A cross-dimensional analysis of LLM answers from a Mathematics education perspective

Ana Rosa Corica
NIEM (UNCPBA) - CONICET
Tandil, Argentina
acorica@niem.exa.unicen.edu.ar

Patricia Sureda
NIEM (UNCPBA) - CONICET
Tandil, Argentina
psureda@niem.exa.unicen.edu.ar

Verónica Parra
NIEM (UNCPBA) - CONICET
Tandil, Argentina
vparra@niem.exa.unicen.edu.ar

Silvia Schiaffino
ISISTAN (UNCPBA/CONICET)
Tandil, Argentina
silvia.schiaffino@isistan.unicen.edu.ar

Daniela Godoy
ISISTAN (UNCPBA/CONICET)
Tandil, Argentina
daniela.godoy@isistan.unicen.edu.ar

***Abstract*—The use of Generative AI tools in educational contexts is attracting a huge amount of attention. While some works have studied the accuracy of responses to different requests, some others have focused on the language used and its suitability for educational purposes. In this work, we examine the responses of multiple chatbots to a set of Math problems for secondary school from different perspectives. Our findings can shed some light on how the chatbots use language to interact with students.**

***Index Terms*—LLMs, Feature Extraction, Math education.**

I. Introduction

The irruption of Generative AI in the educational context enabled numerous discussions on how it can support diverse teaching and learning tasks. Most of the works in this regard, are concerned with the impact that chatbots such as ChatGPT or Gemini are having on schools, some concerning uses and their potential to enhance the teaching of different disciplines.

In addition to these issues, the fast development of Large Language Models (LLMs) is also changing information seeking and discovery tasks in the classrooms. The way in which students interact with chatbots has a strong influence not only on engagement, but also on the adoption of answers and the trust that is conferred to them. These factors, in turn, will shape the parasocial relationship of students with AI models [1].

In this work, we investigate the different characteristics of texts generated by LLMs with respect to the writing style (readability and lexical richness) and tone (sentiment and emotions) to shed some light on the way they interact with students. In particular, we are interested in the behavior of chatbots when answering about Math for being one of the subjects entailing more difficulty, leading even to students' frustration.

The rest of the article is organized as follows. In section II we briefly summarized some related works. In section III we describe the data collection process and the characteristics we have analyzed in chatbots' responses. Then, in section IV we report and analyze the results obtained, and in section V we present our conclusions.

II. Related Works

The use of ChatGPT and similar chatbots in education is being increasingly discussed. Some recent works focus on how the generated text impacts on information-seeking tasks. For example, in [2], [3] the authors present the results of a preliminary exploration aiming to understand whether ChatGPT can adapt to support children in completing information discovery tasks in the education context. They analyze aspects such as readability and language used in responses. The analysis conducted, with feedback from children (9 to 10 years old) indicates that ChatGPT is suitable for 4th grade level. However, the authors acknowledge that it still needs improvement to reach the right level of readability. The work presented in [4], describes the results of an evaluation of the lexical diversity of the text generated by LLMs in English and how it depends on the model parameters, with the purpose of understanding how LLMs use the language. In [5], it is argued that it is crucial to recognize the role of emotions in searching activities that children undertake in the classroom as they are integral to he "information search process because they affect a searcher's attention, memory, performance, and judgments".

III. Materials and Methods

In order to characterize the responses of LLMs when answering about Math, we collected a dataset of problems corresponding to 1st to 6th year of the Argentinian secondary school. This set includes from numeric problems to problems related to geometry or statistics, and they were extracted from the curricular designs of different provinces. Thus, students in a given year are assumed to be able to understand and to have the necessary knowledge to solve them.

For each of the problems, three answers of Gemini[1], Copilot[2], Llama[3] and ChatGPT4o[4] were collected. In this study, 10

[1]https://gemini.google.com/
[2]https://copilot.microsoft.com/
[3]https://llama.ai/
[4]https://chatgpt.com/

problems were collected for every year of the secondary school and, for account for randomness, 3 answers were collected for each of them, so that overall the dataset consists of 480 answers gathered from the chatbots.

We considered features in multiple dimensions to analyze the characteristics of each chatbot's answers. The text of the responses was processed for extracting these features by leveraging on both lexicons, including the Linguistic Inquiry and Word Count (LIWC) [6] and NRC Emotion Lexicon, and available trained models, such as VADER (Valence Aware Dictionary and sEntiment Reasoner) [7]. The dimensions analyzed are the following:

- *Readability*: the readability of a piece of text allows to determine its difficulty regarding the level of instruction needed to understand it. A number of readability formulas are used to quantify this aspect. In this work, we used formulas adapted for Spanish language as Flesch Reading Ease [8], Gutiérrez de Polini's Readability Formula [9] and Szigriszt-Pazos Perspicuity Index [10], implemented in the TextStat library[5].
- *Lexical richness*: this dimension measures the diversity of the vocabulary employed in texts and it is used as proxy for establishing the language sophistication. We used different metrics of lexical richness extracted using the LexicalRichness Python library[6].
- *Sentiment and emotions*: positive, negative and neutral sentiment was estimated using VADER, a lexicon and rule-based sentiment analysis tool which provides sentiment-related scores. Likewise, positive and negative emotions were gleaned with a lexicon-based approach as *EmoLex*, covering the eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust).
- *Semantic categories*: semantic-level categories capture meaning behind words. In this case, word frequencies are assigned to lexicons or phrases that fall into psycholinguistic categories (e.g., those defined in LIWC). For this work, we were interested in observing hints of different cognitive processes.

IV. Results

A. Readability

To assess the readability of chatbots' answers, we used three formulas, Flesch Reading Ease and, specifically for Spanish texts, Gutiérrez de Polinis and Szigriszt-Pazos readability indices. These formulas give a text a score between 1 and 100, with 100 being the highest readability score (easiest to read). Figure 1 shows the readability scores of answers in relation with the level of instruction needed to understand it as proposed by each formula (in the right Y axis). This level is usually considered taking as reference the US school level, so 1st in Argentina is 7th in US education level.

Several observations can be made regarding these results. As assessed with the three formulas, text readability is about

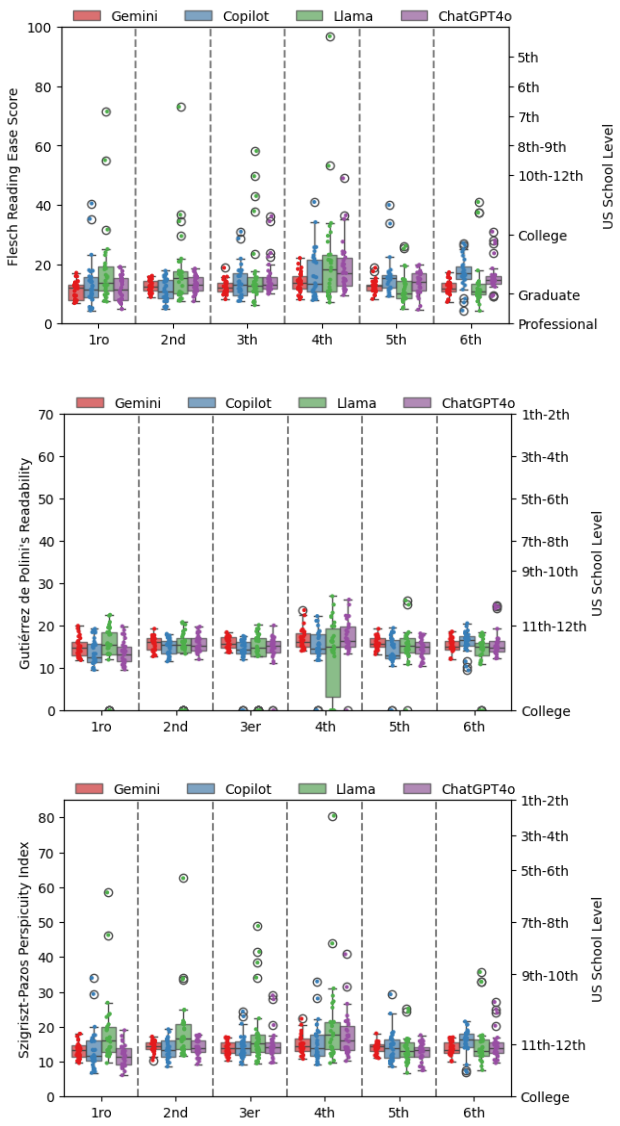

Fig. 1: Readability level of chatbots answers

or higher 11th-12th grade students, i.e. 16-18 years old. Then, it is suitable for 6th year students of secondary schools in Argentina, but excessively complex for lower grades. There are little differences regarding readability among the four chatbots, mean values are close to each other in every year. For problems proposed for the 3 first years, Llama seems to be consistently the one providing easiest to read answers in average (although not by a large margin), but it is also the one with greater variability. Gemini, on the other hand, shows less deviations from the average scores in all cases.

B. Lexical Richness

Lexical diversity is usually measured through the calculation of type-token ratio (TTR) which is the number of unique words divided by their occurrences. Measure of Textual Lexical Diversity (MTLD) and Hypergeometric distribution diversity (HDD) correct for bias introduced when comparing texts of

[5]https://pypi.org/project/textstat/
[6]https://pypi.org/project/lexicalrichness/

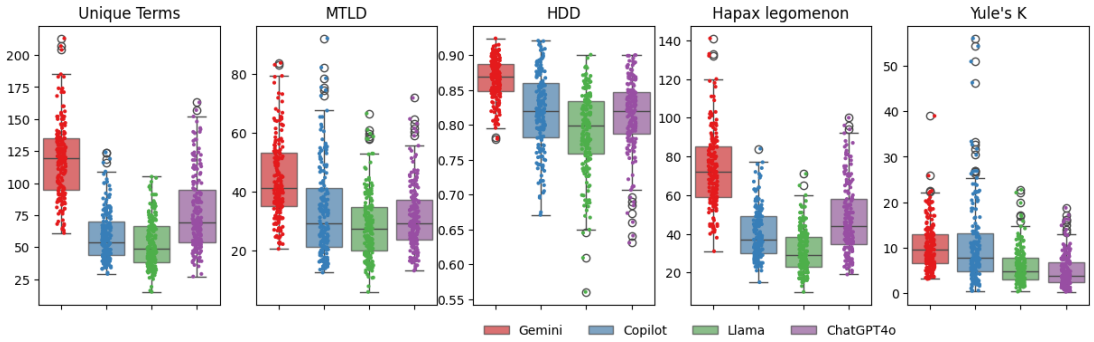

Fig. 2: Lexical richness of chatbots answers

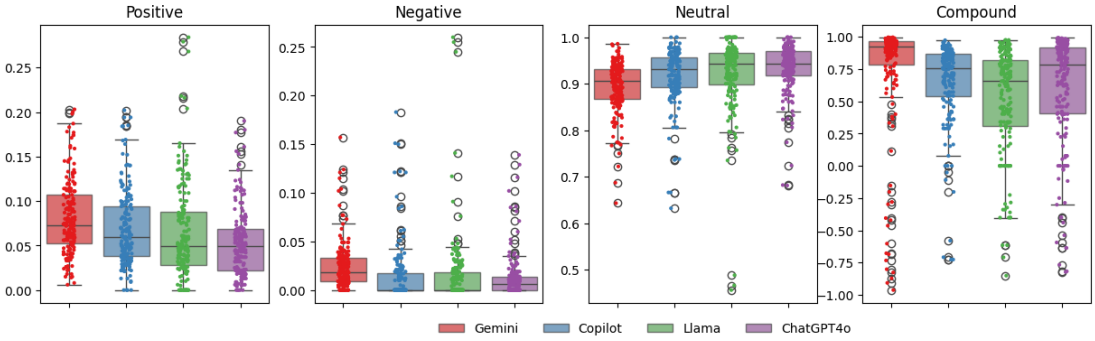

Fig. 3: Sentiment analysis of chatbots answers

different lengths and they are more robust measures than TTR-based measures. In addition, we used Hapax legonema that refers to the number of words appearing only once in the text and Yule's K, which is considered to be highly reliable for being a text length independent measure.

Figure 2 shows the average values of these measures, in addition to the number of unique terms. Gemini is the chatbot showing a richer vocabulary, being superior in lexical richness as assessed for all the metrics than the rest of the chatbots. On the other hand, in almost every case Llama was the one using less unique words, being considered the one with lowest lexical richness according to most of the metrics.

C. Sentiment and Emotion Analysis

Figure 3 shows the sentiment analysis of the collected answers. The three first plots correspond to the distributions of positive, neutral and negative sentiment scores extracted with VADER. The three are calculated as ratios for proportions of text that fall into each category (all adding up to 1). All chatbots exhibit a high level of neutrality in their answers, with a slightly higher average of positive sentiment than negative one. Gemini is the less neutral according to these results, showing greater both positive and negative sentiment scores than the remaining chatbots. ChatGPT4o is, on the other extreme, being the most neutral of all. The compound score depicted in the figure is computed by summing the valence scores of each word in the lexicon, normalized to be between -1 (most extreme negative) and +1 (most extreme positive). Thus, it provides a single uni-dimensional measure of sentiment for a given text that allows us to determine the overall (compound) sentiment that is conveyed or embedded in rhetoric. Compound sentiment scores in the figure show that all chatbots convey a positive sentiment overall. Gemini is the most positive one, while Llama is the least positive.

As regards emotions found in texts, Figure 3 shows that both positive and negative emotions appear in low proportions. Out of the 8 emotions extracted with Emolex, words related to *trust* are the most frequent in the texts, while words related to *joy* appear just as much as negative emotions like *fear* or *sadness*. Overall, the texts are not rich on emotion-related words, being Gemini the one exhibiting more presence of this type words than others.

D. Cognitive processes

Focusing on the presence of cognitive mechanisms operationalized through LIWC, Figure 5 depicts the values obtained for the 5 cognitive processing categories. From these categories it is possible to infer some sort of analytical thinking involved in the LLMs-generated content, extracted from words denoting *insight* (think, realize, perspective), *causation* (because, effect, based), *discrepancy* (should, hope, lack), *tentativeness* (maybe, perhaps,wonder), *certainty* (always, never, clearly) and *differentiation* (despite, although, except).

In this regard, *insight* is the category with higher mean values, implying that answers describe some type of thinking process when solving the problem. Given that the prompts

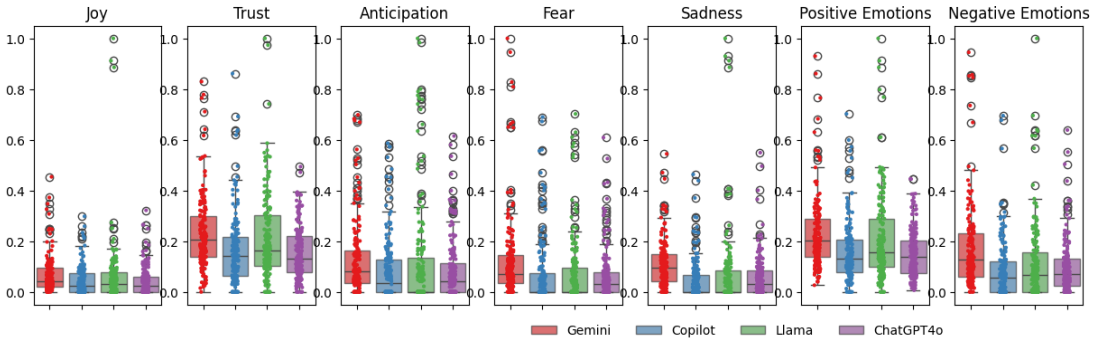

Fig. 4: Emotion analysis of chatbots answers

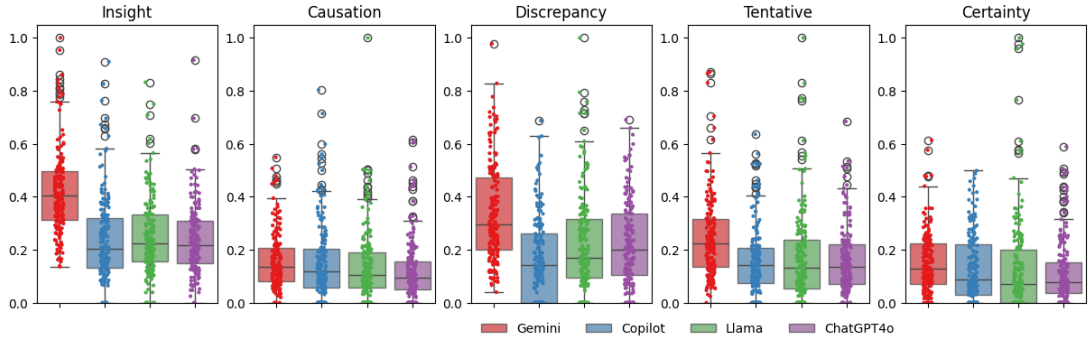

Fig. 5: Cognitive processes in chatbots answers

were Math problems, it draws our attention that words related to *causation* are not strongly present in the texts. This might denote a weak linkage between reasoning steps of the provided solutions. Instead, the category *discrepancy* (denoted by words such as *conflict*, *contradict*, *disagree*, etc.) has relatively higher mean value than expected. The relation between the categories *tentativeness* and *certainty* can also have some interesting interpretation as chatbots seem to resort more tentative terms rather than to those transmitting certainty. This observation has an interesting conclusion, that confidence in the answers is more a user construction than a result of the language used by the chatbots to communicate.

V. Conclusions

This work presents a preliminary exploration of a number of dimensions serving to characterize the texts generated by chatbots in response to Math problems. This study helps us understand how chatbots interact with students, which can be used as an initial step to more guided personalization techniques. In the future, we plan to assess the impact of different prompting techniques to evaluate the responses adaptation to school levels and other interaction goals.

Acknowledgment

This work was supported by PICT-2020-SERIEA-01375, 03-PEIDYT-29C and PIP CONICET 11220220100429CO.

References

[1] T. Maeda and A. Quan-Haase, "When human-AI interactions become parasocial: Agency and anthropomorphism in affective design," in *Proc. of the 2024 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil, 2024, pp. 1068–1077.

[2] E. Murgia, M. S. Pera, M. Landoni, and T. Huibers, "Children on ChatGPT readability in an educational context: Myth or opportunity?" in *Adjunct Proc. of the 31st ACM Conf. on User Modeling, Adaptation and Personalization (UMAP '23)*, 2023, pp. 311—-316.

[3] E. Murgia, Z. Abbasiantaeb, M. Aliannejadi, T. Huibers, M. Landoni, and M. S. Pera, "ChatGPT in the classroom: A preliminary exploration on the feasibility of adapting ChatGPT to support children's information discovery," in *Adjunct Proc. of the 31st ACM Conf. on User Modeling, Adaptation and Personalization (UMAP '23)*, 2023, pp. 22—-27.

[4] G. Martínez, J. A. Hernández, J. Conde, P. Reviriego, and E. Merino-Gómez, "Beware of words: Evaluating the lexical diversity of conversational llms using chatgpt as case study," *ACM Transactions on Intelligent Systems and Technology*, 2024.

[5] M. Landoni, M. S. Pera, E. Murgia, and T. Huibers, "Inside out: Exploring the emotional side of search engines in the classroom," in *Proc. of the 28th ACM Conf. on User Modeling, Adaptation and Personalization (UMAP '20)*, 2020, pp. 136—-144.

[6] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic inquiry and word count: LIWC 2001*. Lawrence Erlbaum Associates, Mahway, 2.

[7] C. J. Gilbert and E. Hutto, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. of the 8th Int. Conf. on Weblogs and Social Media (ICWSM-14)*, 2014.

[8] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," University of Central Florida, Tech. Rep., 1975.

[9] L. E. G. de Polini, "Investigación sobre lectura en Venezuela," in *Primeras Jornadas de Educación Primaria*, 1972.

[10] F. S. Pazos, "Sistemas predictivos de legilibilidad del mensaje escrito: F," Ph.D. dissertation.