A Additional Details on Interactive Visualization

A.1 Re-Training and Re-Evaluating the Model Using C1

During a what-if analysis, when the training/test data is modified, users can click on the buttons to retrain/reevaluate the model. The curve of the previous trained model will be kept in the model panel for comparison. Fig. 8 shows an example of retraining and reevaluating the model after modifying both the training and test data.

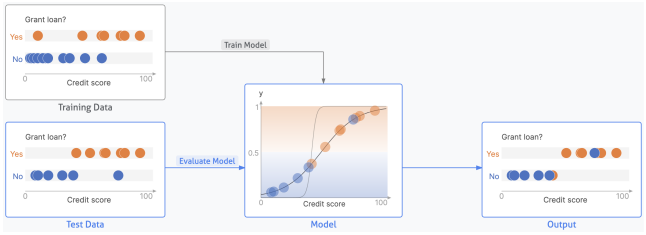


Figure 8: **C1**: updating the training and test data followed by retraining and reevaluating the model.

A.2 Linked Views Between C2 and C3

C2 and **C3** form linked views. Every time users customize the training and test data, the visualizations of the generated data in **C2** and **C3** are updated automatically. The model in **C3** is then retrained and reevaluated, and the changes in the predictions are highlighted in **C3**; see Fig. 9.

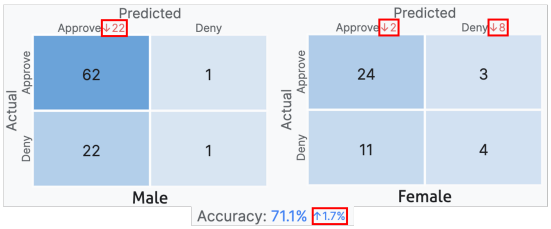


Figure 9: When updating the input data in **C2** and retraining the model, a comparison with the previous model predictions (red boxes) will be displayed in the prediction panel of **C3**.

A.3 In-Processing Bias Mitigation Utilizing C3

In-processing methods use ML models that take fairness into account, typically by adding a fairness term when optimizing the model. We utilize **C3** to visualize the in-processing methods by changing the model in the backend. In addition, we display the changes in the model predictions and accuracy in the prediction panel, compared with the original prediction generated in Sec. 4.2. As shown in Fig. 10, we employ an adversarial debiasing model to regenerate the prediction. This method reverses four males from being approved to being denied, and reverses three females from being denied to being approved, compared with the original logistic regression. However, the prediction accuracy decreases by 2.2%.

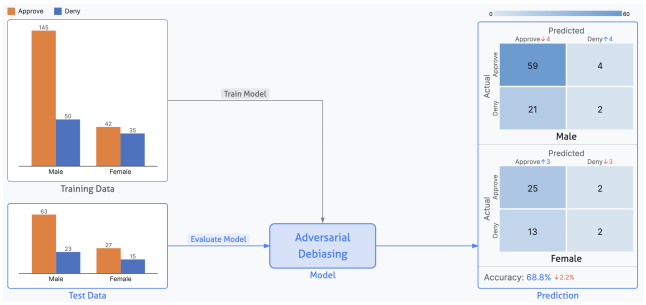


Figure 10: **C3** is used to visualize in-processing debiasing methods.

A.4 Comparing Fairness Metric Values Using C4

To compare the fairness metric values before and after applying a debiasing method, we use **C5** and **C6** together with the fairness metric component **C4**. For example, Fig. 11 shows a metric panel of **C4** that demonstrates the SPD values for predictions generated by the model trained with the original training data and the training data after reweighing. It turns out that the "repaired" data after reweighing results in a SPD value closer to baseline, indicating a mitigation in biases.

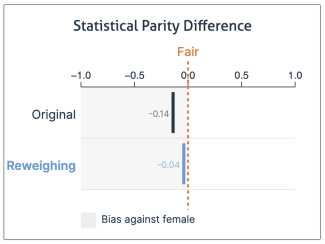


Figure 11: SPD from models trained with and without reweighing.

A.5 Modular Design of Interactive Components

We implemented the interactive components in a modular way for adapting to multiple educational scenarios. Through specified Latex commands, interactive components can be combined with texts and images, and customized for functionalities such as specifying which fairness metrics to display in **C4**. Our educational module then interprets the commands of the Latex file and renders interactive components together with texts and images as a webpage. We utilized the Python library provided by AI Fairness 360 [5] for implementing fairness-related algorithms. We have open-sourced the implementations of these interactive components, available at `https://github.com/tdavislab/FairAI-Education-VisTool.git`.

B Additional User Study Results

We present additional results from our user study.

B.1 Detailed Analysis on Recall Questions

We studied the influence of reading time on the accuracy gain across three conditions. As shown in Fig. 12, the analysis revealed a significant interaction between *TextImg* × *StaticVis* and reading time ($p = 0.0498$), and between *InterVis* × *StaticVis* and reading time ($p = 0.014$). This result implies that *StaticVis* achieves a significantly higher accuracy gain than *TextImg* and *InterVis* when participants spent longer reading time.

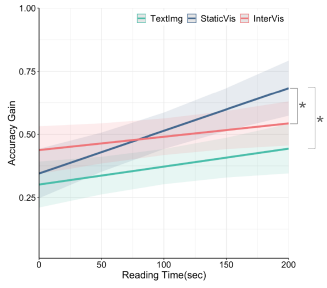


Figure 12: The moderating effect of reading time on the relationship between conditions and accuracy gain. Solid lines represent the mean and the shaded areas represent standard deviations. * indicates a significant difference between conditions ($p < 0.05$).

B.2 Detailed Analysis with Comprehension Questions

We provide below a detailed analysis of the comprehension testing. The comprehension questions were intended to evaluate participants' comprehension of the learning material, which required a deeper

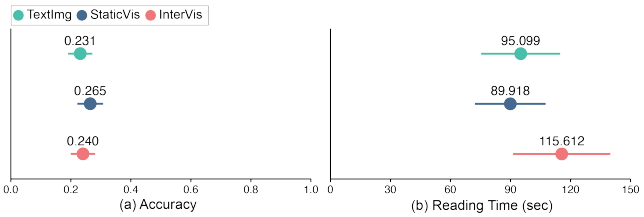


Figure 13: (a) Comprehension test accuracy. (b) Reading time before and during the comprehension test. Error bars show 95% CIs.

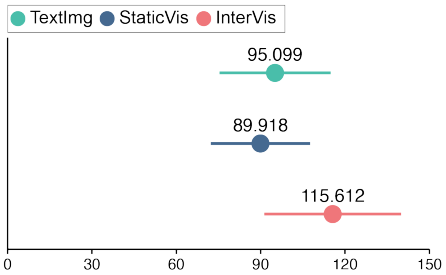


Figure 14: Reading time (in seconds) before and during the comprehension test under three conditions.

understanding than the recall questions. We calculated the accuracy of the comprehension test as the performance metric; see Fig. 13 (a). A one-way ANOVA indicated that there was no significant difference in the mean accuracy across the three conditions ($F(2,379) = 0.660$, $p = 0.518$). We also analyzed the influence of reading time (see Fig. 13 (b)) and visual learning ability on accuracy but found no significant results.

Influence of reading time. We examined the impact of reading time on the relationship between the three conditions and the accuracy of the comprehension questions. In this context, reading time refers to the duration that participants spent on the material before and during the comprehension test. Eight significant outliers were excluded based on the Mahalanobis Distance ($p < 0.001$). Fig. 14 shows the distribution.

A one-way ANOVA test revealed no significant difference among the three conditions ($F(2,372) = 1.259, p = 0.285$). We also ran a linear regression (twice) with two referent conditions, *TextImg* and *StaticVis*, respectively. The result suggests that reading time did not significantly impact the relation between the three conditions and the accuracy.

Influence of visual learning ability. To test whether the visual learning ability influenced the relationship between the three conditions and the accuracy of the comprehension questions, we ran a linear regression model (twice) with *TextImg* and *StaticVis* as referent conditions, respectively. The analysis did not reveal a significant interaction between visual learning ability and accuracy; p-values of condition pairs are: *TextImg* $\times$ *StaticVis*: 0.486; *TextImg* $\times$ *InterVis*: 0.912; *InterVis* $\times$ *StaticVis*: 0.426.

Highlighted results. In summary, we found that neither reading time nor visual learning ability significantly influenced the relationship between the three conditions and the accuracy of comprehension questions.

B.3 Detailed Analysis of Impression Questions

We show in Fig. 15 (a) the participants' ratings of the learning material based on three impression questions using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The questions focused on whether the visualizations were effective (Q1), engaging (Q2), and recommended (Q3). We performed a one-way ANOVA for the rating of each question and found no significant differences across the three conditions, in particular: $F(2,379) = 1.538$, $p = 0.216$, for Q1; $F(2,379) = 0.616$, $p = 0.540$, for Q2; and $F(2,379) = 0.878$, $p = 0.417$ for Q3.

B.4 Detailed Analysis with Response Time

We examined the impact of the response time of recall and comprehension questions on the three dependent variables across all three conditions: accuracy gain from the recall test, accuracy from the comprehension test, and the ratings of the impression questions. The response time was defined as the time participants spent on the recall and comprehension questions, excluding any time spent revisiting the learning material during the tests.

Recall questions. We calculated the response time as the time participants spent on the recall questions' webpage. After removing seven significant outliers through the Mahalanobis distance ($p < 0.001$), we plotted the distribution of response time in Fig. 16 (a). A one-way ANOVA showed that there was no significant difference in response time across three conditions ($F(2,372) = 2.214, p = 0.111$). We then used a linear regression to examine whether response time influenced the relationship between the three conditions and the accuracy gain on the recall test. Our analysis revealed no significant interactions between the response time and the three conditions.

Comprehension questions. We measured the response time as the time spent on the comprehension questions' webpage. We removed eight significant outliers using the Mahalanobis distance ($p < 0.001$), and plotted the distribution of response time in Fig. 16 (b). Again, we conducted a one-way ANOVA and found no significant differences in response time across the three conditions ($F(2,371) = 0.358, p = 0.700$). Furthermore, our linear regression did not find any significant influence of response time on the accuracy of the comprehension test across the three conditions.

Impression questions. We investigated the impact of the total response time of recall and comprehension tests on participants' ratings of the three impression questions. Fig. 16 (c) shows the distribution after removing 12 significant outliers ($p < 0.001$). A one-way ANOVA did not reveal any significant differences in response time across the three conditions ($F(2,367) = 0.742, p = 0.477$). Our linear regression models indicated that response time did not significantly affect participants' ratings of the three impression questions across the three conditions.

Highlighted results. In summary, our analyses suggested that response time did not significantly influence participants' performance on the recall and comprehension tests, as well as their impressions of the learning material.

B.5 Revisiting Time

We examined the influence of the revisiting time, representing the time spent on revisiting learning material during the recall and comprehension tests, on the three dependent variables: accuracy gain on the recall test, accuracy in the comprehension test, and the ratings of the impression questions. We did not find significant outliers in the revisiting time data.

Recall questions. The distribution of the revisiting time during the recall test is shown in Fig. 17 (a). A one-way ANOVA test showed no significant differences in revisiting time across the three conditions ($F(2,379) = 0.276, p = 0.759$). We further used linear regression to investigate whether the revisiting time affected the accuracy gain on the recall test across the three conditions, and we found no significant interactions between the revisiting time and the conditions.

Comprehension questions. Similarly, we conducted a one-way ANOVA on the revisiting time during the comprehension test, see Fig. 17 (b). We found no significant differences across the three conditions ($F(2,379) = 0.735, p = 0.480$). Our linear regression analysis also did not reveal any significant influence of revisiting

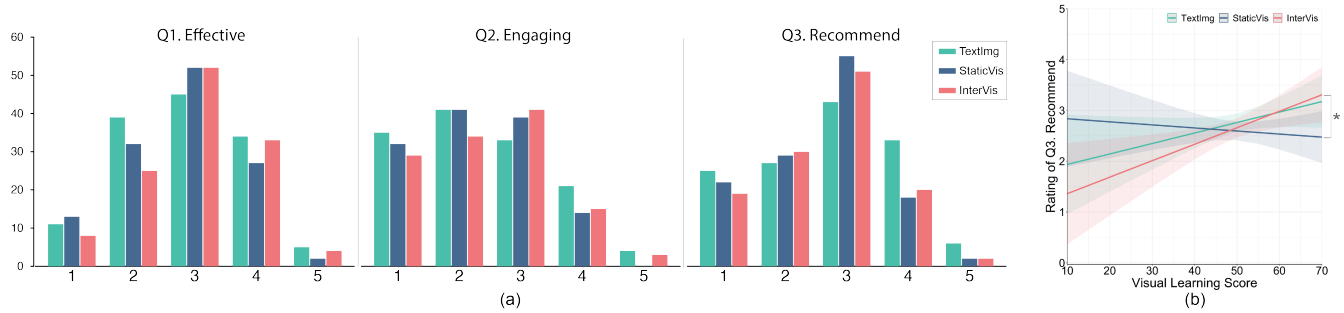


Figure 15: (a) Participants' ratings on three impression questions (1 = strongly disagree, 5 = strongly agree). (b) The moderating effect of visual learning on the relation between conditions and rating of Q3 (Recommend). * indicates a significant difference between conditions ($p < 0.05$).

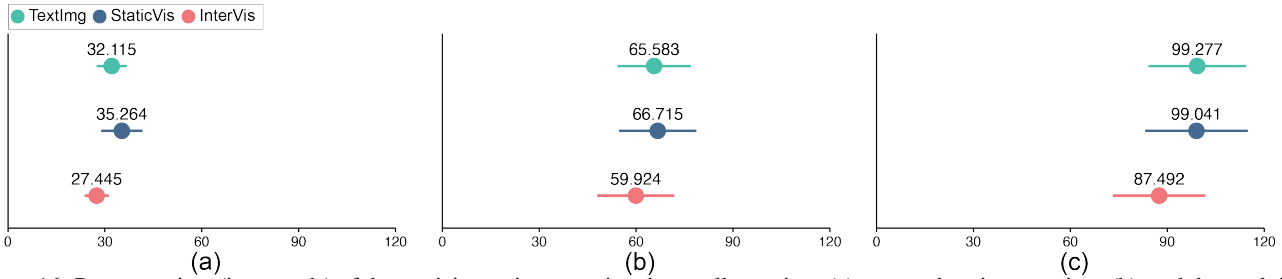


Figure 16: Response time (in seconds) of the participants in answering the recall questions (a), comprehension questions (b), and the total time for answering both sets of questions (c).

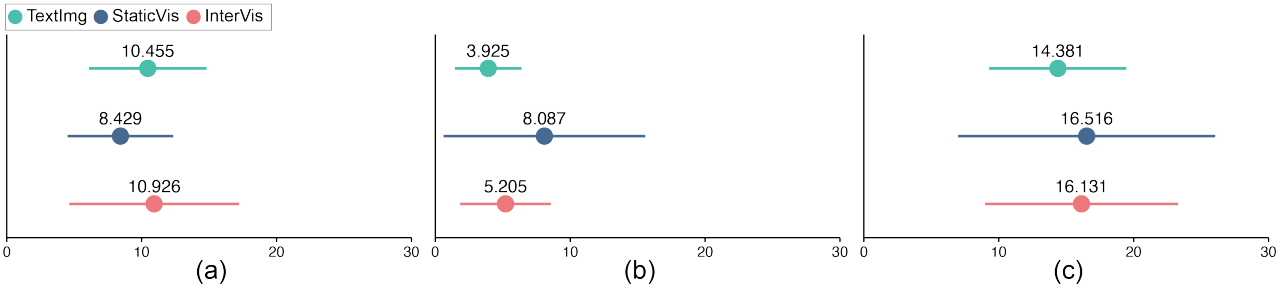


Figure 17: Revisiting time (in seconds) during the recall test (a), the comprehension test (b), and the total revisiting time for both tests (c).

time on the accuracy of the comprehension test across the three conditions.

Impression questions. We analyzed the impact of the total revisiting time of recall and comprehension tests on participants' ratings of three impression questions; see Fig. 17 (c). A one-way ANOVA revealed no significant differences in the revisiting time across the three conditions ($F(2,379) = 0.930, p = 0.911$). A linear regression showed that revisiting time did not significantly affect participants' ratings on the three impression questions across the three conditions.

Highlighted results. In conclusion, the time that participants spent on revisiting the learning material during the recall and comprehension tests did not significantly influence the performance of participants on the recall and comprehension questions, nor their impression of the learning material among three conditions.

C Questions

We present screenshots of all questions used in the user study, which were shared by all three conditions. The background questionnaire is shown in Fig. 18.

Questions from the pre-test and the recall test are shown in Fig. 19 and Fig. 20, respectively. The only distinction between them is that the recall test includes a button enabling users to revisit the learning material while responding to the questions.

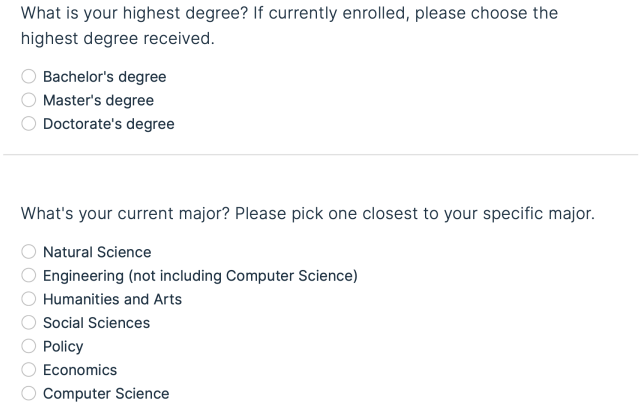


Figure 18: Multiple-choice questions for the background.

The comprehension questions are shown in Fig. 21, and the impression questions are included in Fig. 22. Finally, Fig. 23 and Fig. 24 describe the 22 questions in the learning style questionnaire. To determine participants' visual learning scores, we utilized the 11 questions highlighted in orange, including seven perceptive

Disparate Impact for measuring the fairness of a Machine Learning model (ML) is calculated based on the prediction of the ML model. Do you agree with this statement?

- Agree
- Disagree
- I have no idea

For the fairness metric, Disparate Impact, which of the following values means a higher level of fairness?

- 0.8
- 0.3
- I have no idea

If the fairness of the ML model increase, the accuracy of the ML model decreases. Do you agree with this statement?

- Agree
- Disagree
- I have no idea

Figure 19: Multiple-choice questions from the pre-test.

questions and four imaginative questions. We did not highlight any questions during the user study.

Note: When answering the following questions, you can return the learning material by clicking here.

Disparate Impact for measuring the fairness of a Machine Learning model (ML) is calculated based on the prediction of the ML model. Do you agree with this statement?

- Agree
- Disagree
- I have no idea

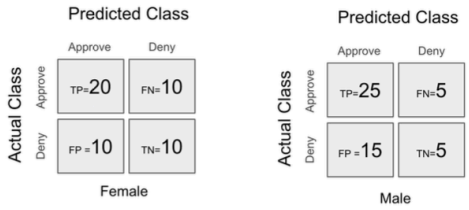
For the fairness metric, Disparate Impact, which of the following values means a higher level of fairness?

- 0.8
- 0.3
- I have no idea

Figure 20: Multiple-choice questions from the recall test.

Note: When answering the following questions, you can return the learning material by clicking here.

Suppose the following two confusion matrices are the model prediction on a loan application for the male and female groups, respectively. Suppose the male group is the privileged group. Please answer the following 2 Questions.



How will the value of **Disparate Impact** change if we moved 5 items from False Negative (**FN**) to True Negative (**TN**) for the female group? That is, will **Disparate Impact** increase, decrease or not change if False Negative (**FN**) became 5 and True Negative (**TN**) became 15 in the female group. Please indicate your choice from the options provided below.

- Increase
- Decrease
- No change

To calculate **Disparate Impact** (DI), do we need to know the ground truth (the actual outcome of data)?

- Yes
- No

How will the value of **Disparate Impact** change if we moved 5 items from False Negative (**FN**) to True Positive (**TP**) for the male group? That is, will **Disparate Impact** increase, decrease or not change if False Negative (**FN**) became 0 and True Positive (**TP**) became 30 in the male group. Please indicate your choice from the options provided below.

- Increase
- Decrease
- No change

Figure 21: Multiple-choice questions from the comprehension test.

How effective was the material shown to you in illustrating the concept? Please rate it on a rating scale of 1-5.

Not at all effective (1)	Not effective (2)	Neutral (3)	Effective (4)	Very effective (5)
○	○	○	○	○

How engaging did you find the information shown? Please rate it on a rating scale of 1-5.

Not at all engaging (1)	Not engaging (2)	Neutral (3)	Engaging (4)	Very engaging (5)
○	○	○	○	○

Would you recommend this website to others to learn about this topic? Please rate it on a rating scale of 1-5.

Not at all recommend (1)	Not recommend (2)	Neutral (3)	recommend (4)	Very recommend (5)
○	○	○	○	○

Figure 22: Impression questions that test participants' impression of the learning material.

At the **end** of this study, please rate the following statements on a scale of 1-6. There are no wrong or right answers. (1, strongly disagree; 2, moderately disagree; 3, disagree a little; 4, agree a little; 5, moderately agree; 6, strongly agree)

Most of the time, I ...	strongly disagree	moderately disagree	disagree a little	agree a little	moderately agree	strongly agree
...prefer to study alone.	○1	○2	○3	○4	○5	○6
...enjoy competing.	○1	○2	○3	○4	○5	○6
...create a mental picture of what I study.	○1	○2	○3	○4	○5	○6
...prefer to study with other students.	○1	○2	○3	○4	○5	○6
...compete to get the highest grade.	○1	○2	○3	○4	○5	○6
...create a mental picture of what I see.	○1	○2	○3	○4	○5	○6
...learn better when someone represents information in a pictorial (e.g., picture, flowchart) way.	○1	○2	○3	○4	○5	○6
...learn practical tasks better than theoretical ones.	○1	○2	○3	○4	○5	○6
...learn better when I study with other students.	○1	○2	○3	○4	○5	○6
...compete with other students.	○1	○2	○3	○4	○5	○6
...create a mental picture of what I read.	○1	○2	○3	○4	○5	○6
...learn better when someone uses visual aids (e.g., whiteboard, power point) to represent a subject.	○1	○2	○3	○4	○5	○6
...learn better when I am involved in a task.	○1	○2	○3	○4	○5	○6
...focus more on the details of a subject.	○1	○2	○3	○4	○5	○6
...consider the details of a subject more than its whole.	○1	○2	○3	○4	○5	○6

Figure 23: Questions from the learning style questionnaire: first page.

...learn better when I watch an educational program.	○1	○2	○3	○4	○5	○6
...learn better when I watch a demonstration.	○1	○2	○3	○4	○5	○6
...create a mental picture of what I hear.	○1	○2	○3	○4	○5	○6
...remember the details of a subject.	○1	○2	○3	○4	○5	○6
...learn better when I study alone.	○1	○2	○3	○4	○5	○6
...remember specific details of subjects.	○1	○2	○3	○4	○5	○6
...learn better when studying practical, job-related, subjects.	○1	○2	○3	○4	○5	○6

Figure 24: Questions from the learning style questionnaire: second page.