THE UNIVERSITY OF BRITISH COLUMBIA
Faculty of Land and Food Systems
Animal Welfare Program

vetmeduni

Cow welfare assessment using ChatGPT Vision

Sky (Kehan) Sheng[1], Mahshid Heydarirad[1], Borbala Foris[2]

1: Animal Welfare Program, Faculty of Land and Food Systems, Vancouver, BC, Canada
2: Centre for Animal Nutrition and Welfare, University of Veterinary Medicine, Vienna, Austria

INTRO & AIM

Welfare evaluation by trained assessors is time-consuming, infrequent, costly, and covers only a fraction of the herd…

Automated visual assessment of animal-based measures in cattle could contribute to a more efficient and consistent evaluation.

Let's use GPT-4o to assess cleanliness (dirty or clean) for 3 body parts (hind leg, hindquarter, udder), as it is done in the Welfare Quality Protocol (WQP)!

METHODS

User → **GPT-4o**

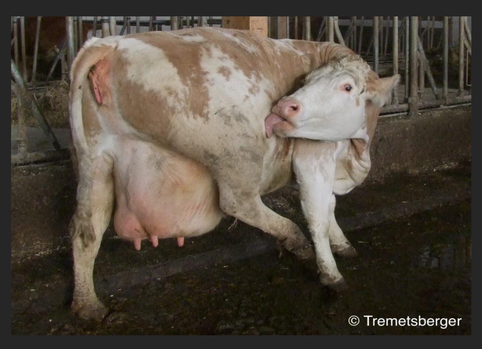
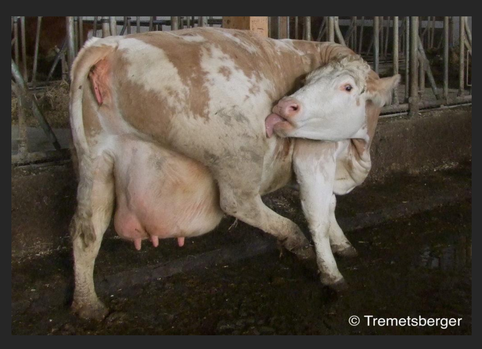
Prompt
"Act as an animal welfare assessor with 20 years of experience…"

Training text
WQP assessment instructions

Training images
2 images per category, labeled by 5 auditors

Test images
4 images per category, each tested in 10 rounds

Example of 4 image provision methods tested

	1. Original	2. Original boxed	3. Segment	4. Segmented body part
Training				
Test				
Response	**Dirty**. The lower hind legs of the cow show significant dirt accumulation, with continuous plaques of dirt visible above the coronary band. (10/10)	**Dirty**. The lower hind leg of the cow, including the hock, shows significant dirt accumulation. (10/10)	**Dirty**. The dirt is substantial and covers a significant portion of the lower hind legs, similar to the example images. (10/10)	**Clean**. The hind leg in the image shows no significant dirt or plaques. There is only minor splashing. (9/10)

RESULTS

	Accuracy				Precision				Recall				Kappa			
	hindleg	hindquarter	udder	overall	hindleg	hindquarter	udder	overall	hindleg	hindquarter	udder	overall	hindleg	hindquarter	udder	overall
original	0.61	0.62	0.39	0.54	0.56	0.57	0.43	0.52	1.00	1.00	0.75	0.92	0.25	0.25	-0.25	0.08
original boxed	0.71	0.62	0.42	0.59	0.63	0.57	0.45	0.55	1.00	1.00	0.75	0.92	0.50	0.25	-0.25	0.17
segment	0.50	0.54	0.49	0.51	0.50	0.52	0.49	0.50	1.00	1.00	0.97	0.99	0.00	0.00	0.00	0.00
segmented body part	0.71	0.62	0.52	0.62	0.63	0.57	0.52	0.58	1.00	1.00	0.75	0.92	0.50	0.25	0.00	0.25
overall	0.63	0.60	0.46	0.56	0.58	0.56	0.47	0.54	1.00	1.00	0.81	0.94	0.31	0.19	-0.12	0.12

CONCLUSION

GPT-4o is often biased towards labeling cows as dirty, potentially due to background dirt and difficulty identifying the correct body part. Bounding boxes and body part segmentation improve model performance.

A larger, high-resolution, standardized dataset with clear views of the relevant body parts could further improve model performance.