

Background

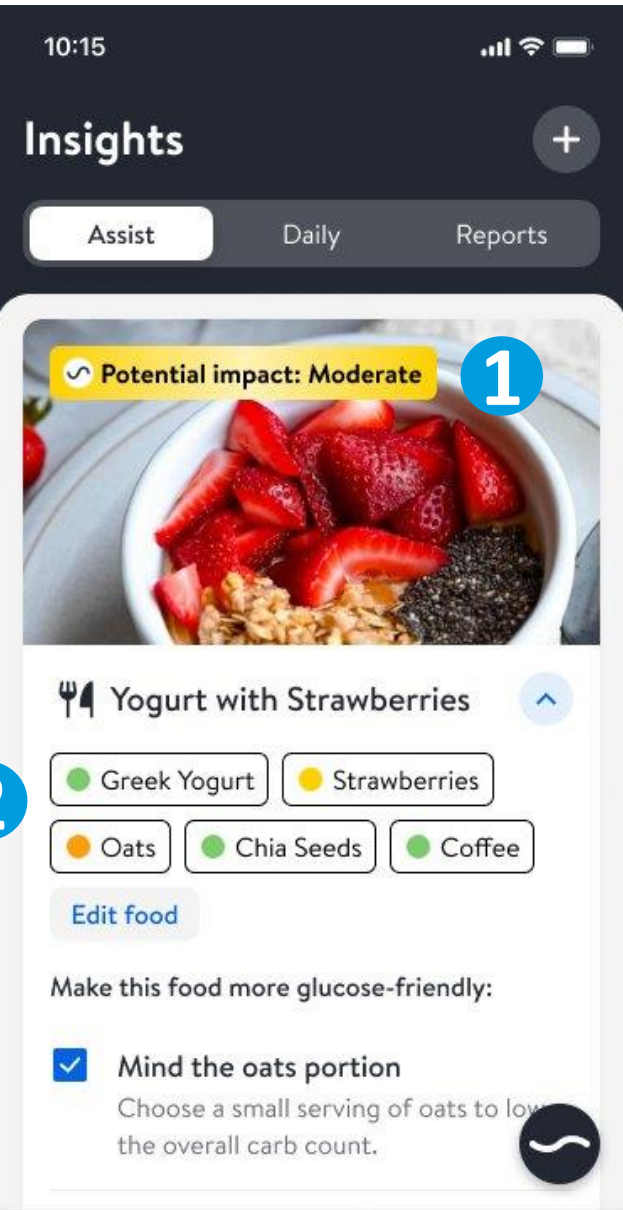


Figure 1: Screenshot of the Libre app's 'Insights' feature showing meal analysis results.

LibreAssist is a generative AI feature in the Libre app that provides personalized meal guidance based on user-input images or text descriptions. In this study, we evaluate the model’s performance at analyzing:

- 1. Glycemic Impact Score:** Predicting the glycemic impact of the overall meal, classified as minor, moderate or major.
- 2. Component Glycemic Impact Score:** Estimating the glycemic impact of each meal component, classified as minor (green), moderate (yellow) or major (orange).

Given that there are about 40.1 million people living with diabetes in the United States¹ and approximately 19,200 Certified Diabetes Care & Education Specialists (CDCES)², or roughly 1 CDCES for every 2000 individuals with diabetes, LibreAssist offers potential to enhance diabetes care by providing timely, personalized meal guidance that complements CDCES-delivered education at scale.

Methods

Data Collection & Annotation

- A total of 1971 images and 400 text descriptions were collected during a pilot study. Additional images were supplemented to ensure the dataset was representative of the typical American diet, based on the top 20 food groups in the NHANES “What We Eat in America” study³. The dataset was split into a working dataset (983 images, 200 text inputs) and validation dataset (988 images, 200 text inputs) to ensure robustness to new data.
- 34 experts with CDCES and Registered Dietician (RD) certifications were selected to annotate the meals. 5 experts were randomly assigned to each image or text description.

Overall Glycemic Impact Score

- Model performance on classifying the overall glycemic impact of the meal was evaluated by comparing the average percent agreement between the model and annotators to the average percent agreement among the annotators. Agreement reflected the percentage of other annotators whose chosen label (minor,

moderate, major) matched the model or individual annotator’s label. Statistical non-inferiority was tested using a 10% margin for images and 20% for text.

- To calculate agreement rate, we used a leave-one-out method, calculated by removing one annotator and comparing both the model and held-out annotator to the remaining four annotators. This process is repeated for all five annotators, producing an average agreement rate for the model and annotators
- Model accuracy metrics were also computed, including precision and recall, treating the majority vote of annotators as truth labels.
- Additionally, the proportion of images on which the model-selected glycemic impact matched at least one annotator was calculated, useful for identifying the most significant errors.
- To account for the the stochastic nature of GenAI models, variability on overall glycemic impact score was evaluated in two ways. First, Krippendorff’s Alpha was calculated to compare inter-annotator variability (between the 5 annotators) to inter-model

Results

- On overall glycemic impact score, model agreement with annotators was non-inferior to agreement among annotators using both image and text inputs (Figure 1, Table 1), demonstrating the model’s ability to provide meal analysis comparable to expert CDCES/RDs.
- Model accuracy versus the majority vote of annotators was strong on images and text inputs (Table 2). However, substantial inter-annotator variability was observed, making majority vote accuracy a less precise measure of model performance compared to leave-one-out agreement rate.
- The model agreed with at least one other annotator on the overall glycemic impact score on 94.8% of images and 96.0% of text inputs (Table 2).
- Model to annotator agreement on component-level

glycemic impact scores was non-inferior to the agreement rate among annotators, demonstrating that the model can provide reliable feedback on individual food items.

- Inter-annotator Krippendorff’s Alpha (between the 5 annotators) was 54.9%, while the inter-model Krippendorff’s Alpha (between 5 repeated model runs) was 92.8%, suggesting that multiple runs of the model yield less variability than annotations performed by different experts.
- Run-to-run variability in glycemic impact agreement rate was low. Ten unique runs of the model yielded differences in agreement rates ranging 0.6% with an IQR of 0.3%. On each run, the model was non-inferior to the annotators. The narrow range indicates strong repeatability of model performance given a large sample size (Figure 2).

Figure 2: Distribution of Difference in Agreement Rate on Repeated Runs Figure shows the distribution of the difference (model-annotators) in agreement rate on overall glycemic impact score over 10 repeat runs of the model on the same set of images (N=1000).

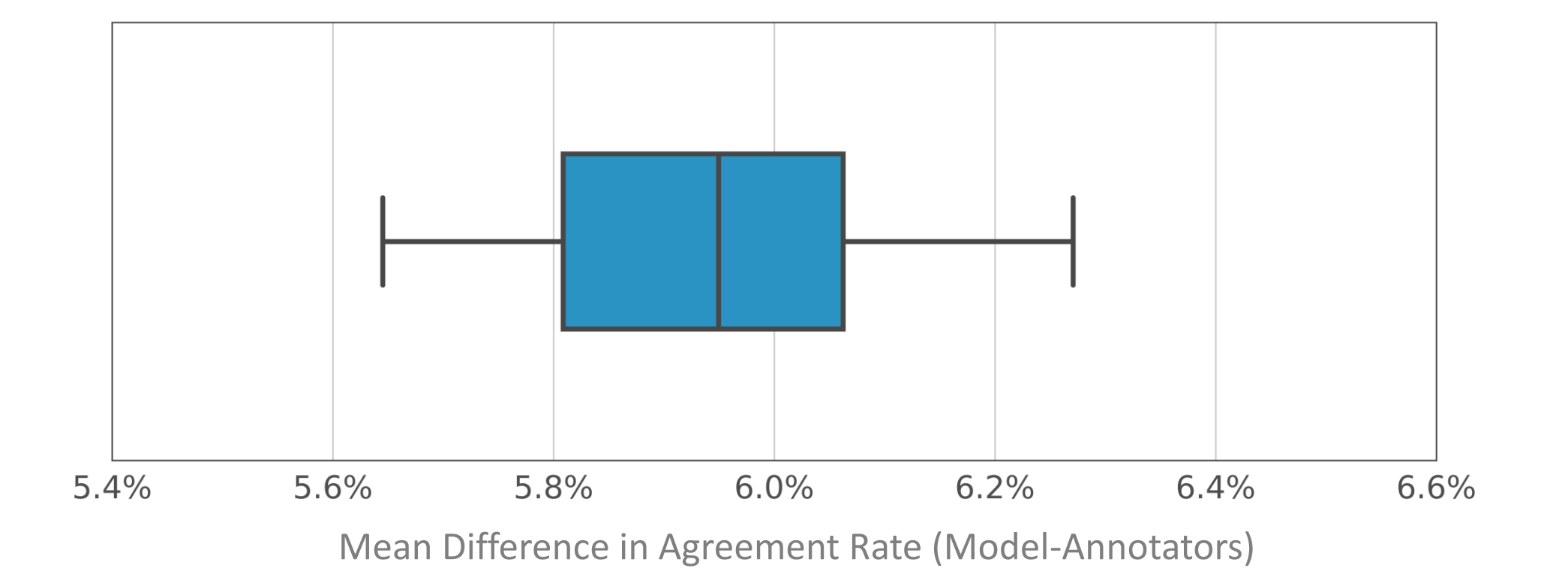


Table 1: Agreement Rate on Overall Glycemic Impact Score Table shows the mean agreement rate among annotators and between the model and annotators on overall glycemic impact score, with 95% confidence intervals.

Input set	Mean (Among Annotators)	Mean (Model vs. Anno)	Difference (Model - Annotators)	CI Lower	CI Upper	N
Images	58.6%	62.9%	4.3% ^a	2.5%	6.0%	987
Text	63.9%	68.1%	4.2% ^a	0.5%	7.6%	200

Table 2: Majority Vote Accuracy & Agreement with at Least 1 Annotator Table shows the rate of agreement with at least one annotator and model accuracy, precision and recall relative to the majority vote of annotators on overall glycemic impact score.

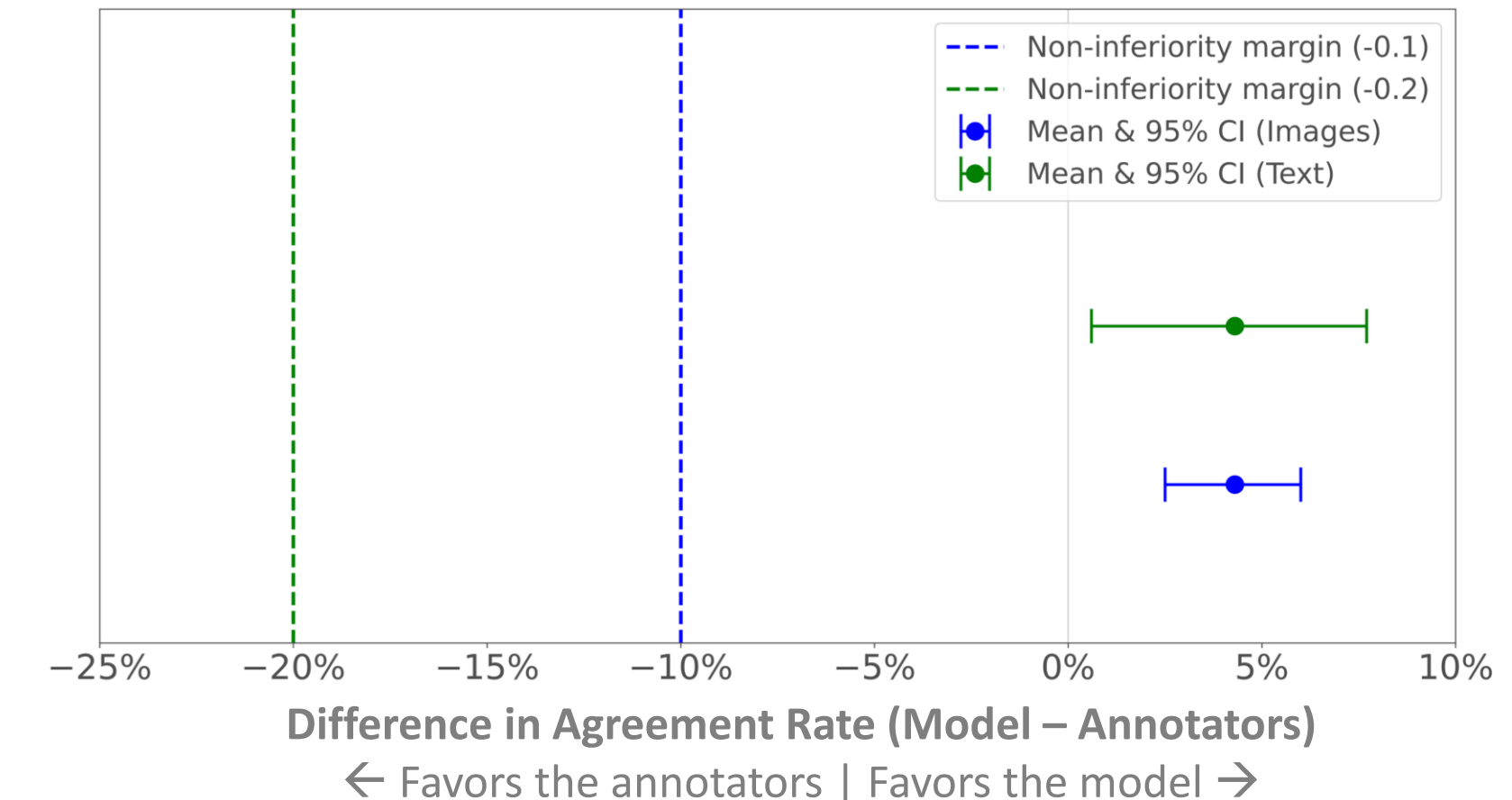
Input Set	Agree ≥ 1 annotator	Majority-vote accuracy	Majority-vote precision	Majority-vote recall
Images	94.8%	74.5%	75.2%	74.7%
Text	96%	79%	81.2%	80%

Table 3: Agreement Rate on Component-level Glycemic Impact Score Table shows the mean agreement rate on component-level glycemic impact scores for the annotators and model on components selected by at least 3 annotators.

Input set	Mean (Among Annotators)	Mean (Model vs. Annotators)	Mean Difference (Model - Annotators)	CI Lower	CI Upper	N
Images	69.3%	69.8%	0.5% ^a	-1.1%	2.1%	743
Text	74.6%	74.0%	-0.6% ^a	-4.4%	2.8%	140

^a Indicates statistically non-inferiority using bootstrapped confidence intervals at α=0.05, non-inferiority margin of -10% for images and -20% for text.

Figure 1: Difference in Agreement Rate on Glycemic Impact Score. Mean and 95% confidence interval on the difference in model agreement rate with annotators versus agreement among annotators is shown, with a 10% non-inferiority margin. Lower bound CI does not cross the margin, showing non-inferiority.



Conclusions

- The LibreAssist meal inference model performed comparably to experts on both meal analysis tasks.
- These findings demonstrate that LibreAssist provides glycemic impact predictions comparable to expert dietitians, supporting its role in enhancing CDCES care by providing real-time individualized meal feedback at scale.
- One limitation is that the study was intended for a target population of users in the U.S., so non-western cuisines are underrepresented in the dataset. Further study is needed before applying the model to cuisines outside the U.S.

References

1. Centers for Disease Control and Prevention, National Diabetes Statistics Report, c2026 [cited 2026 Feb 5], 2. Certification Board for Diabetes Care and Education, Count by State of Health Professionals Holding the Certified Diabetes Care and Education Specialist Certification, c2024 [cited 2026 Feb 5], 3. Dwyer J, Picciano MF, Raiten DJ; National Health and Nutrition Examination Survey. Estimation of usual intakes: What We Eat in America-NHANES. J Nutr. 2003 Feb;133(2):609S-23S. doi: 10.1093/jn/133.2.609S.

Acknowledgements

This study was funded by Abbott