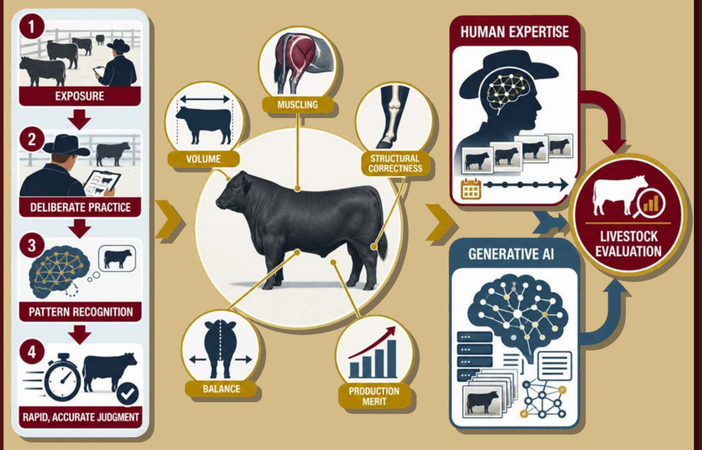
Can ChatGPT Judge Beef Cattle? Baseline AI Performance in Livestock Evaluation

Audie L. Cherry, Jesse D. Backstrom
Texas State University, Department of Agricultural Sciences

Introduction

- The ag industry is well positioned to benefit from AI, with increased research and innovation occurring in animal systems (Edwards-Callaway et al., 2025; Lopez et al., 2025; Wang et al., 2025)
- Livestock Evaluation is also a prominent Career Development Event in 4-H and FFA (Chrestman & Bryan, 2020)
- Mixed reports of professional development and training needs for School-Based Agricultural Education (SBAE) teachers (Dreher & Robinson; 2024; McGuire, 2022)
- Teachers likely to use online resources to learn and teach in livestock evaluation (Harris, 2008)

Theoretical Framework



Expertise Theory (Chase & Simon, 1973; Ericsson et al., 1993)
Conceptual Inspirations from Lee's (2026) AI in Education Research

Purpose and Objectives

Purpose:
Evaluate the baseline performance of a large language model (LLM)/multimodal AI (i.e., ChatGPT version GPT-5.5 Thinking; OpenAI, 2026) when judging beef cattle.

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.
2. Characterize the implications of ChatGPT's baseline beef evaluation performance for SBAE livestock evaluation instruction.

Methods

I will be using this chat to upload sets of animals for livestock judging. Each time, I want you to: 1. Judge the class, putting the animals in order with cuts. Give Brief Reasons. 2. Keep a running table of your judging classes with columns for species, breed, gender class, placements with cuts, and any other pertinent information. 3. Do NOT cross reference any images or text to any websites. I am seeing how you place them compared to others, so I do not want to get skewed data.

PROMPT

ChatGPT 5.5-Thinking Plus

Red angus heifers

- Non-Experimental, Within-Subject Design
- *n* = 36
- No Corrective feedback
- Second Prompt after four rounds

Class 1 – Red Angus Heifers

Placing: **2–3–4–1**
Cuts: **3–2–4**

SURE•CHAMP

Brief Reasons:

Top Pair (2 over 3):
I placed 2 over 3 as she is the more complete, balanced heifer that combines femininity with structural correctness. She is cleaner through her front third, smoother in her shoulder, and more extended in her neck. Additionally, she appears more level from

Did you use data to "train yourself" in livestock evaluation?

Short answer: **not in the way you're thinking.**

I wasn't trained like a judging team member who studies official placings, reasons cards, or specific contest datasets. There's no internal lookup table of livestock classes or memorized "correct answers" I'm referencing in real time.

What actually drives my evaluation:

- I was trained on a **broad mixture of text data** (books, articles, general knowledge), which includes agriculture and animal science concepts.
- From that, I learned **principles**, not specific classes:
  - Structural correctness (angles, feet/leg set, mobility)
  - Muscle expression (width, shape, dimension)
  - Balance and phenotype
  - Maternal indicators in females (capacity, femininity, broodiness)
- When you send a class, I'm applying those **first-principles heuristics** to what I visually observe—essentially pattern recognition + domain logic.

What I am NOT doing:

- Not pulling from past judging contests or databases
- Not matching your images to known sets
- Not optimizing toward "official placings"
- Not learning or updating from your inputs during this session

Reality check (important for your use case):

This is closer to a **well-trained evaluator applying theory consistently**, not a **contest-trained judge calibrated to a specific circuit**.

That means:

- You'll get **internally consistent, principle-based placings**
- But not necessarily **industry-consensus placings every time**

Which, frankly, aligns well with what you said—you're testing independent judgment, not conformity.

Use QR for References

Results

MARK THAT CARD – RED ANGUS HEIFERS

Posted: April 29th, 2022

Results provided by Texas A&M Livestock Judging Team

Placing: 3 1 2 4

Cuts 5-2-3

Official Placing: 3-1-2-4 — Livestockjudging.com

Cut 1	Cut 2	Cut 3
5	2	3

2341 — 36 Score out of 50

Table 1

ChatGPT

Class Grouping	(*n*) Classes	(*n*) Correct Pair Splits	*M* of 50	*SD*	Range
Breed			34.25	9.56	21-48
Angus	8	1	34.00	-	-
Brahman	1	0	26.67	10.69	15-36
Charolais	3	0	35.00	7.07	27-46
Crossbred	7	1	35.20	6.42	26-42
Hereford	5	0	37.00	-	-
Limousin	1	0	35.50	2.12	34-37
Red Angus	2	0	34.00	-	-
Shorthorn	1	0	33.63	9.05	17-42
Simmental	8	0	34.25	9.56	21-48
Sex					
Bulls	4	1	35.25	13.52	15-43
Bred Heifers	4	0	32.25	5.68	24-37
Heifers/Heifer Calves	27	0	33.48	7.03	17-48
Steers	1	1	46	-	-
Total	36	2	33.89	7.75	15-48

Conclusions, Implications, Recommendations

- From baseline performance, ChatGPT cannot be considered an expert in livestock evaluation
- There may be improvements with corrective feedback
- It currently should not be used as a SBAE resource for Livestock Evaluation
- Further research should be conducted with corrective feedback and to understand AI limitations better