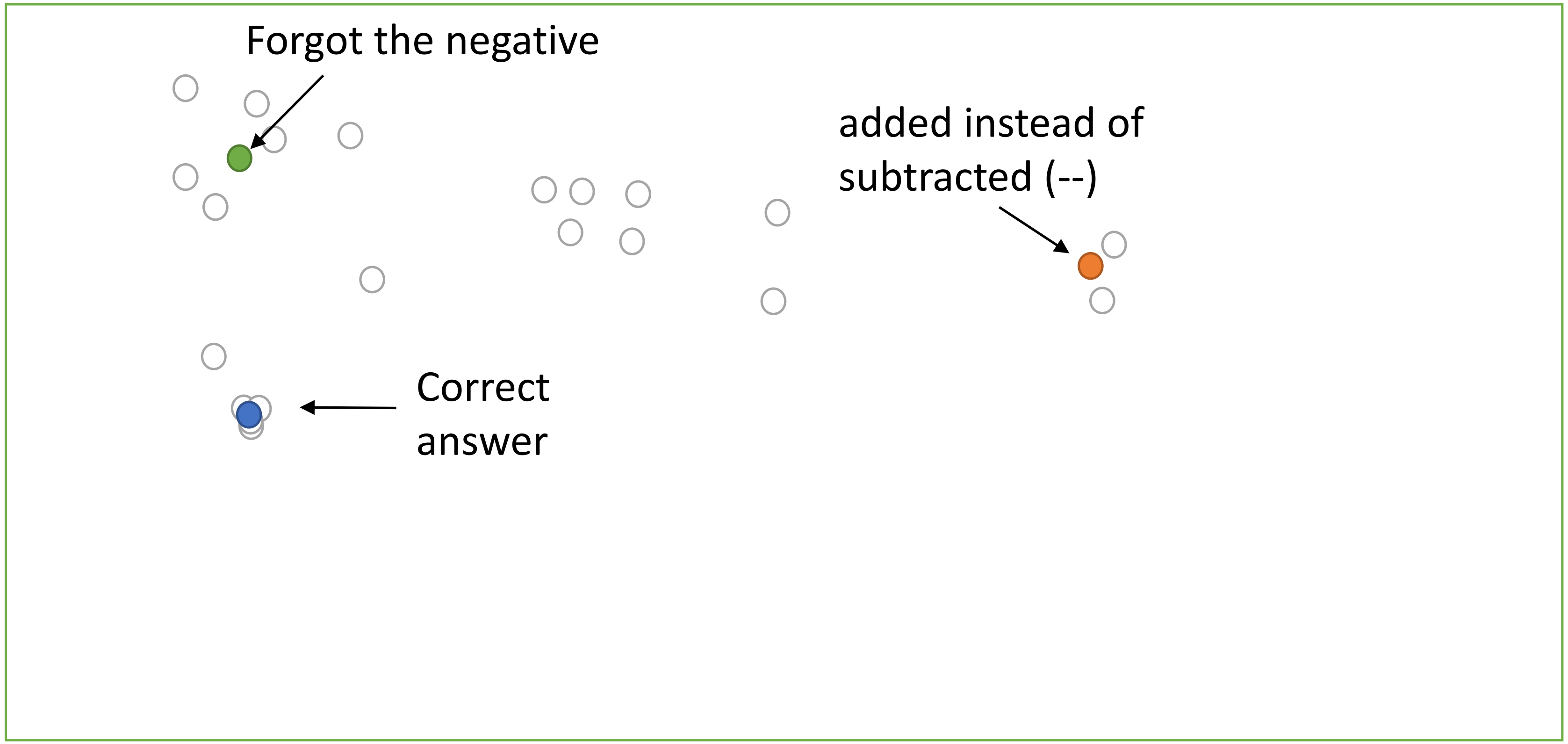
Naïve Baselines for Comparison

1. **Random:** Compute a random sample of *N* terms from the vocabulary
2. **Frequency:** Predict the *N* terms that appear most frequently in the diagnoses from the training set

Results



- Our approach outperforms frequency baseline by ~2x
- Roughly 18%-27% of original terms recovered
- Accuracy rivals that of traditional word-to-word machine translation (Mikolov et al., 2013), which produced 25% accuracy translating between English and Vietnamese.



Limitations

- Open-text diagnoses had high variation
  - > should limit taggers to a shared taxonomy
- W2v embedding was not cross-exercise
  - > use a datasets with high exercise coverage per student
- Single word misconception “hint” may not be enough

Thank You!

Computational Approaches to
Human Learning (CAHL) research lab