Can ChatGPT Judge Beef Cattle? Baseline AI Performance in Livestock Evaluation

Dr. Audie L. Cherry
Assistant Professor of Instruction
Texas State University
audiec@txstate.edu

Dr. Jesse Backstrom
Assistant Professor
Texas State University
jbackstrom@txstate.edu

Can ChatGPT Judge Beef Cattle? Baseline AI Performance in Livestock Evaluation

Introduction

Agriculture is well positioned to benefit from AI-driven innovation that can transform how food, fiber, and natural resources are produced and marketed. A growing body of research demonstrates potential within animal production systems, including live animal applications (Edwards-Callaway et al., 2025; Lopez et al., 2025; Wang et al., 2025). However, less attention has been devoted to whether AI can either replicate or improve upon the visual and analytical decision-making processes used by human livestock evaluators. Livestock evaluators assess structural correctness, muscling, balance, body capacity, and overall production merit based largely on visual appraisal. As such, livestock evaluation provides a unique setting for examining whether AI systems can perform tasks traditionally reserved for trained human experts.

Livestock Evaluation is among the most prominent Career Development Events (CDE) in both FFA and 4-H contests. These evaluations routinely require relating animal form to function and predicting future performance based on observable phenotypic traits, as well as placing/ranking a group (usually four) of similar animals based on predicted performance (Chrestman & Bryan, 2020). The literature presents mixed findings on the livestock evaluation professional development and training needs of School-Based Agricultural Education (SBAE) teachers (Dreher & Robinson; 2024; McGuire, 2022). Additionally, SBAE teachers are more likely to use online resources for CDE preparation (Harris, 2008).

The purpose of this exploratory study was to evaluate the baseline performance of a large language model (LLM)/multimodal AI (i.e., ChatGPT version GPT-5.5 Thinking; OpenAI, 2026) to evaluate beef cattle. This study was part of a larger study and guided by the following objectives: 1) Determine the accuracy of ChatGPT in evaluating beef cattle through baseline multimodal evaluation without corrective feedback by comparing its results to expertly placed classes and 2) Explore the instructional implications of ChatGPT's baseline beef evaluation performance for SBAE livestock evaluation instruction.

Theoretical Framework

This study was examined through the lens of Expertise Theory (Chase & Simon, 1973; Ericsson et al., 1993), which suggests that expertise is developed through extensive experience, repeated exposure, and deliberate practice, which allows them to recognize complex patterns and make rapid, accurate judgements Livestock evaluation represents an example of expertise development. Successful evaluators learn to identify and integrate numerous visual indicators related to muscling, volume, structural correctness, balance, and overall production merit. Through repeated exposure to animals and judging scenarios, evaluators develop the ability to quickly recognize patterns and make comparative assessments that often distinguish experts from novices. Because modern AI systems are trained on vast amounts of information and can process visual and textual inputs, they provide a unique opportunity to examine whether machine-based systems can approximate expert pattern-recognition processes traditionally developed through years of human experience. This study also drew conceptual inspiration from Lee et al. (2026), who explored generative AI performance in an educational assessment context.

Methods

Expertly placed beef judging classes (four animals per class) were used as the standard against which ChatGPT Plus evaluations were compared. All classes came from Sure Champ

(n.d.). an online training tool that uses nationally recognized judges to establish official placings. Our sample included all beef classes in the Sure Champ repository (n = 36), excluding two classes that incorporated Expected Progeny Differences because this study focused on phenotypic traits. Although appropriate for this exploratory study, the resulting sample may limit generalizability. After a constructed prompt with evaluation instruction, images of each class, along with the class designation (e.g., Simmental Bulls), were uploaded directly to ChatGPT Plus for evaluation. Procedural reliability was supported through standardized prompts, consistent image-uploading procedures, and instructions preventing ChatGPT from cross referencing images. ChatGPT was treated like a CDE participant, and its placings were scored and compared to those of the expertly (human) placed classes (just as in a CDE) on Sure Champ's website. LivestockJudging.com (n.d.) was used as the standard for scoring the ChatGPT placings.

Results

ChatGPT placed the classes (n = 36) and provided cuts. Scores were given to each placed class based on the human expert class placing. A participant score of 50 is considered a perfect class placing. The range of ChatGPT scores was 15–48. The overall mean score of ChatGPT's placings was 33.89 (SD = 7.75). The 95% confidence interval was 31.27–36.51.

Ten classes met the 80% practical benchmark (40 points or higher) used to interpret baseline evaluation performance. Of the 36 classes evaluated by ChatGPT, only two had the same top and bottom pairs (the top two animals are the same, even if in the wrong order) as Sure Champ (n.d.). Further, the cuts given by ChatGPT showed a consistent pattern of 2-2-3 (n = 24; 66.67%), regardless of class placing. No cuts given by Sure Champ were 2-2-3. Grouping by breed or sex did not have the statistical power to provide reliable correlational data.

Conclusions/Implications/Recommendations

One component of this study design was, though prompted to examine resources about livestock evaluation beforehand, ChatGPT did not receive corrective feedback during evaluation, limiting its ability to adjust its judging approach (Objective 1). Further, we provided ChatGPT with only one image per class, which likely weakened its capabilities. However, this design was appropriate because trainees also evaluated only a single image for each class. Moreover, this study represented the initial stage of a larger research effort, and developing a larger sample of live animal images was not feasible. While these are limitations, they reflect practical constraints necessary to maintain the study's integrity as a baseline performance evaluation. We did not find ChatGPT to be accurate in evaluating cattle classes without corrective feedback (i.e., baseline). Further, when examining Objective 2, we found that there is *potential* for ChatGPT and similar AIs to enhance livestock evaluation instruction, but, as it relates to this study, further refinement is needed before it could be considered an expert for beef evaluation, especially compared to the expertise of a human whose expertise developed through experience, exposure, and practice.

We suggest ChatGPT and similar AI platforms may improve with corrective feedback and refinement, which may make it a valuable resource for SBAE teachers to enhance their CDE prep and classroom instruction. These findings may also inform future AI-assisted livestock evaluation research beyond the classroom. Therefore, we recommend future studies compare various types of prompting (e.g., baseline, rubric-guided, corrective feedback, etc.) to determine whether performance gains are attributable to feedback, prompt structure, or model capability. It is recommended and reiterated that researchers, practitioners, SBAE teachers, and others in the livestock industry recognize the current limitations of AI before over-reliance on the technology.

References

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Chrestman, G., & Bryan, K. A. (2020). *4-H Livestock Judging Manual*. Mississippi State University Extension. https://extension.msstate.edu/sites/default/files/publications/P2289_web.pdf

Dreher, C., & Robinson, J. S. (2024). School-based agricultural education teachers' interest in and confidence for preparing students to compete in career development events. *Journal of Career and Technical Education, 39*(1). https://doi.org/10.21061/jcte.473

Edwards-Callaway, L., Loh, H.Y., Kautzky, C., & Sullivan, P. (2025). A comparison of artificial intelligence and human observation in the assessment of cattle handling and slaughter. *Animals*, 15, 1325. https://doi.org/10.3390/ani15091325

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. https://psycnet.apa.org/doi/10.1037/0033-295X.100.3.363

Harris, C. R. (2008). Career development event participation and professional development needs of Kansas agricultural education teachers. *Journal of Agricultural Education, 49*(2), 130-138. https://doi.org/10.5032/jae.2008.02130

Lee, D. Y. H., Parker, A. J., Norbury, C. F., & Shanks, D. R. (2026). Validating AI-assisted evaluation of open science practices in brain sciences: ChatGPT, Claude, and human expert comparisons. *Royal Society Open Science, 13*(2), 250381. https://doi.org/10.1098/rsos.250381

Livestockjudging.com. (n.d.). *Judging score calculator*. Retrieved May 27, 2026, from https://www.livestockjudging.com/scorecalculator/

Lopez, A., Garcia, V., Valles, D., & Drewery, M. L. (2025). 149 Developing and validating artificial intelligence models to predict cattle behavior, movement, and emotion. *Journal of Animal Science*, 103, Issue Supplement_2, 81. https://doi.org/10.1093/jas/skaf170.094

McGuire, M. R. (2022). *Secondary preservice agricultural education teachers' professional knowledge bases & collective pedagogical content knowledge* [Master's Thesis]. Purdue University. Available from ProQuest One Academic. (2838333458). https://www.proquest.com/dissertations-theses/secondary-preservice-agriculture-education/docview/2838333458/se-2

OpenAI. (2026). *ChatGPT* (GPT-5.5 Thinking) [LLM]. https://chatgpt.com

Sure Champ. (n.d.). *Judging archives*. BioZyme, Inc. Retrieved May 27, 2026, from https://surechamp.com/category/judging/

Wang, S., Jiang, H., Qiao, Y., Jiang, S., Lin, H., & Sun, Q. (2022). The research progress of vision-based artificial intelligence in smart pig farming. *Sensors, 22*(17), 6541. https://doi.org/10.3390/s22176541