Can an LLM build an application to automatically assess students' interpretations of diagrams?

Grecia Garcia Garcia[1] **and Peter C-H. Cheng**[1]

Abstract. RIST is a theory of the mental structures of external representations, and RISN is its notation [1, 5]. In this exploratory study, we developed a pilot web application with the assistance of Claude's Sonnet 4.6 to: (a) investigate whether RIST/N can be used for pedagogic evaluation of students' diagrammatic interpretations, and (b) identify the challenges that must be addressed to make such a system effective for educators. This application processes student interpretations of diagrams and provides feedback to the teacher on where students are experiencing difficulties. The system has three inputs: (i) a diagram, (ii) a student's written interpretation of that diagram, and (iii) the teacher's normative RISN model of the diagram. The application constructs a RISN model of the student's interpretation, compares it with the teacher's model, and derives feedback for tailoring lessons. Initial trials show some promise that a more robust system could be built. However, issues such as consistency in constructing RISN models, transparency in the AI assistant's internal processes for building the system and improving the usability of the output remain to be addressed.

1 INTRODUCTION & VISION

Despite the ubiquity of diagrams, charts and graphs in education, formal training on their use is lacking. Thus, it is not surprising that, when students are formally evaluated in examinations, a variety of interpretations and misconceptions can surface (e.g., [6], [8]).

Our vision is to create a diagnostic tool for teachers that helps evaluate students' ability to interpret graphs and charts, independently of their understanding of the topic content. We aim to do this in a way that is grounded in cognitive theory of how external representations are interpreted. To achieve this, we use an AI assistant to help develop a pilot web application. This application is not a tutoring system (such as [7] or many LLM applications [9]), but a tool that supports educators to diagnose the quality of *students' graph interpretations*.

The cognitive theory we adopt is Cheng's cognitive *Representational Interpretative Structure Theory* (RIST) and its accompanying notation (RISN) [1, 3, 4, 5]. It proposes that an interpretation of an external representation is a hierarchical network comprised of four kinds of mental schemas: *Representation*, *R-scheme*, *R-dimension* and *R-symbol*. Figure 1 shows an example of a RISN model for interpreting a simple three bar Income by Age-profile chart. It contains: a *Representation* schema of the chart (Figure 1, E1), *R-schemes* that coordinate data, age and data points (E2, D3), *R-dimensions* for each of those dimensions (B4, E4, G4), and *R-symbols* for specific data points (B6, D6, F6). Crucially, each schema coordinates information between a target concept (internal representation) and its graphical representation (external representation). Different interpretations create different schema networks. Stockdill et al. [10] identified recurring patterns of schemas in these network structures, called *idioms*.

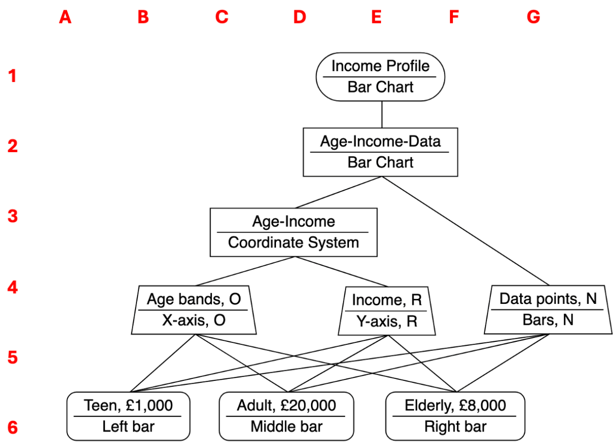

Figure 1. A simple example of a RISN model for a line graph. The red index grid was added for clarity.

RIST/N can be used to model interpretations of various diagrams [10], evaluate the usability of representations [3], study cognitive artefacts such as indexing systems [4], and observe changes in knowledge structures during learning [2]. However, it has not yet been applied in an educational context, as a key challenge is training people to create RISN models. Thus, in this work, we use a Large Language Model (LLM) as the underlying engine for automatic creation and comparison of RISN models. This is an aspect that has not been explored and would be beneficial to automate.

Our vision is as follows. We start by modelling a student's written interpretation as a hierarchical RIST/N network to identify omissions or misinterpretations of concepts and diagram relationships. An LLM is used: (a) to generate a RISN model of a student's interpretation; (b) to compare it to the teacher's target RISN model; and (c) to provide a report to the teacher of missing schemas and global and local misinterpretations. Figure 2 illustrates this pipeline. Here we will consider three questions about such a system:

Q1. Can an application be created so that it can produce reasonable RISN models from a student's written description of a diagram?

Q2. Can it successfully compare a student's RISN model with a teacher's target RISN model?

[1] Department of Informatics, University of Sussex, BN1 9QH, UK. Email: {`g.garcia-garcia`, `p.c.h.cheng`}`@sussex.ac.uk`.

Q3. Can it provide reliable, comprehensible and useful summaries of the differences between the student and the teacher models verbally and graphically?

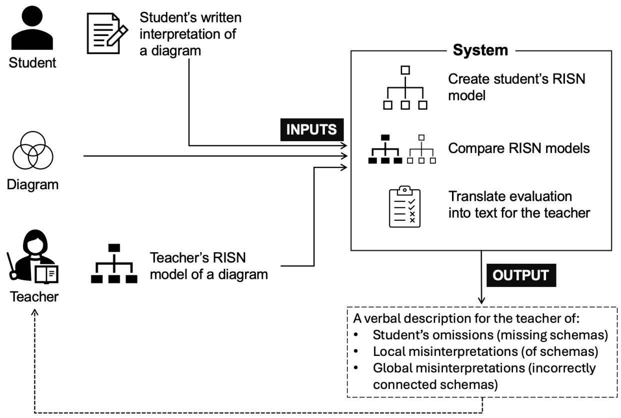

Figure 2. Pipeline for using RIST/N in an educational setting.

2 WEB APPLICATION

We developed a pilot web application with the assistance of Claude's Sonnet 4.6. It was used both during the development of the web application and was integrated into the system for background processing. Published papers on RIST/N were fed to Claude, and it implemented the application in JavaScript and HTML.

The system was refined through multiple iterations. We provided examples of simple to complex line graphs (e.g., speed/distance graphs, Playfair's trade balance line graph) and a variety of interpretative statements (including misconceptions such as the graph-as-picture misconception [6]). We reviewed the generated RISN models and reports for the teacher and prompted the AI assistant to modify the code when issues arose. To support development, we added an optional feature to generate synthetic interpretations of the diagrams at novice, intermediate and expert knowledge levels – this was used in addition to testing the system with actual users' interpretations collected from previous research projects.

In use, the web application receives a diagram, a student's written interpretation of a diagram, and the teacher's normative RISN model (Figure 2, left). Internally, the application generates a student's RISN model from the student's interpretation, compares it with the teacher's model, and produces a summary of the differences between them (Figure 2, top right), making calls to Claude as needed. The output is a report highlighting omitted concepts as well as global and local misinterpretations (Figure 2, bottom right).

3 SYSTEM REVIEW

We reviewed each of the key steps of the pipeline: examined the generated RISN models to ensure they fit with RISN's rules, verified whether the student's written interpretation matched the generated RISN model, and conducted heuristic evaluations of the generated teacher's feedback. The following list describes some of the main issues.

Creating a RISN model from a student's written interpretation. The creation of the student's RISN model was inconsistent, as different models were generated from the same input. Despite this, most generated models were valid (i.e., well-constructed), and the validity of RISN models improved with minor refinements of the prompt regarding RISN construction rules. To address the production of inconsistent models, however, we may need to add rules regarding the generation of RISN models. For instance, we may need to specify that RISN schemas should be added in the order they appear in the written description, that schemas should be created statement by statement rather than allowing the LLM to infer relationships automatically, and to limit the topic knowledge that the LLM introduces in the creation of models.

Comparing a student's and a teacher's RISN models. Inspecting the Claude-written code, we found two methods were used to compare models: (1) matching strings of text from the schemas, and (2) sending the RISN models to Claude to perform structural comparisons. This caused inconsistencies in the output. For example, two strings of text could match, but if their schemas were in different locations within the student's and the teacher's models, then it was marked as a mismatch. Also, semantically equivalent strings were marked as different (e.g., 'life stage' and 'age group' would be marked as different). Although this could be remediated by providing further explicit instructions on how to compare models, this highlights the potential imprecision of using an AI assistant to process data in the background.

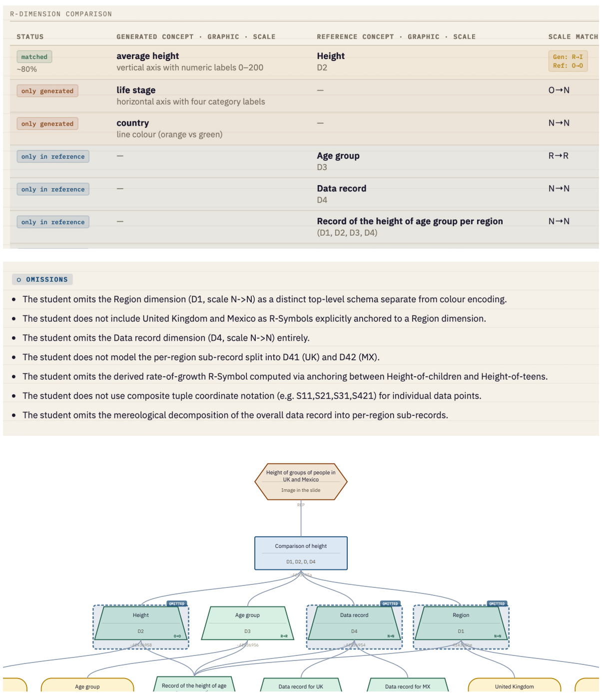

Figure 3. Snapshots of the feedback output for teachers. Formats from top to bottom: table, text summary, and annotated graph.

Output generated for the educator. Different formats were explored for presenting feedback for the teacher: (a) a table comparing the schemas and idioms from both the student's and teacher's models; (b) a text summary of omissions, local and global misinterpretations; and (c) the teacher's RISN model annotated with missing concepts, local and global misinterpretations. See Figure 3. In terms of usability, the table presentation provided the best way to quickly review missing concepts, while the text summary was too dense to be practical, and the annotated graph was inconsistent with the text output.

4 DISCUSSION & FUTURE WORK

In this exploratory study, we developed a pilot web application that uses an LLM to create and compare RISN models for supporting educators in understanding students' mental structures of their diagrammatic interpretations. In particular, it focuses on supporting educators in identifying whether students have omitted relevant concepts and their relationships, including both local and global misinterpretations.

Consider the three questions posed. For (Q1), we found that the system can generate reasonable RISN models, but identical inputs can produce different models. Addressing this requires clearer rules for deriving RISN models along with constraints on additional topic knowledge introduced by the LLM.

For (Q2), it was found that the application could compare the student's and teacher's RISN models. However, Claude applied two different comparison methods, which resulted in inconsistent feedback. Our intention was not to use different methods; however, this highlights that (a) prompt design requires careful attention, including the explicit specification of implicit logical constraints (e.g., using the same comparison method across all stages), and (b) code generated by agents must be reviewed.

On the other hand, this issue highlighted the need to develop methods for comparing and identifying (equivalent) network structures within RISN models. Further investigation would then need to use a wider variety of diagrams and students' (mis)interpretations to identify which method is most accurate. Once an effective method is developed, Claude could be instructed to include it as part of the system.

(Q3) concerns the reliability, comprehensibility and usefulness of the summaries of the differences between the student and the teacher models, verbally and graphically. This pilot study represents an initial step towards addressing this question. Of the various reporting formats explored, tables appeared to be the most useful, as they provided a quick visual way to highlight missing concepts. Nevertheless, alternative formats that offer additional value for the teacher may need to be explored. For example, feedback could be more concise and focused through brief three-point summaries of critical issues.

This study revealed that there are ample opportunities for future work. For example:

- We presented the example using a single graph interpretation as an input. However, the broader vision is for the system to collect graph interpretations from multiple students to enable the teacher to provide immediate feedback during the class. This is similar to using a short poll during a class or workshop, where participants' responses help guide subsequent activities.
- The evaluation of this system should include feedback from educators on the reliability and practicality of recommendations, as well as its potential for implementation in classroom settings.
- The system should be tested more extensively using a wide variety of diagrams, students' interpretations (from novices to experts), and types (correct, incorrect, partial interpretations, and misconceptions). These should be collected from real people, as the extent to which synthetic data may have influenced the generation of RISN models remains unclear.
- Finally, an alternative scenario could be explored in which students *also* receive feedback on how their interpretations differ from the teacher's expectations. Although this approach will require consideration of other factors (e.g., students' metacognitive skills, confidence and engagement on the topic), it could provide valuable opportunities for students to generate explanations and deepen their understanding of the use of graphs and diagrams.

REFERENCES

[1] Cheng, P. C-H. (2020). A sketch of a theory and modelling notation for elucidating the structure of representations. In A. Pietarinen, P. Chapman, L. Bosveld-de Smet, V. Giardino, J. Corter, & S. Linker (Eds.), *Lecture Notes in Computer Science: Vol. 12169. Diagrammatic Representation and Inference* (pp. 93-109). Cham: Springer.

[2] Cheng, P. C-H., & Garcia Garcia, G. (2025). What and how schema networks are acquired during the learning of line graphs: Modelling using representational systems theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.

[3] Cheng, P. C-H., Garcia Garcia, G., Raggi, D., & Jamnik, M. (2024). A human information processing theory of the interpretation of visualizations: demonstrating its utility. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). Association for Computing Machinery, New York, NY, USA.

[4] Cheng, P. C-H., Garcia Garcia, G., Raggi, D., & Jamnik, M. (2024). Index systems: Enumerating their forms and explaining their diversity with representational interpretive structure theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46.

[5] Cheng, P. C-H., Stockdill, A., Garcia Garcia, G., Raggi, D., & Jamnik, M. (2022). Representational interpretive structure: theory and notation. In: V. Giardino, S. Linker, R. Burns, F. Bellucci, J.-M. Boucheix, & P. Viana, (Eds.), *Lecture Notes in Computer Science: Vol. 13462. Diagrammatic Representation and Inference* (pp. 54-69). Springer, Cham.

[6] Garcia Garcia, G. & Cox, R. (2009). Children who Interpret Graphs as Pictures. In V. Dimitrova, R. Mizoguchi, B. du Boulay & A.C. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED 2009* (pp. 695-697). IOS Press.

[7] García-Méndez, S., de Arriba-Pérez, F. & Somoza-López, M.d.C. A. (2025). Review on the use of large language models as virtual tutors. *Science & Education*, *34*, 877-892.

[8] Kerslake, D. (1977). The understanding of graphs. *Mathematics in School*, *6*(2), 22-25.

[9] Lyu, W., Wang, Y., Chung, T., Sun, Y., & Zhang, Y. (2024). Evaluating the effectiveness of LLMs in introductory computer science education: A semester-long field study. *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)* (pp. 63-74). Association for Computing Machinery, New York, NY, USA.

[10] Stockdill, A., Garcia Garcia, G., Cheng, P. C-H., Raggi, D., & Jamnik, M. (2022). Cognitive modeling of interpretations of representations. In J. Macbeth, L. Gilpin, & M. T. Cox (Eds.), *Proceedings of the Tenth Annual Conference on Advances in Cognitive Systems* (pp. 36-56). Wright State University, Dayton, OH.