In this study, the evaluation of AI models was conducted by collecting text-based analytical outputs generated by three AI models and structuring them into a questionnaire-based assessment.

To ensure the scientific validity of the questionnaire, both reliability and validity measures were incorporated. The Cronbach’s alpha coefficient was used to assess reliability, yielding a final value of 0.93, indicating high internal consistency. For validity, the questionnaire included evaluation criteria that maintained a strong conceptual alignment with the analytical outputs of the AI models.

Questionnaire Structure and Evaluation Criteria

The questionnaire was designed based on four key evaluation criteria:

1. Accuracy

1.1. The model’s response provides a highly accurate and reliable description of the chart.

1.2. A human analyst reviewing the same data is likely to reach a similar conclusion as the model.

1.3. By reading the model’s text, one can easily visualize the chart.

1. Clarity and Readability

2.1. The model has used simple and clear sentences.

The information in the model’s text is presented in a clear, step-by-step, and easy-to-follow manner.

The model has avoided unnecessary repetition of words or sentences.

1. Completeness

All peaks and troughs in the chart are accurately reflected in the model’s analysis.

The model has provided sufficient explanations to support the reported trends.

The model has taken into account external factors or context that may influence data analysis.

4. Creativity and Insightfulness

The model has used analogies or metaphors to make the data easier to understand.

The model has highlighted points that a human analyst might have overlooked during the chart review.

The model has successfully identified and analyzed hidden connections

between different sections of the chart.

Each criterion was assessed through three positively framed declarative statements, where participants provided their responses on a Likert scale ranging from 1 to 5. A score of 5 indicated strong agreement, whereas a score of 1 represented strong disagreement. To minimize response variability due to cognitive bias, the scale was restricted to discrete whole numbers, ensuring consistent and structured feedback.

Each participant evaluated 72 statements related to the AI models. The target population for this study was academic professionals, specifically individuals who held at least a bachelor’s degree in engineering and had either formally studied SD or acquired personal knowledge through independent research. Despite their familiarity with SD, all participants were provided with a brief introduction to the concept of causal relationships, SD models, and the graphical AI-generated outputs to establish a uniform understanding.

To prevent bias in participant responses, the names of the AI models were concealed during the evaluation process.

Ultimately, 14 participants fully and accurately completed the questionnaire, and their responses formed the basis for the final evaluation and ranking of the AI models.