ABBOTT DIABETES CARE

SARAH VAUGHAN, M.S., SHREYA GUPTA, M.S., JEFFREY NISHIDA-BOUCHER, M.S., BYRON OLSON, PH.D.

Evaluation of a GenAI Glycemic Impact Prediction System Relative to Expert Dieticians

23 | April | 2026

Disclosures

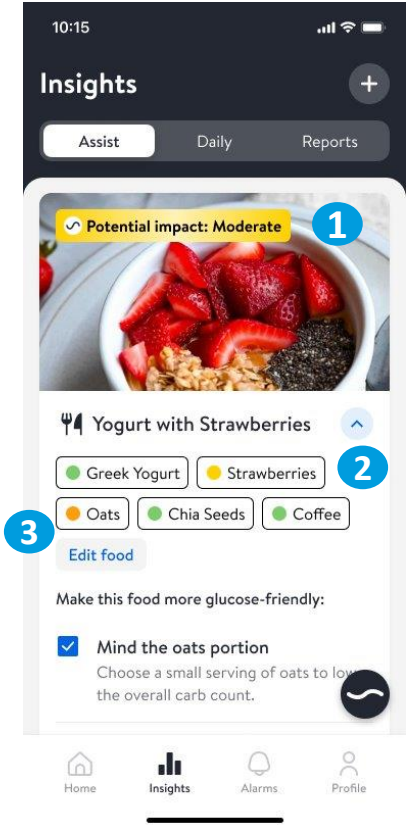
This study was funded by Abbott.

Sarah Vaughan, M.S., Shreya Gupta, M.S., Jeffrey Nishida-Boucher, M.S., Byron Olson, Ph.D. are full time employees of Abbott Diabetes Care.

Proprietary and confidential — do not distribute

BACKGROUND

LibreAssist Meal Inference Feature



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 3 aspects of a meal:

1. **Glycemic Impact Score:** Predicting the glycemic impact of the overall meal, classified as minor, moderate or major.
2. **Meal Components:** Identifying the individual components of the meal, including foods that could be easily removed, swapped out, or skipped.
3. **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 States[1] and approximately 19,200 Certified Diabetes Care and 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” study[3].



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.



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.

RESULTS

Agreement Rate on Overall Glycemic Impact

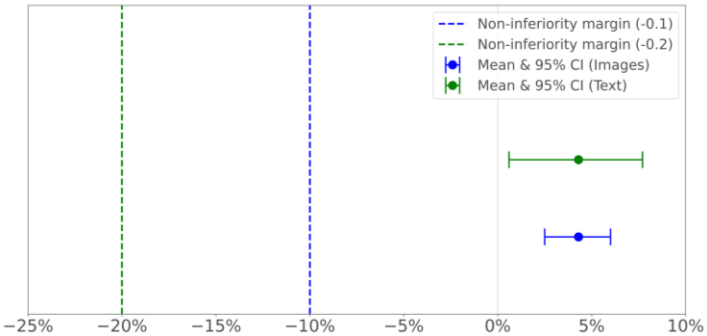
- Model performance on classifying the overall glycemic impact was evaluated by comparing the average % agreement between the model and annotators to the average % agreement among the annotators. Agreement reflected the % of other annotators whose chosen label (minor, moderate, major) matched the model or individual annotator's label.
- 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 agreement with annotators was non-inferior to agreement among annotators on both images and text inputs (Figure 1, Table 1), indicating the model's ability to provide meal analysis comparable to expert CDCES/RDs.

Table 1: Agreement Rate on Overall Glycemic Impact Score

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

[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



RESULTS

Majority Vote & Agreement with ≥1 Annotator

- Model accuracy metrics, including precision and recall, were also calculated, treating the majority vote of all 5 annotators as the truth labels. Model accuracy versus the majority vote of the annotators was 74.5% for images (precision: 75.2%, recall: 74.7%), and 79% for text inputs (precision: 81.2%, recall: 80%) as shown in Table 2. No systematic bias were seen across the three labels.
- However, substantial inter-annotator variability was observed, making majority vote accuracy a less precise measure of model performance compared to agreement rate between the model and annotators.
- 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).

Table 2: Majority Vote Accuracy & Agreement with at Least 1 Annotator 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%

Proprietary and confidential — do not distribute

RESULTS

Identification of Individual Meal Components

Diverse terminology may be used to describe the same food (ex. “chicken nugget”, “fried chicken”, and “breaded chicken” refer to conceptually equivalent components). All component strings were normalized by querying the FatSecret nutritional database[4].

Component identification was evaluated in 2 ways:

- % of components selected by all other annotators that were selected by the model, compared to the corresponding % for each annotator.
- % of components selected by the model that were not selected by ≥ 1 annotator versus the % of annotator-selected components that were not identified by ≥ 1 other annotator.

Model performance at identifying meal components was non-inferior to the annotators on both metrics (Table 3). The model identified components that experts typically agree on, without introducing extraneous ingredients, achieving a suitable balance between sensitivity and specificity.

Table 3: Component Identification Performance

Input Set	Metric description	Mean (Annotators)	Mean (Model)	Difference (Model - Annotators)	CI Lower	CI Upper	N
Images	% components selected out of in-consensus components[b]	75.8%	88.5%	12.7%[a]	10.0%	15.6%	904
Images	% components that no other annotator selected	11.3%	11.2%	-0.1%[a]	-1.3%	1.3%	986
Text	% components selected out of in-consensus components[b]	80.6%	93.3%	12.7%[a]	7.6%	18.3%	157
Text	% components that no other annotator selected	8.9%	7.6%	-1.4%[a]	-3.7%	1.1%	171

[a] Indicates statistically non-inferiority using bootstrapped confidence intervals at α=0.05, non-inferiority margin of -10% for images and -20% for text.
[b] In-consensus components refers to the set of components selected by at least 4 other annotators.

RESULTS

Component-level Glycemic Impact

- Component-level glycemic impact prediction was also evaluated by comparing percent agreement between the model and annotators to the percent agreement among annotators.
- Because annotators do not always identify the same components, agreement was computed only for components identified by the model and at least 3 annotators. Agreement rates were averaged over the in-consensus components for each image or text input.
- Model to annotator agreement on component-level glycemic impact scores was non-inferior to the agreement rate among annotators on images (difference: 0.5%, 95% CI: -1.1%, 2.1%) and text (difference: -0.6%, 95% CI: -4.4%, 2.8), demonstrating that the model can provide reliable feedback on individual food items.

Table 4: Agreement Rate on Component-level Glycemic Impact Score

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.

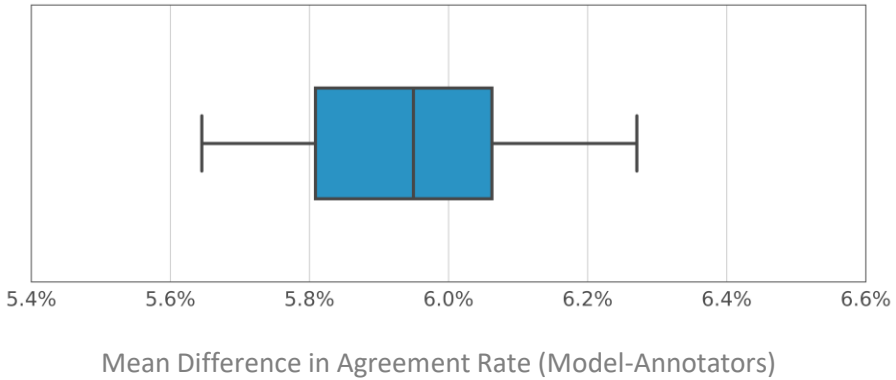
RESULTS

Model Variability

To account for the the stochastic nature of LLMs, model variability on overall glycemic impact score was evaluated in two ways.

- Inter-annotator Krippendorf's Alpha (between the 5 annotators) was 54.9%, while the inter-model Krippendorf'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).



Conclusions

- The LibreAssist meal inference model performed comparably to experts across all three meal analysis tasks.
- On overall glycemic impact, model agreement with annotators was non-inferior to agreement among annotators on both image and text inputs.
- For meal-component identification, the model identified components that experts typically agree on, without introducing extraneous ingredients, achieving a suitable balance between sensitivity and specificity.
- 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.
- 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. **4**. FatSecret Platform API for Food and Nutrition Data, FatSecret, c2025 [cited 2025 Dec 9], Available from: https://platform.fatsecret.com/



Abbott