

# The Role of Natural Language Processing in Advancing Competency-Based Education and Mathematics Learning in Fourth Graders

19/04/23 (RELELA)  
Felipe Urrutia Vargas



JYVÄSKYLÄN YLIOPISTO  
UNIVERSITY OF JYVÄSKYLÄ



UNIVERSIDAD DE CHILE  
INSTITUTO DE ESTUDIOS  
AVANZADOS EN EDUCACIÓN



# Abstract

La presente charla tiene como objetivo explorar cómo el procesamiento del lenguaje natural (NLP) puede ser utilizada para comprender la educación en la enseñanza primaria.

En la actualidad, las computadoras se han convertido en una herramienta esencial para el cálculo y la simulación, por lo que resulta importante desarrollar enfoques pedagógicos que mejoren la capacidad de comprensión y razonamiento de los estudiantes.

Para lograr este objetivo, se ha propuesto la utilización de técnicas de inteligencia artificial adecuadas para comprender cómo y dónde centrarse en la enseñanza.

En esta charla, presentaré gran parte de mi investigación sobre el impacto de la minería de datos, el aprendizaje automático y las técnicas de NLP en la investigación educativa, tanto pura como aplicada.

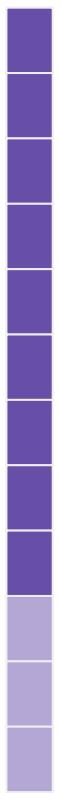
Ilustrare las principales tareas de NLP utilizadas en la minería de datos educativos y describiré los desafíos y los conocimientos adquiridos al proponer nuevos métodos y tareas de NLP en educación. Además, expondré la utilidad de NLP para descifrar las perspectivas actuales de la educación basada en competencias y demostrar el potencial de las técnicas de NLP para entender los desafíos de la argumentación en matemáticas en estudiantes de cuarto grado en Chile y su relación con la estimación del rendimiento en pruebas estandarizadas.

Explicare la parte metodológica y detallare los datos y técnicas de análisis utilizados. Presentare los resultados obtenidos de manera clara y discutiré las implicaciones prácticas de los hallazgos.

Finalmente, destacare la importancia del uso de NLP en educación para mejorar las competencias de los estudiantes y sugeriré posibles vías para futuras investigaciones.

Los asistentes a la charla tendrán la oportunidad de comprender el nuevo paradigma en NLP y conocer cómo conectar NLP con grafos del conocimiento, cuáles son los beneficios que nos aporta para construir IA aplicables en contextos educativos y cómo podría ayudar a desarrollar una nueva técnica para comprender conceptos en educación.

# Timeline



45 min

15 min  
preguntas



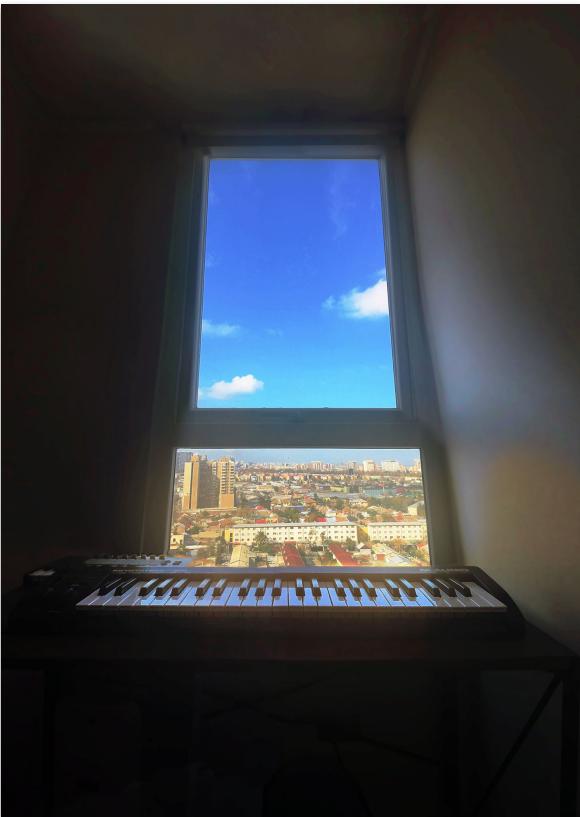
Presentación personal

Minería de Datos  
Educacionales (EDM)

Trabajos de  
Investigación con  
NLP en Educación

Trabajo futuro

# Presentación personal



# Presentación personal

Licenciado en Ciencias de la Ingenieria, mención Matemáticas

## Estudiante



*Alumno regular*  
Ingeniería  
matemática,  
FCFM



*Alumno regular*  
Magister en Ciencias,  
mención Computacion,  
FCFM

## Investigador asistente



Enero 2022 - Diciembre 2023

Araya, Roberto.

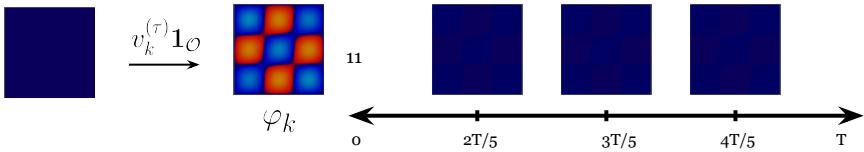


Abril 2022 - Abril\* 2023

Mannonen, Joonas; Hämäläinen, Raija;  
Lehesvuori, Sami.

# Source Reconstruction for Heat Equation

Noviembre 2022-hoy  
Urrutia, Felipe & Axel Osse



**Figure 1.**  
Illustration of our methodology for source reconstruction from known global measurements.  
**(Solid lines)** Sequence of steps of our method and  
**(Dashed lines)** sequence of steps of the [1] method.

**Theorem 1.** Exist  $\beta$  a real number different to  $\gamma\lambda_k$ ,  $V^{(\tau)}$  the Volterra operator (2), such that a control  $v_k^{(\tau)}$  of the form

$$v_k^{(\tau)}(x, t) = c_k(\tau; \beta)\psi_k(x)V^{(\tau)}e^{\beta t} \quad (11)$$

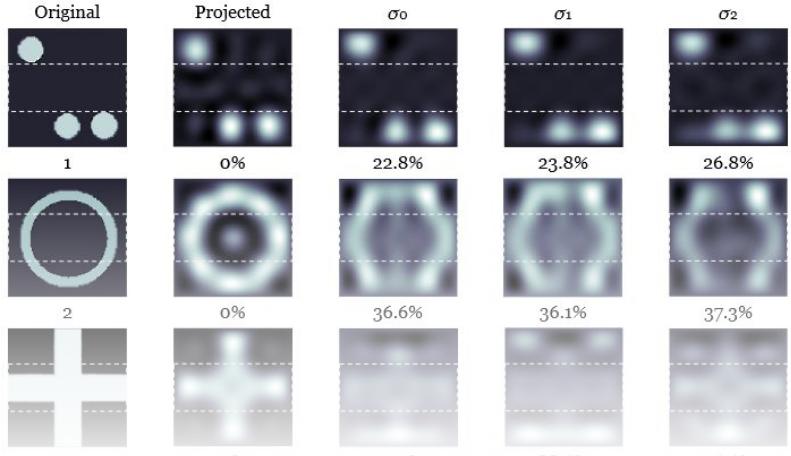
satisfies that the function  $\phi$  is a solution of the problem (9) and satisfies the initial temperature constraint (10). Moreover,  $V^{(\tau)}e^{\beta t}$  is known (Lemma 1) and the  $c_k$  term is given by

$$c_k(\tau; \beta) = \frac{\gamma\lambda_k - \beta}{\beta\sigma(0)(e^{\beta\tau} - e^{\gamma\lambda_k\tau}) + \int_0^\tau (\sigma(s) + \beta\sigma'(s))(e^{\beta\tau+\gamma\lambda_k s} - e^{\beta s+\gamma\lambda_k s})ds}. \quad (12)$$

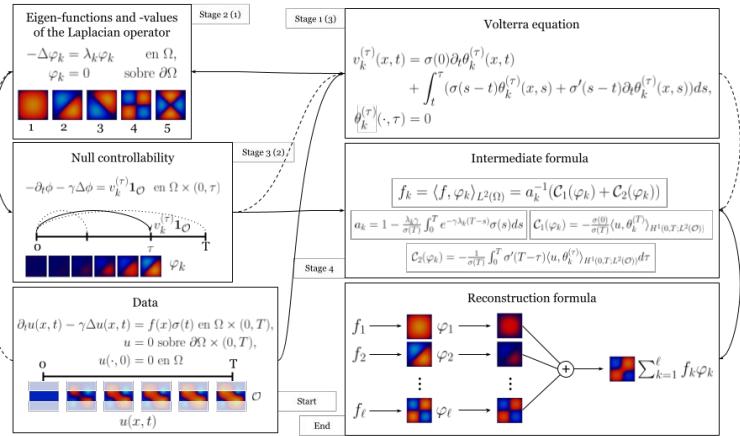
**CMM**

Center for Mathematical Modeling

$$\theta_k^{(\tau)}(x, t) = c_k(\tau; \beta_\tau)\psi_k(x)e^{\beta_\tau t}.$$



**Figure 3.** The given figure depicts the process of reconstructing various sources, represented by the function  $f(x)$ , where  $x$  belongs to a two-dimensional region denoted by  $\Omega = (0, 1)^2$ . The reconstruction is done based on the local measurements taken from the observatory denoted by  $O = (0, 1) \times (0.3, 0.7)$ , which is bounded by dotted lines. The measurements are affected by centered-Gaussian noise (with 0.5 standard deviation). The accuracy of the reconstructions is evaluated by computing the  $L^2$  relative error relative to the first 45 eigenfunctions of the projected source. Three different cases are considered, as represented in the third, fourth, and fifth columns of the figure, where  $\sigma$  takes the values  $\sigma_1$ ,  $\sigma_2$ , and  $\sigma_3$ , respectively (see Figure 2).



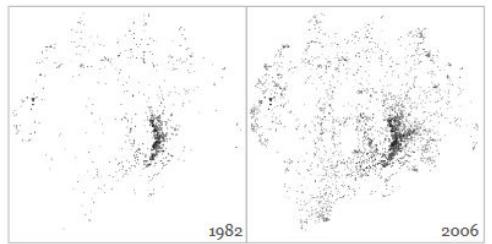
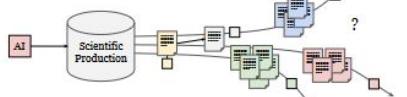
**Figure 4.** Sample of the first eight eigen-functions of the Laplacian operator together with their associated pseudo-functions. (Left) Eigen-functions. (Right) Pseudo-functions.

# Ciencia de la ciencia en IA

Enero 2023-hoy  
Urrutia, Felipe & Andres Abeliuk

## Motivation

The impact of artificial intelligence in various fields has generated great interest in investigating the topics and concepts addressed by the main currents of research in this field. Understanding the similarities and differences in the scientific production in artificial intelligence in Chile and the rest of the world would allow identifying areas of opportunity for new research and projects. Previous works [1] and [2] provide tools and methodologies to analyze the disciplinary organization of scientific publications and predict research trends. Studying these questions could contribute to the advancement of Artificial Intelligence (AI) and consolidate Chile's position in this field.

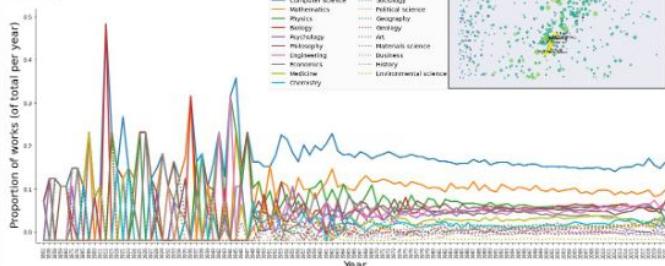
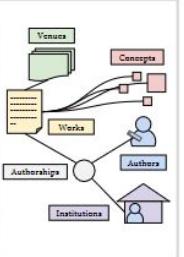


## Introduction

The OpenAlex [3] dataset is a vast collection of scholarly entities and their relationships, including works, authors, sources, institutions, concepts, and publishers.

We use Node2Vec [4] to extract meaningful information from such a large dataset, mapping the graph of concepts to a high-dimensional vector space, allowing for efficient computation and analysis of the relationships between concepts.

By leveraging the power of Node2Vec, the OpenAlex dataset offers a wealth of information and insights into the scholarly world.

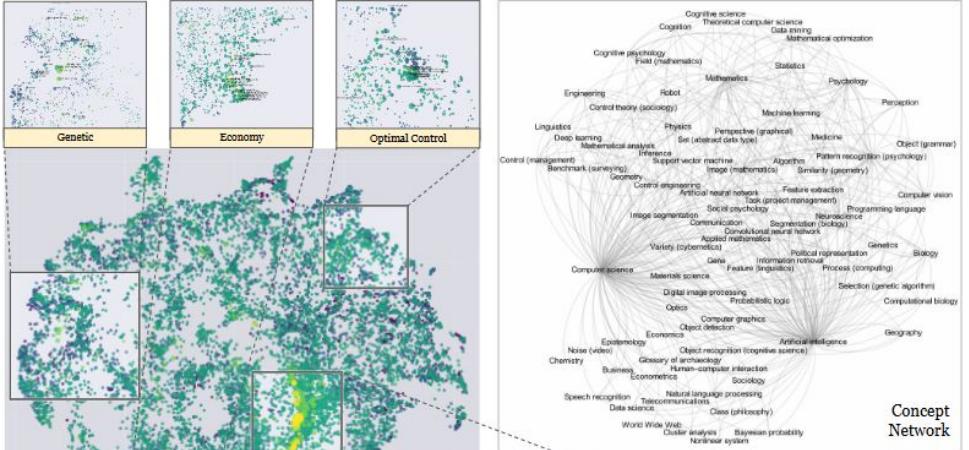


# Science of science in Artificial Intelligence

Urrutia, Felipe<sup>1</sup> & Andres Abeliuk<sup>2</sup>



Concept Network



## Conclusion

### Results

- ◆ The last 10 years, ~28% of AI research has focused on Computer Science, 10% on Mathematics, and 20% on Biology, Psychology, Philosophy, Physics, and Engineering.
- ◆ We have only partially answered the first research question, and we have not yet begun to answer the second research question.

### Future work

- Exploring the disciplinary organization of AI research, particularly in Chile, and identifying areas of opportunity for new research and projects.
- Comparing the research output of Chilean researchers to that of their international counterparts could provide valuable insights into the similarities and differences between the scientific production of AI in Chile and the rest of the world.

## References

- [1] Peng, H., Ko, Q., Balakrishnan, C., Kemere, D. M., & Ahn, Y. Y. (2021). Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances*, 7(17), eabk004.
- [2] Krenn, M., & Zellinger, A. (2020). Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences of the United States of America*, 117(30), 18332–18337.
- [3] Peters, J., Pfeiffer, H., & Orlitzky, R. (2020). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2005.01839*.
- [4] Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 825–834).

# Minería de Datos Educacionales (EDM)

# Minería de Datos Educativos



1. Elementos
2. Grupos de actores
3. Tareas

**Educational data mining: a review of the state of the art**

2010  
Romero, Cristóbal & Sebastián Ventura.  
*IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618.

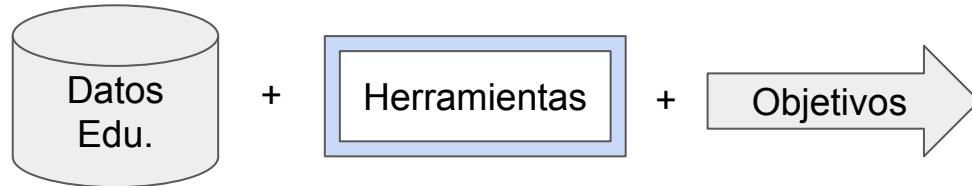
# Minería de Datos Educativos



- Respuestas en ejercicios
- Grabaciones de clases
- Redes sociales
- Libros de contenido para clases
- Contabilidad de recursos
- Mallas curriculares
- Variables de comportamiento
- Planificación de clases

Romero, C.; Ventura, S. (2010). *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618.

# Minería de Datos Educacionales



## Algoritmos estadísticos

- Análisis de correlaciones

## Aprendizaje de máquinas

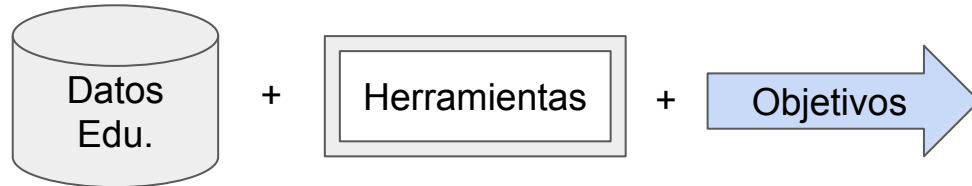
- Regresión
- Clasificación
- Clustering

## Minería de datos

- Reglas de asociación

Romero, C.; Ventura, S. (2010). *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618.

# Minería de Datos Educacionales



## Investigación **pura**

- *Entendimientos* de fenómenos educativos
- *Descubrimiento* de nuevos fenómenos educativos

## Investigación **aplicada**

- Mejorar procesos de aprendizaje
- Guiar el aprendizaje del estudiante
- Apoyar a los profesores con herramientas

# Minería de Datos Educacionales



## Grupos de actores

- Estudiantes
- Educadores
- Investigadores educativos
- Directores

## Tareas

- Recomendaciones para estudiantes
- Modelado de estudiantes
- Desarrollo de mapas conceptuales
- Predecir el desempeño del estudiante

Romero, C.; Ventura, S. (2010). *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618.

# Trabajos de Investigación con NLP en Educación

**Araya, R.; Ulloa, O.; Jimenez, A.; Mannonen, J.; Lehesvuori, S.; Hämäläinen, R.**

# Research

**Do Written Responses to Open-Ended Questions on Fourth-Grade Online Formative Assessments in Mathematics Help Predict Scores on End-of-Year Standardized Tests?**

2022

Urrutia, F., & Araya, R.  
*Journal of Intelligence*,  
10(4), 82.

**Automatically Detecting Incoherent Written Math Answers of Fourth-graders**

2023  
Urrutia, F., & Araya, R.  
(pre-print) *Educational Studies in Mathematics*

**Who's the Best Detective? LLMs vs. MLs in Detecting Incoherent Fourth Grade Math Answers**

2023  
Urrutia, F., & Araya, R.  
(pre-print) *Journal of Educational Computing Research*

**A Methodology for Enhanced Explanations of Incoherence Detection in Fourth Grade Student Writing**

2023  
Urrutia, F., & Araya, R.  
*MKR 2023 (submitted)*

**Untangling Incoherent Argumentation in Fourth Graders' Written Answers to Open-ended Math Questions**

2027?  
Urrutia, F., & Araya, R.

**Mapping the main streams and foci of competence-based education research: A review with direct citation network analysis and topic modeling with latent semantic analysis**

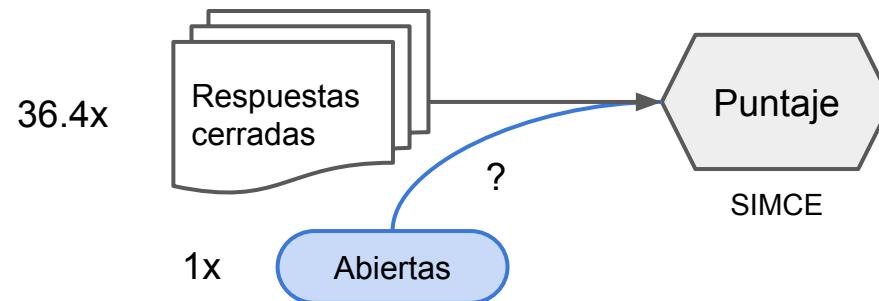
2023  
Mannonen, J., Urrutia, F., Lehesvuori, S., Hämäläinen, R., & Araya, R.  
(proofreading)  
*Educational Research Review*

Urrutia, F.; Araya, R. J. Intell. 2022, 10, 82.

Do Written Responses to Open-Ended Questions on Fourth-Grade Online Formative Assessments in Mathematics Help Predict Scores on End-of-Year Standardized Tests?

2022  
Urrutia, F., & Araya, R.  
*Journal of Intelligence*,  
10(4), 82.

**RQ:** ¿En qué medida las respuestas cortas y escritas de los estudiantes a las preguntas abiertas diseñadas por el profesor en las pruebas formativas semanales en línea ayudan a mejorar las predicciones del rendimiento en las evaluaciones nacionales estandarizadas de opción múltiple de final de año?



**Do Written Responses to Open-Ended Questions on Fourth-Grade Online Formative Assessments in Mathematics Help Predict Scores on End-of-Year Standardized Tests?**

2022

Urrutia, F., & Araya, R.  
*Journal of Intelligence*,  
10(4), 82.

*Respuestas a preguntas abiertas*

1

María y su marido cocinaron ayer una tortilla, la dividieron en 6 partes iguales. María se comió  $\frac{2}{6}$  y su marido  $\frac{3}{6}$ .  $\frac{3}{6}$ . ¿Qué fracción de la tortilla quedó?

1/6

2

Catalina compró 12 cebollas. De las 12 cebollas, utilizó  $\frac{1}{4}$  de ellas para hacer unas deliciosas empanadas. ¿Cuántas cebollas utilizó para las empanadas? Explica cómo supiste el resultado

Necesito 3 y lo sé porque he dividido  $12:4=3x1=3$

4

Pablo tarda 5 horas en viajar de Santiago a La Serena. Su amigo Pedro viajó de La Serena a Santiago y tardó 300 minutos. ¿Cuál de los dos niños tardó menos? Explica tu respuesta

ambos tomaron el mismo tiempo porque he multiplicado  $5 \times 60 = 300$  y 300 minutos son 5 horas

3

Camilo tiene que recoger 60 bolas. Hasta ahora ha recogido 23. Para saber cuántas bolas le quedan por recoger, restan 23 de 60. ¿Es correcto el ejercicio de Camilo? Justifica tu respuesta

está bien porque he añadido  $37+23$  y medio 60

5

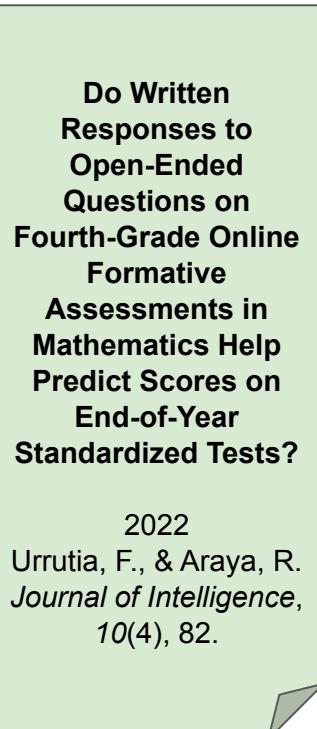
¿Qué es una línea de simetría? Explica con tus propias palabras y dame un ejemplo

una línea de simetría es una línea que separa dos imágenes iguales edades

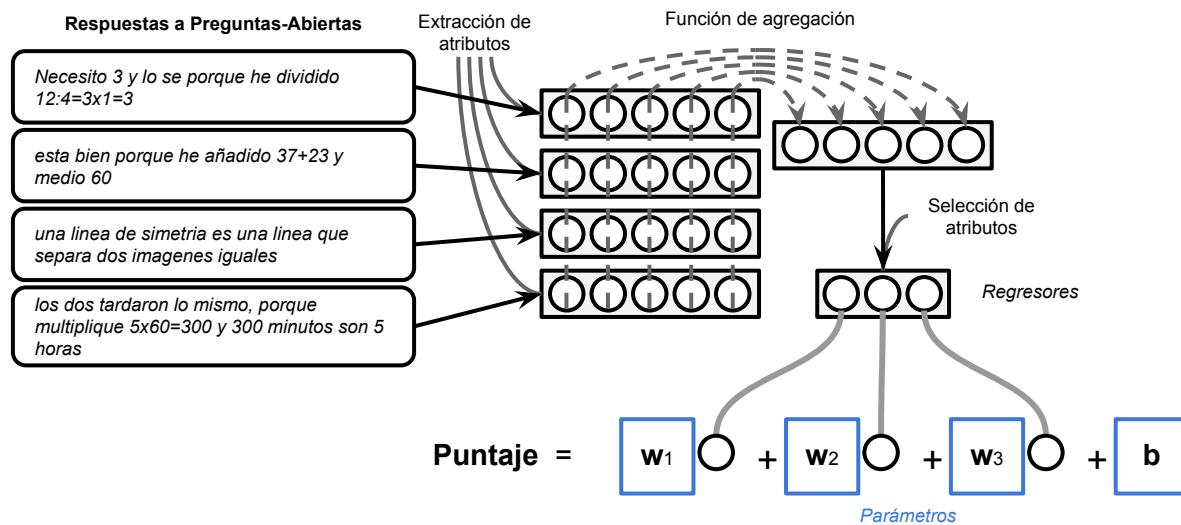
0

Pamela tiene 25 flores y su amiga le regala 17 flores. Escribe con palabras el número total de flores que tiene Pamela

cuarenta y dos



## Modelo Open-ended



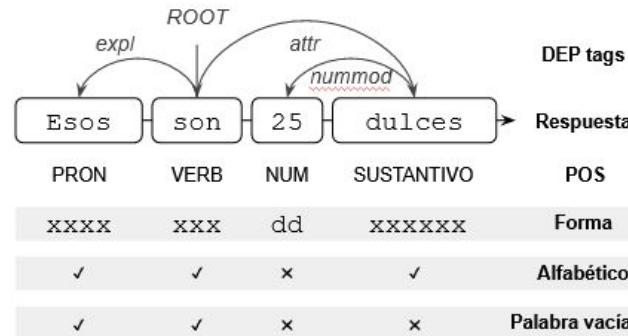
Basado en Fig. 1. Urrutia, F.; Araya, R. J. Intell. 2022, 10, 82.

Do Written Responses to Open-Ended Questions on Fourth-Grade Online Formative Assessments in Mathematics Help Predict Scores on End-of-Year Standardized Tests?

2022

Urrutia, F., & Araya, R.  
*Journal of Intelligence*,  
10(4), 82.

## Variables basadas en respuesta abiertas



Pregunta de tipo **Calcular con explicar**

Laura está en el almacén y compra 268 pesos de pan. Si cancela con 500 pesos, ¿cuánto le dieron de vuelto? Explica con tus palabras cómo llegaste al resultado

no, yo creo que fue Pedro :v

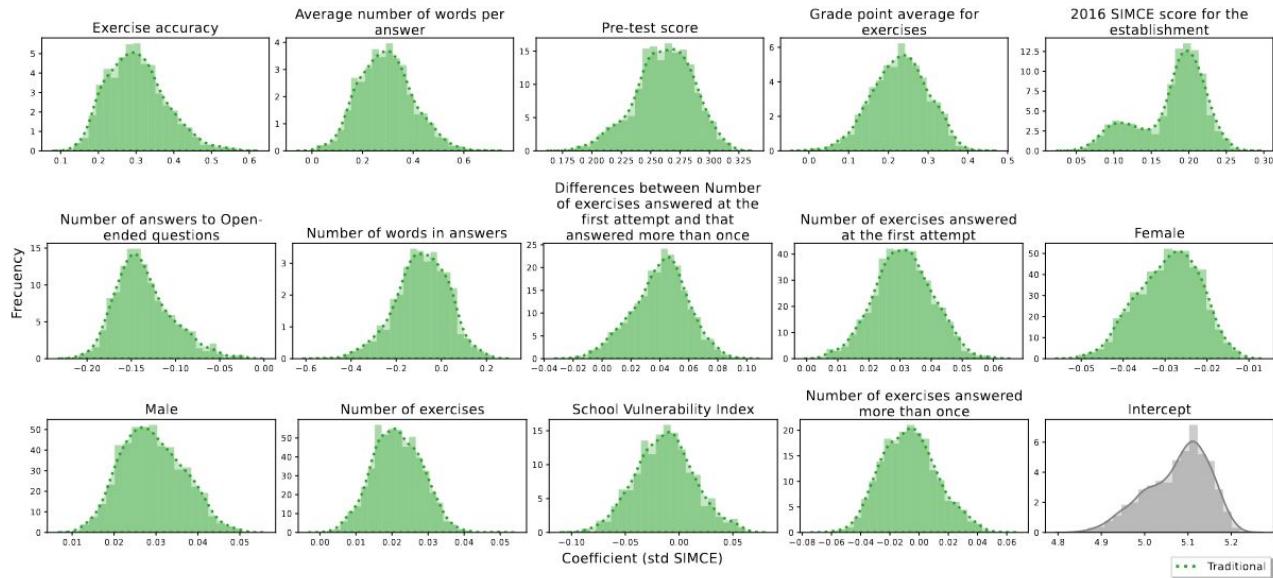
Respuesta es **Incoherente**

Urrutia, F.; Araya, R. J. Intell. 2022, 10, 82.

**Do Written Responses to Open-Ended Questions on Fourth-Grade Online Formative Assessments in Mathematics Help Predict Scores on End-of-Year Standardized Tests?**

2022

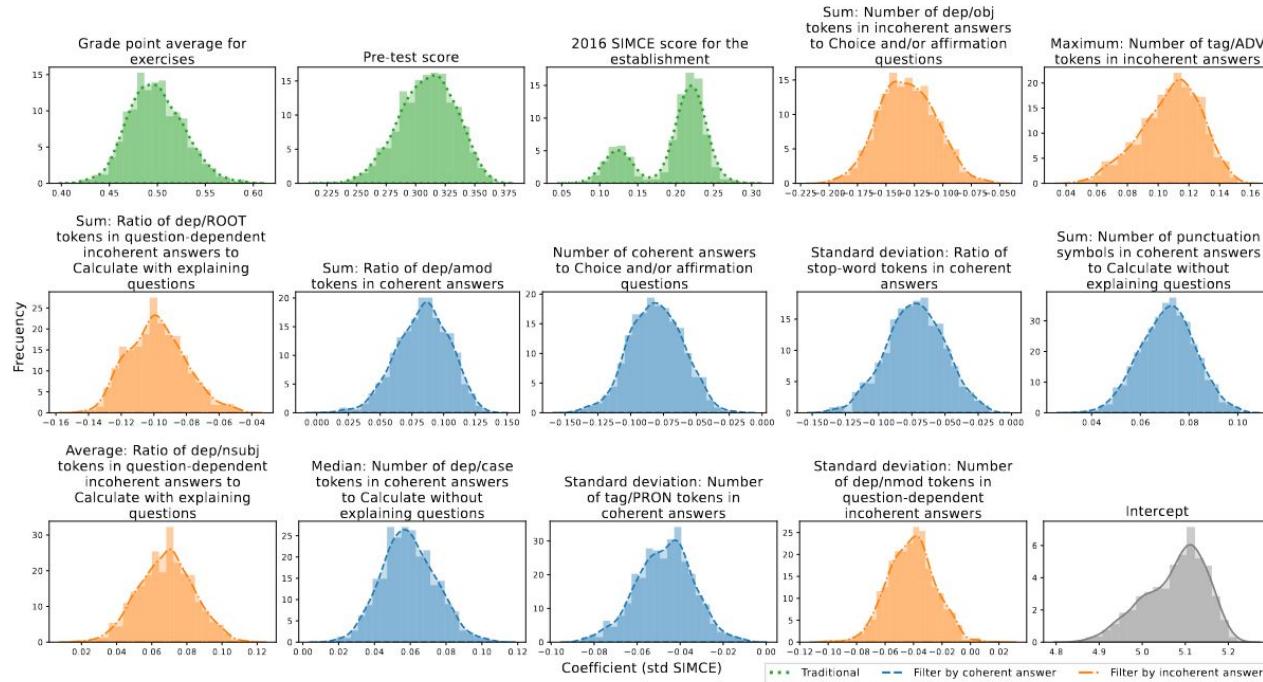
Urrutia, F., & Araya, R.  
*Journal of Intelligence*,  
10(4), 82.



**Figure 2.** Distribution of each coefficient (std SIMCE) for the baseline model. These are obtained from 250 four-fold cross-validations. (Green) Traditional regressors.

## Do Written Responses to Open-Ended Questions on Fourth-Grade Online Formative Assessments in Mathematics Help Predict Scores on End-of-Year Standardized Tests?

2022  
Urrutia, F., & Araya, R.  
*Journal of Intelligence*,  
10(4), 82.

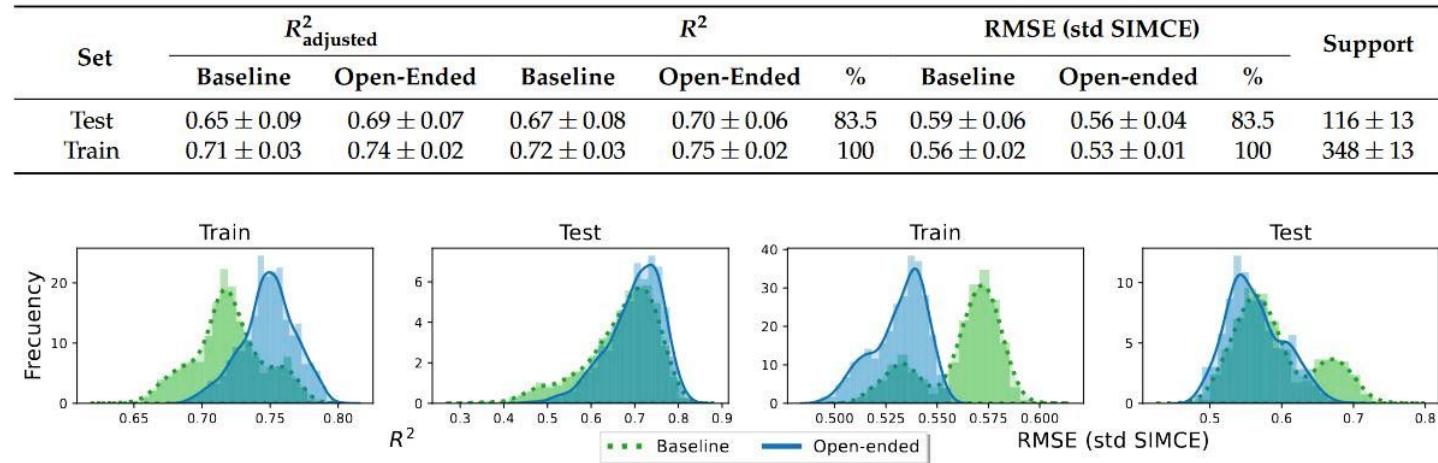


**Figure 3.** Distribution of each coefficient (std SIMCE) for the Open-ended model. These are obtained from 250 four-fold cross-validations. (Green) Traditional regressors. (Blue) Filter by coherent answers. (Orange) Filter by incoherent answers.

**Do Written Responses to Open-Ended Questions on Fourth-Grade Online Formative Assessments in Mathematics Help Predict Scores on End-of-Year Standardized Tests?**

2022

Urrutia, F., & Araya, R.  
*Journal of Intelligence*,  
10(4), 82.



Tab. 5 y Fig. 5. Urrutia, F.; Araya, R. J. Intell. 2022, 10, 82.

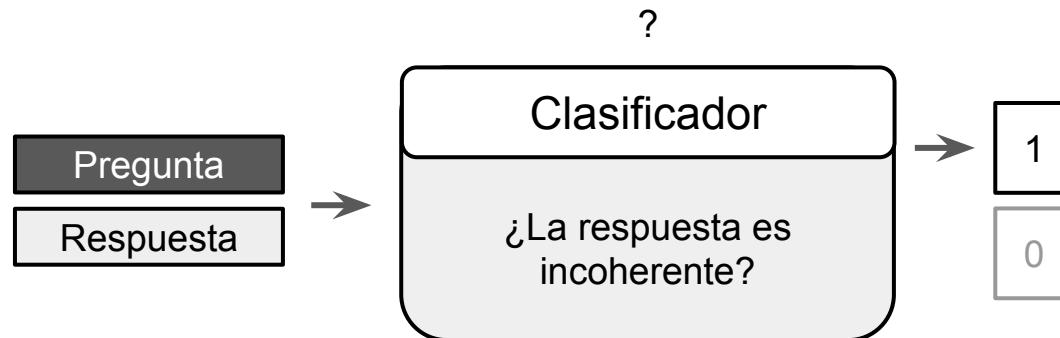
# Urrutia, F.; Araya, R. (pre-print) ESM. 2023

Automatic detection  
of incoherent written  
responses to  
open-ended  
mathematics  
questions of fourth  
graders

2023

Urrutia, F., & Araya, R.  
(pre-print) *Educational  
Studies in  
Mathematics*

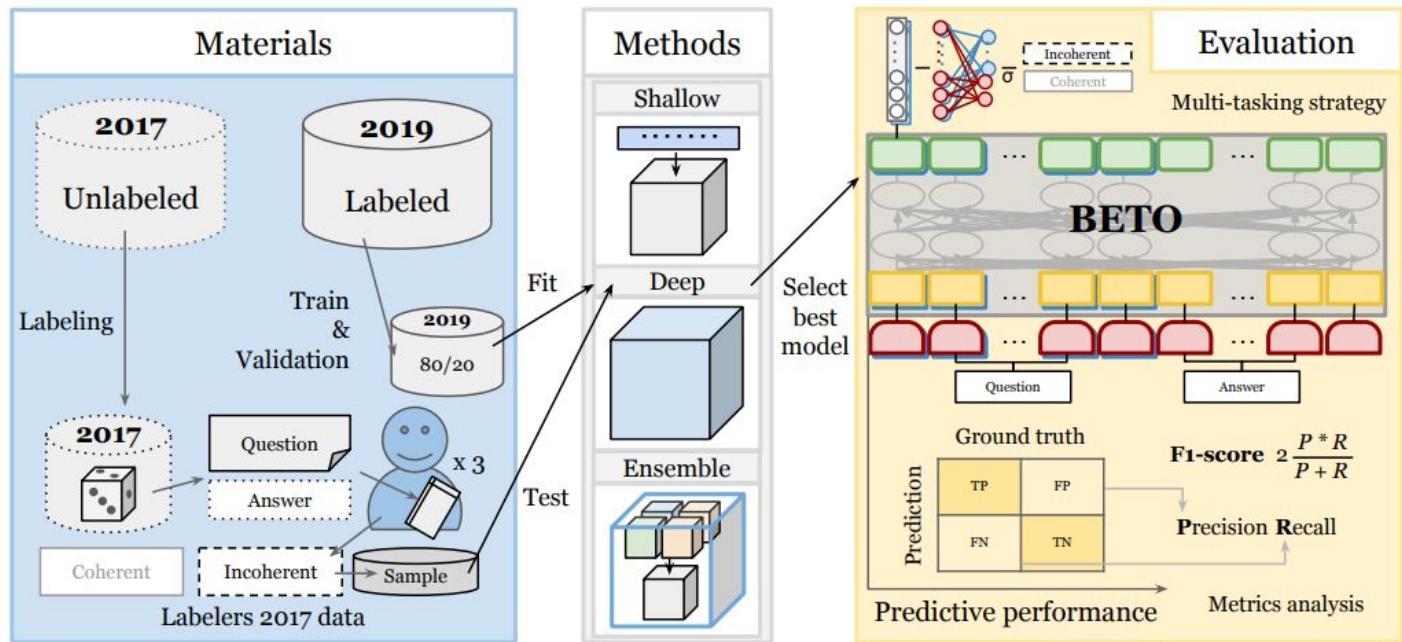
**RQ:** ¿Hasta qué punto se puede construir un clasificador automático que detecte en tiempo real las respuestas incoherentes dadas por los alumnos de cuarto grado a las preguntas abiertas de matemáticas diseñadas y escritas al momento por el profesor en una plataforma online?



# Urrutia, F.; Araya, R. (pre-print) ESM. 2023

Automatic detection  
of incoherent written  
responses to  
open-ended  
mathematics  
questions of fourth  
graders

2023  
Urrutia, F., & Araya, R.  
(pre-print) *Educational  
Studies in  
Mathematics*



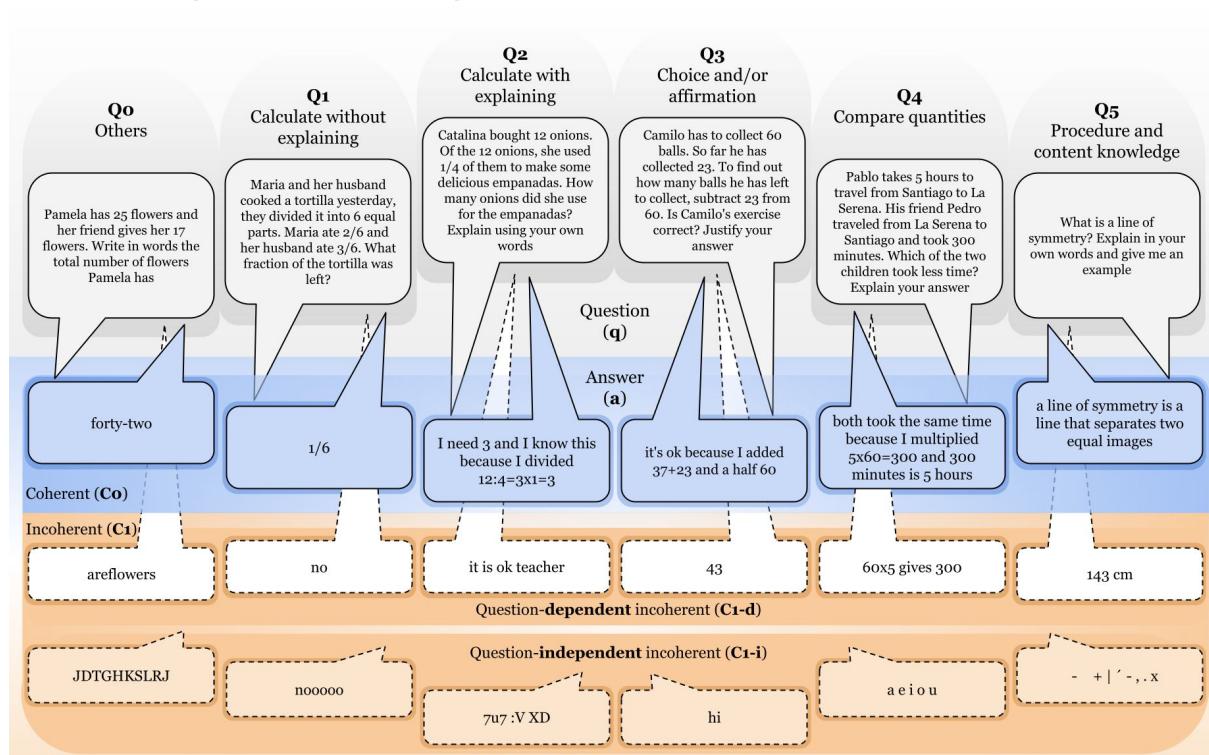
**Fig. 1** (Left) Datasets and labeling. (Center) Three types of ML algorithms. (Right) Performance evaluation.

# Urrutia, F.; Araya, R. (pre-print) ESM. 2023

## Automatic detection of incoherent written responses to open-ended mathematics questions of fourth graders

2023

Urrutia, F., & Araya, R.  
(pre-print) *Educational  
Studies in  
Mathematics*



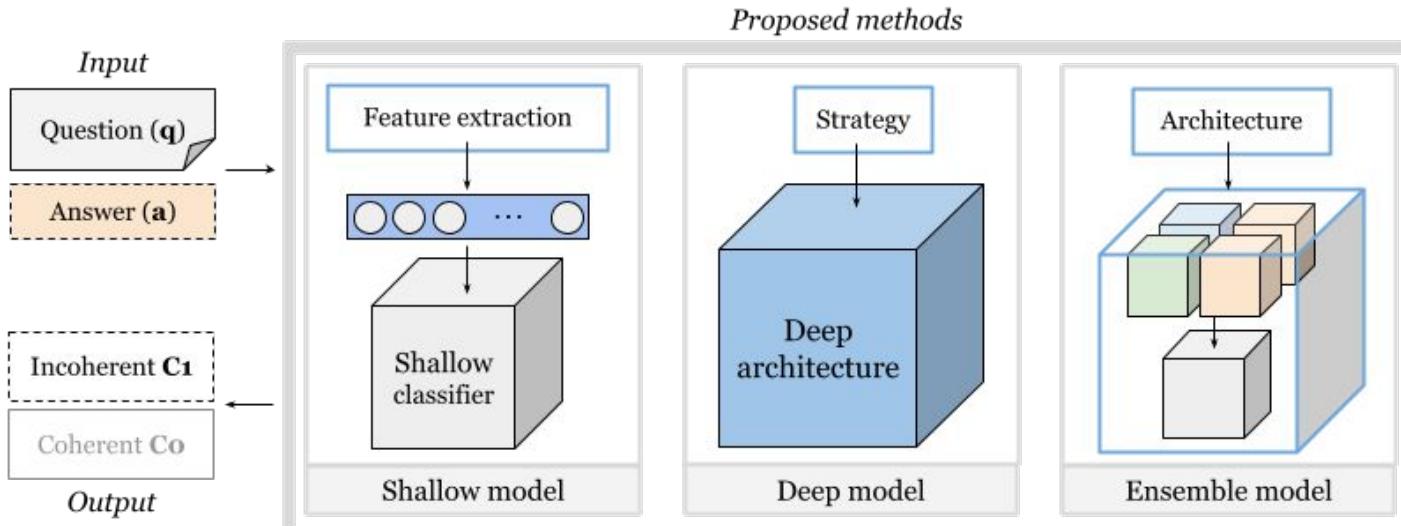
**Fig. 2** Example of open-ended math exercise questions and fourth graders answers. The first row contains the questions, separated by type and indexed with  $Q_i$ , where  $i$  is a number between 0 and 5. The second row has coherent answers, and the third row has incoherent answers dependent on the question. The fourth row contains incoherent answers that are independent of the question. Note: Examples originally in Spanish.

# Urrutia, F.; Araya, R. (pre-print) ESM. 2023

Automatic detection  
of incoherent written  
responses to  
open-ended  
mathematics  
questions of fourth  
graders

2023

Urrutia, F., & Araya, R.  
(pre-print) *Educational  
Studies in  
Mathematics*



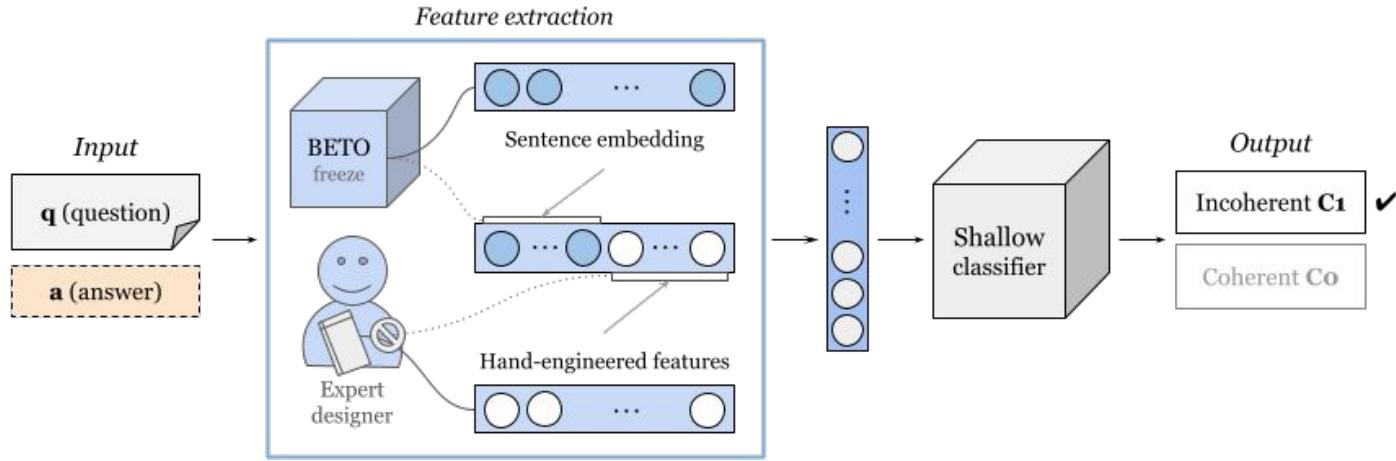
**Fig. 3** Shallow models are low complexity and suitable for simple tasks with limited data, capturing only shallow patterns. Deep models are more complex, capturing more intricate patterns. Ensemble models use a combination of shallow and deep models, depending on the task.

# Urrutia, F.; Araya, R. (pre-print) ESM. 2023

Automatic detection  
of incoherent written  
responses to  
open-ended  
mathematics  
questions of fourth  
graders

2023

Urrutia, F., & Araya, R.  
(pre-print) *Educational  
Studies in  
Mathematics*



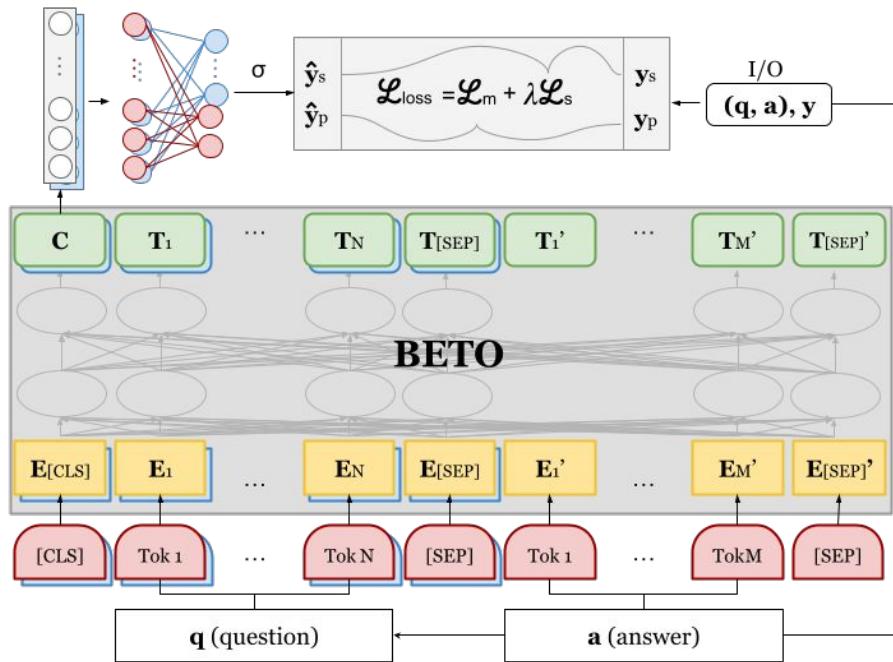
**Fig. 4** Classify incoherence in open-ended answers using various feature extraction methods, including handcrafted features and word embeddings. Interpretable features such as word count, question-answer overlap, and linguistic knowledge were used to identify incoherence (Table 4). The study utilized the Spanish version of BERT (BETO) for vector representations of text (Figure 5) and two classification models, Support Vector Machines (SVMs) and eXtreme Gradient Boosting (XGBoost). We evaluate the model and fit the parameters following Tables 1 and 3.

# Urrutia, F.; Araya, R. (pre-print) ESM. 2023

Automatic detection  
of incoherent written  
responses to  
open-ended  
mathematics  
questions of fourth  
graders

2023

Urrutia, F., & Araya, R.  
(pre-print) *Educational  
Studies in  
Mathematics*



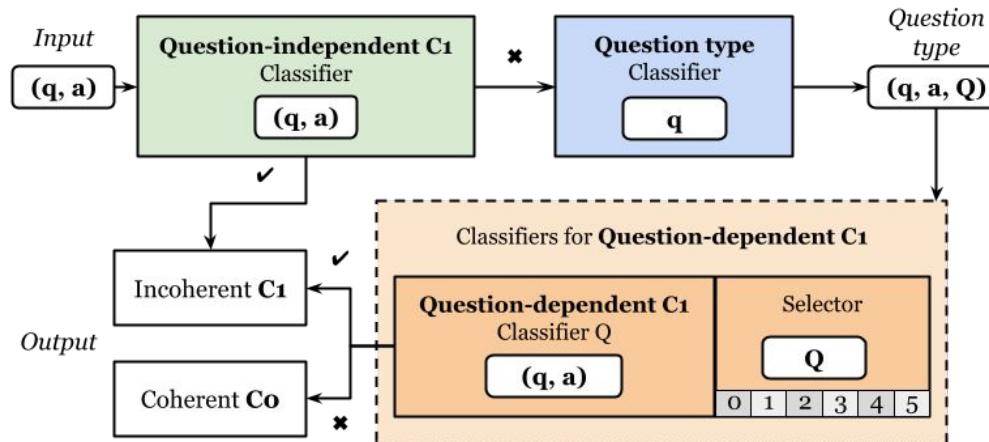
**Fig. 5** The BETO model was trained to classify incoherence in open-ended question answers using Multi-Tasking and Fine-Tuning strategies with a fixed  $\lambda = 0$ . The model was pre-trained on two self-supervised tasks, Masked language model and Next sentence prediction, using the sources of the OPUS Project and Spanish Wikipedia datasets. Fine-Tuning involved adjusting the pre-trained model by adding a last linear layer, while Multi-Tasking trained the model on multiple tasks simultaneously. Further Pre-Training involved retraining the model with intrinsic tasks specific to the domain of the main task data. For the BETO Multi-Tasking experiment, a BETO model was trained directly without further pre-training [44], using specific parameters (Table 2).

# Urrutia, F.; Araya, R. (pre-print) ESM. 2023

Automatic detection  
of incoherent written  
responses to  
open-ended  
mathematics  
questions of fourth  
graders

2023

Urrutia, F., & Araya, R.  
(pre-print) *Educational  
Studies in  
Mathematics*



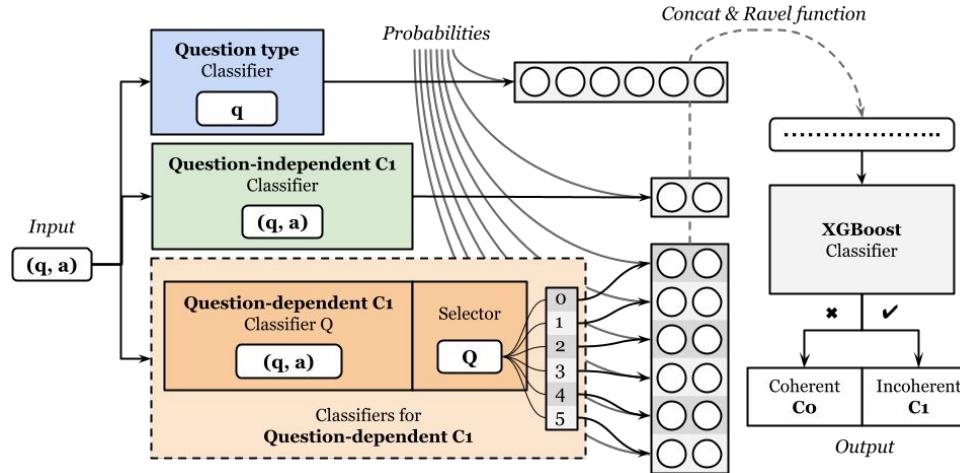
**Fig. 6** The Logical architecture is an ensemble model consisting of eight classifiers designed to identify incoherent answers to different types of questions. Question-dependent incoherence (C1-d) requires further analysis than question-independent incoherence (C1-i). The model includes a C1-i classifier, a Question type classifier (QT), and a C1-d classifier per type of question. The QT classifier is a BETO model trained with fine-tuning strategy, while the C1-i and C1-d classifiers are XGBoost models using hand-crafted features and BETO sentence embeddings. By considering the type of question, the Logical architecture can determine whether an answer is coherent or not.

# Urrutia, F.; Araya, R. (pre-print) ESM. 2023

Automatic detection  
of incoherent written  
responses to  
open-ended  
mathematics  
questions of fourth  
graders

2023

Urrutia, F., & Araya, R.  
(pre-print) *Educational  
Studies in  
Mathematics*



**Fig. 7** The General architecture incorporates the same classifiers as the Logical architecture but requires additional training to properly integrate them. This approach uses the probabilities associated with each classifier to generate a large probability vector, which is then inputted into a tree-based model to make the best decision and determine whether the answer is coherent with the question. The General architecture is an ensemble model that includes the same eight classifiers as the Logical architecture (Figure 6). To optimize the XGBoost model's parameters (Table 3), we use a grid search method over the validation set to choose the best generalizing model.

# Urrutia, F.; Araya, R. (pre-print) ESM. 2023

**Automatic detection  
of incoherent written  
responses to  
open-ended  
mathematics  
questions of fourth  
graders**

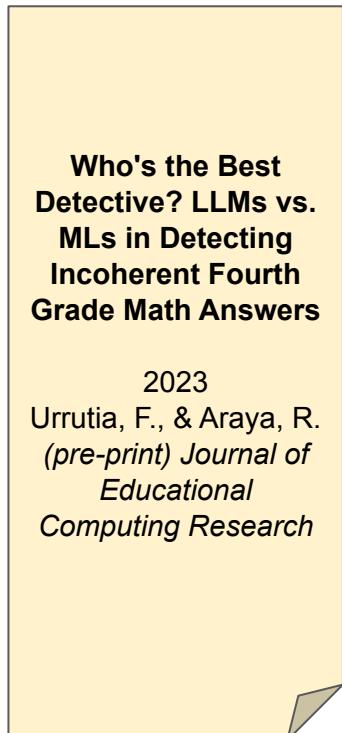
2023  
Urrutia, F., & Araya, R.  
(pre-print) *Educational  
Studies in  
Mathematics*

**Table 5** Predictive performance of the models on the test set, validation set, and train set. The category column corresponds to baselines and the three families of proposed models: shallow, deep, and ensemble models. The model column refer to the names of the experiment models, and the numerical values are the metrics associated with each dataset and model.

Category	Model	Test set			Validation set			Train set		
		P	R	F	P	R	F	P	R	F
Baseline	Dummy (most frequent)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Dummy (stratified)	20.24	12.50	15.45	15.06	15.72	15.38	13.48	13.57	13.53
	Dummy (uniform)	21.65	52.21	30.60	13.57	49.48	21.30	13.00	48.90	20.54
	NB+BOW	100.00	2.21	4.32	60.99	28.61	38.95	80.06	55.00	<b>65.20</b>
	Rule-based	69.79	49.26	<b>57.76</b>	50.46	56.96	<b>53.51</b>	48.07	53.38	50.58
Shallow	XGBoost + IIF	70.00	72.06	71.01	89.63	80.15	84.63	94.80	92.40	93.59
	<b>XGBoost + Mix</b>	70.92	73.53	<b>72.20</b>	93.97	84.28	<b>88.86</b>	99.93	99.74	<b>99.84</b>
	SVM + BETO embeddings	60.78	68.38	64.36	88.05	72.16	79.32	93.27	80.13	86.20
Deep	BETO fine-tuning	74.83	78.68	<b>76.70</b>	81.23	84.79	82.98	98.21	99.94	<b>99.07</b>
	<b>BETO multi-tasking</b>	76.19	82.35	<b>79.15*</b>	84.50	84.28	<b>84.39</b>	96.25	100.00	98.09
Ensemble	Logical	66.45	75.74	70.79	88.25	90.98	89.59	92.67	98.51	95.50
	<b>General (XGBoost)</b>	78.79	76.47	<b>77.61</b>	94.49	92.78	<b>93.63</b>	99.93	99.74	<b>99.84</b>

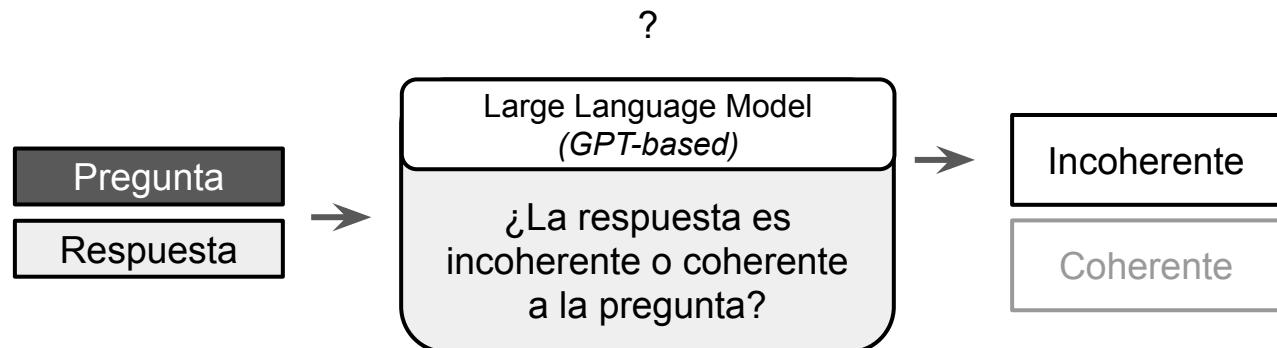
Note: Precision (P), Recall (R), F1-score (F); Support: 136 (test), 388 (val), 1540 (train).

# Urrutia, F.; Araya, R. (pre-print) JECR. 2023



**RQ1:** How do LLMs fare in detecting incoherent fourth-graders responses to typical math word problems?

**RQ2:** How does the incoherence detection performance of LLMs compare to that of ML classifiers?

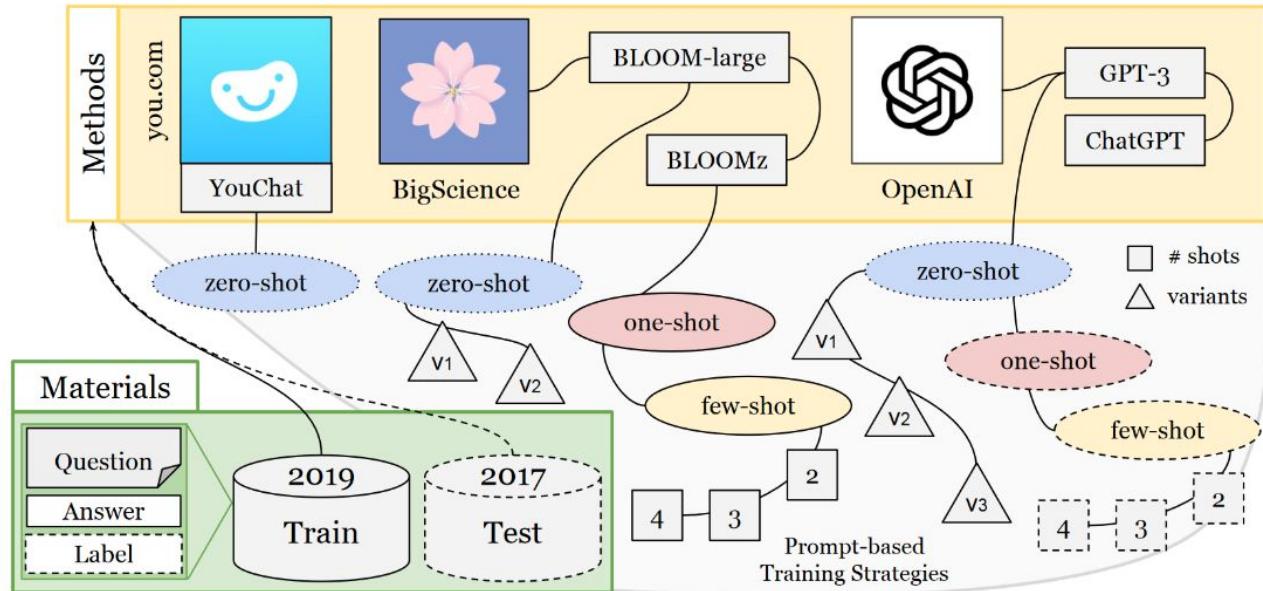


# Urrutia, F.; Araya, R. (pre-print) JECR. 2023

Who's the Best Detective? LLMs vs. MLs in Detecting Incoherent Fourth Grade Math Answers

2023  
Urrutia, F., & Araya, R.  
(pre-print) Journal of  
Educational  
Computing Research

Figure 1. The four LLMs used in this study. Three of them with prompting of zero, one, two or three shots. Below are the two databases with questions and answers. Those of 2019 will be used to train the ML models. The 2017 one is used for testing.



# Urrutia, F.; Araya, R. (pre-print) JECR. 2023

**Who's the Best Detective? LLMs vs. MLs in Detecting Incoherent Fourth Grade Math Answers**

2023  
Urrutia, F., & Araya, R.  
(pre-print) *Journal of Educational Computing Research*

## Template

Clasificar incoherencia en respuestas a preguntas abiertas de alumnos de cuarto grado en matemáticas.  
Se indicará con <E Incoherente /E> si la respuesta entre <R /R> es incoherente a la pregunta entre <P /P>. Con <E Coherente /E> si la respuesta entre <R /R> es coherente a la pregunta entre <P /P>.

## Task description

## Example 1

<P Agustín desea guardar unos dados de 1 cm cúbico en una caja que mide 5 cm de largo, 2 cm de ancho y 2 cm de alto. Su hermano dice que caben en la caja 15 dados. ¿Es correcto lo que señala el hermano de Agustín? Explica con tus palabras. /P>. <R esta mal por caben 20 dados /R>. <E Coherente /E>

## Example 2

...

## Example N

<P {Question} /P>.  
<R {Answer} /R>.  
<E \_\_\_\_\_

## Input (Question, Answer)

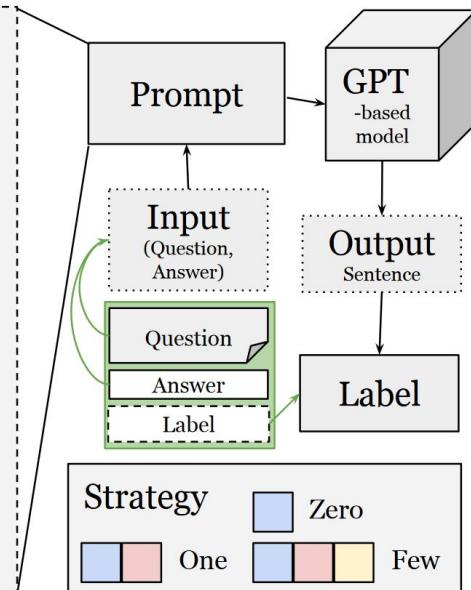


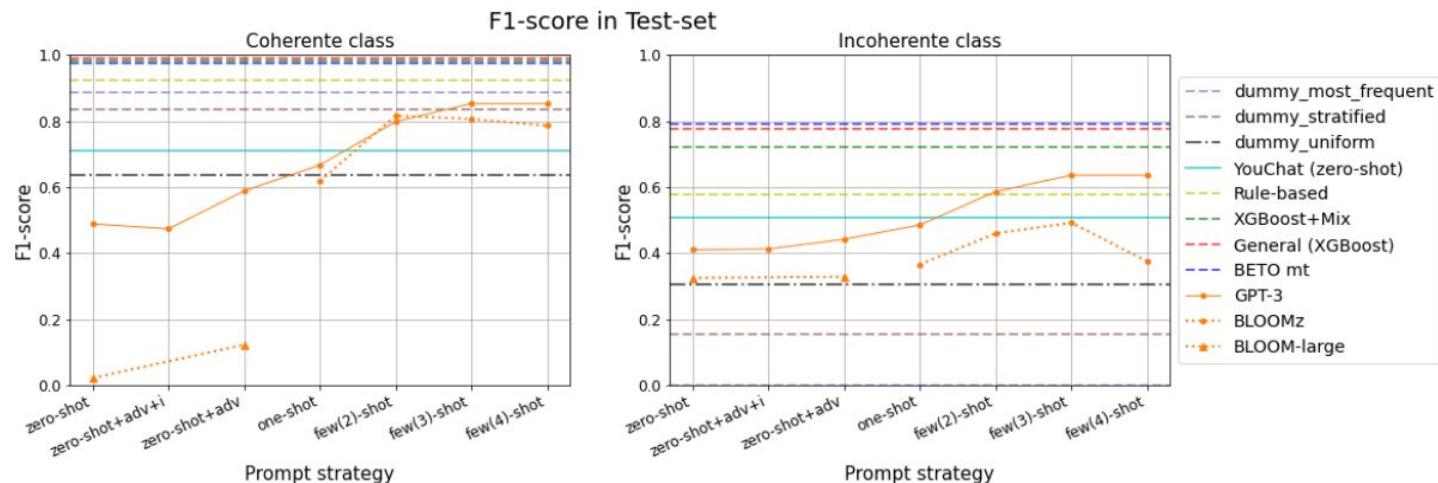
Figure 2. (Left) Structure of the Prompt-based Training Strategies. (Right) Input to the GPT and

# Urrutia, F.; Araya, R. (pre-print) JECR. 2023

Who's the Best Detective? LLMs vs. MLs in Detecting Incoherent Fourth Grade Math Answers

2023  
Urrutia, F., & Araya, R.  
(pre-print) Journal of Educational Computing Research

Figure 3. Comparison of performance between GPT and ML models. (Left) F1-score of the Coherent class. (Right) F1-score of the Incoherent class.



# Urrutia, F.; Araya, R. (pre-print) JECR. 2023

**Who's the Best Detective? LLMs vs. MLs in Detecting Incoherent Fourth Grade Math Answers**

2023  
Urrutia, F., & Araya, R.  
*(pre-print) Journal of Educational Computing Research*

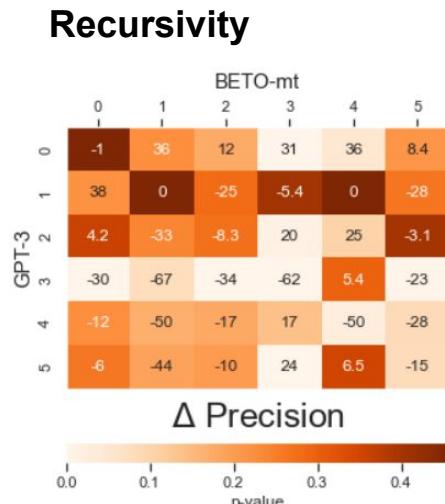


Figure 4. Differences between Precision for the Incoherent class. **(Diagonal)** Difference of Precision between GPT-3 and BETO-mt models for each question type. **(Under diagonal)** Differences of Precision between questions for the GPT-3 model. **(Over diagonal)** Differences of Precision between questions for the GPT-3 model and BETO-mt model.

Pregunta: El agricultor Braulio distribuye su cosecha de 35 choclos en 7 camionetas. Un trabajador dijo que en cada camioneta pondrán 4 choclos. ¿Está en lo correcto el trabajador? Explica con tus palabras.  
Respuesta: esta malo por que 7 camionetas  $\times$  5 es 35 choclos.  
¿Es coherente la respuesta a la pregunta?

La respuesta no es coherente con la pregunta ya que se ha proporcionado una respuesta opuesta a la pregunta original. En lugar de explicar si el trabajador está en lo correcto al decir que cada camioneta tendrá 4 choclos, se ha proporcionado una respuesta que indica que 7 camionetas multiplicadas por 5 son 35 choclos. Por lo tanto, se necesita una respuesta que aborde directamente la pregunta original para determinar si el trabajador está en lo correcto o no.

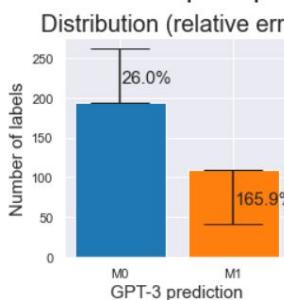
# Urrutia, F.; Araya, R. (pre-print) JECR. 2023

**Who's the Best Detective? LLMs vs. MLs in Detecting Incoherent Fourth Grade Math Answers**

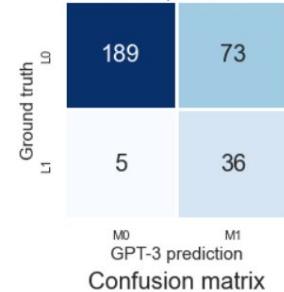
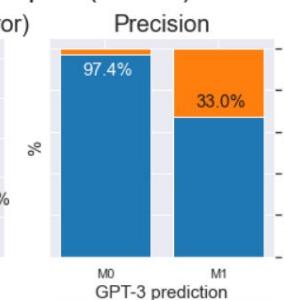
2023  
Urrutia, F., & Araya, R.  
(pre-print) *Journal of Educational Computing Research*

## Recursivity

Distribution (relative error)



Precision



New prompt report (GPT-3)

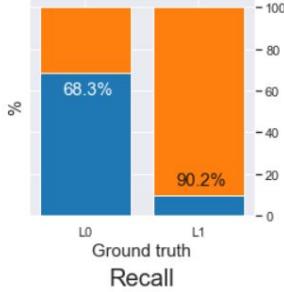
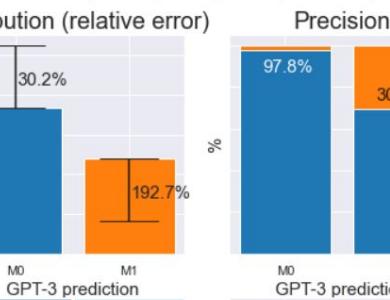
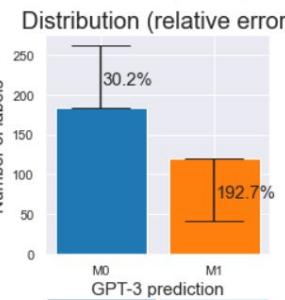


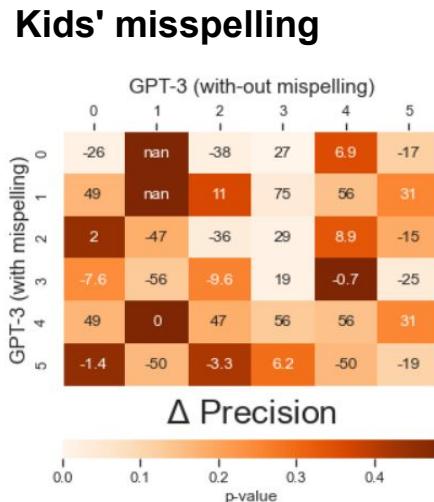
Figure 5. Comparison of results of the GPT-3 model in question type Q3. (Left) GPT-3 with 4 shots. (Right) GPT-3 with new three-question zero-shot prompt.

# Urrutia, F.; Araya, R. (pre-print) JECR. 2023

**Who's the Best Detective? LLMs vs. MLs in Detecting Incoherent Fourth Grade Math Answers**

2023

Urrutia, F., & Araya, R.  
(pre-print) *Journal of Educational Computing Research*



Pregunta: D es decena, ¿Por qué 2D de manzanas es lo mismo que 20 manzanas? Explica.

Respuesta: porque la desenavale 10.

¿Es coherente la respuesta a la pregunta?



La respuesta es incoherente y no tiene sentido. Afirmar que la "desenavale" 10 no es una explicación lógica para la equivalencia de 2D de manzanas con 20 manzanas. La respuesta correcta es que "D" representa la cantidad de unidades de diez en un número. Por lo tanto, 2D de manzanas significa 2 veces la cantidad de unidades de diez, lo que se traduce en 20 manzanas.

*Figure 6. Differences between Precision for the Incoherent class. (Diagonal) Difference of Precision between GPT-3 with and with-out misspellings for each question type. (Under diagonal) Differences of Precision between questions for the GPT-3 model in answers with misspellings. (Over diagonal) Differences of Precision between questions for the GPT-3 model in answers with-out misspellings.*

Parte 1 ...

to be continued ...

# The Role of Natural Language Processing in Advancing Competency-Based Education and Mathematics Learning in Fourth Graders

19/04/23 (RELELA)  
Felipe Urrutia Vargas



JYVÄSKYLÄN YLIOPISTO  
UNIVERSITY OF JYVÄSKYLÄ



UNIVERSIDAD DE CHILE  
INSTITUTO DE ESTUDIOS  
AVANZADOS EN EDUCACIÓN

