

MooBot: RAG-based Video Querying System for Dairy Cattle Behavior and Health Insights

Unmesh Raskar
University of Wisconsin–Madison
Madison, WI
uraskar@wisc.org

Omkar Prabhune, Hien Vu, and Younghyun Kim
Purdue University
West Lafayette, IN
{oprabhun, vu64, younghyun}@purdue.edu

Abstract

We introduce MooBot, a RAG-based video querying system powered by GPT-4o designed to bridge the gap between what complex cattle video data can provide and what dairy farmers need through a natural language web interface. MooBot applies computer vision inference on barn videos to detect cows, identify individuals, and classify their behaviors, transforming visual data into a structured schema containing useful insights. Our results demonstrate the potential of MooBot for enhancing accessibility to video-derived insights in precision livestock farming, bringing advanced computer vision analytics within reach of farmers without requiring technical expertise.

1. Introduction

Dairy farming is a highly data-driven industry, where managing cattle health and behavior efficiently is critical to optimizing production. Despite advancements in video surveillance and computer vision technologies, farmers still face challenges in making sense of the massive amounts of visual data generated from barn cameras. Traditional methods of video analysis are often manual and time-consuming, and require specialized expertise.

Recent research has applied deep learning models to classify cow behaviors from video and wearable sensors. However, these solutions are typically siloed, designed for offline use, and require domain-specific interfaces or knowledge. They lack the ability to support natural, conversational interaction across multiple data types and do not provide personalized, context-rich insights that align with the practical needs of farmers.

We introduce MooBot, a Retrieval Augmented Generation (RAG) [13]-based video querying framework powered by GPT-4o [9] designed to query barn video data through computer vision inference. By applying trained vision models to detect cows, identify individual cows, and classify behaviors, MooBot transforms raw video into structured

MooBot: Dairy Farm Chatbot

Ask me anything about your cattle!



Which is the most crucial cattle behavior change for early signs of heat stress?

B

The most crucial behavior change to monitor for early signs of heat stress in cattle is reduced feed intake. When cows experience heat stress, they often eat less and might also show an increase in water consumption. Additionally, observing a decrease in lying time can also indicate discomfort or distress related to heat. Monitoring these behaviors can help you address heat stress before it significantly impacts milk production and overall cow health.



Show me cows that were lying down in ventilation area between 10 am to 12 pm

B

Cows based on your query: ['C08', 'C09', 'C13']



Figure 1. MooBot web interface for querying dairy barn videos.

data that can be queried using natural language. When integrated with additional sensor streams such as 3D UWB (Ultra-Wideband) location data, IMMU (inertial and magnetic measurement unit) acceleration, milk yield, and environmental readings, MooBot offers personalized, context-aware insights to farmers.

To our knowledge, MooBot is the first system to support open-ended, natural language queries over integrated video and sensor data in dairy farming settings. MooBot’s ability to generate accurate responses from complex video data potentially enables solutions to improve dairy barn management practices. We detail the architecture of MooBot and its vision-centric pipeline, followed by an evaluation of

System	Domain	Modality	LLM interface	Insights provided
CowView [19]	Livestock	Location + Sensors	No	Raw tracking data
Dairy3D [12]	Livestock	Video + Location	No	Visualization only
Milchbot [7]	Livestock	Textual	Scripted chat	Static advice
BeefBot [23]	Livestock	Textual (RAG)	Yes (Llama3-8B)	QA from documents
ChatDB [8]	General	Tables (SQL)	Yes (LLM-SQL)	Textual insights
FinGPT [14]	Finance	Market data	Yes	Investment insights
MooBot (Ours)	Livestock	Video + Sensors	Yes (GPT-4o)	Behavioral insights from barn data

Table 1. Comparison of prior systems and MooBot.

both the computer vision models’ performance and the system’s query capabilities in a real-world setting. The paper highlights the potential of combining computer vision with natural language interfaces to bridge the gap between complex visual data and actionable insights, showing how they can support decision-making in livestock monitoring. In the following sections, we describe the system design, experimental setup, and key findings from our pilot deployment with dairy science researchers.

2. Related Work

Recent computer vision systems for livestock monitoring have primarily focused on animal detection, tracking, and behavior classification using single- or multi-camera setups [5, 15, 18]. These methods demonstrate strong performance on behavior classification and identity recognition tasks in barn environments. However, they offer low-level behavior labels but not “useful insights” such as summaries or contextual alerts for farm management. To enhance behavior recognition, recent work has explored spatio-temporal models. For example, a C3D-ConvLSTM (Convolutional Long Short-Term Memory) architecture can achieve up to 90% accuracy for behaviors such as feeding and grooming [16].

To improve robustness and context-awareness, multimodal approaches have emerged that fuse video with IMUs, UWB tracking, and environmental sensors [2, 10]. Commercial systems like CowView [19] and Dairy3D [12] enable real-time monitoring but require specialized infrastructure and technical dashboards.

Conversational interfaces in dairy farming remain underexplored. Milchbot [7] and the Chatbot E-system [4] enable natural language interaction through mobile applications, while more recent systems like BeefBot [23] use RAG to answer domain-specific questions from textual sources. However, these systems lack support for video understanding or multimodal sensor reasoning to utilize various recent implantable and wearable technologies [3, 20]. General-purpose assistants such as FarmerChat [17] and AgriAssist [6] offer advice across agricultural topics but do not integrate with real-time on-farm analytics. Earlier systems like Agribot [1] laid the groundwork for agricultural chatbots, yet lacked multimodal capabilities.

Outside of agriculture, several recent systems adopt large language model (LLM)-querying architectures over structured data. For example, FinGPT [14] offers financial insight generation via LLMs on structured datasets. Research efforts such as ChatDB [8], AutoSQL [24], and DB-GPT [22] explore LLM-to-SQL translation or RAG pipelines over knowledge bases. These systems highlight the feasibility of using LLMs as a conversational interface over data schemas, but most lack multimodal reasoning or spatio-temporal analytics.

We summarize this comparison in Tab. 1. MooBot builds upon these paradigms by integrating video inference and multimodal sensor fusion into a structured schema, enabling LLM-powered insights specifically tailored for dairy barn monitoring.

3. MooBot: System Overview

MooBot is designed as a modular RAG-based video querying pipeline that enables farmers to query complex barn video data using natural language. As shown in Fig. 2, the system comprises three core components:

- **Video pre-processing module:** Processes video feeds from the barn to detect individual cows, identify them based on their unique coat patterns, and predict their behaviors. These insights along with other multimodal data are stored in the master schema, which will be used to generate response to farmers’ queries.
- **Query processing module:** Translates farmers’ natural language queries to structured schema queries and executes it on the master schema, which is a structured, time-aligned table aggregating cow behaviors, sensor streams, and environment data—which will be used to generate responses to farmers’ queries.
- **Response generation module:** Converts the query results into a clear, easy-to-understand natural language response tailored for the farmer.

This section provides a conceptual overview of MooBot’s end-to-end pipeline, highlighting how data flows through each component of MooBot.

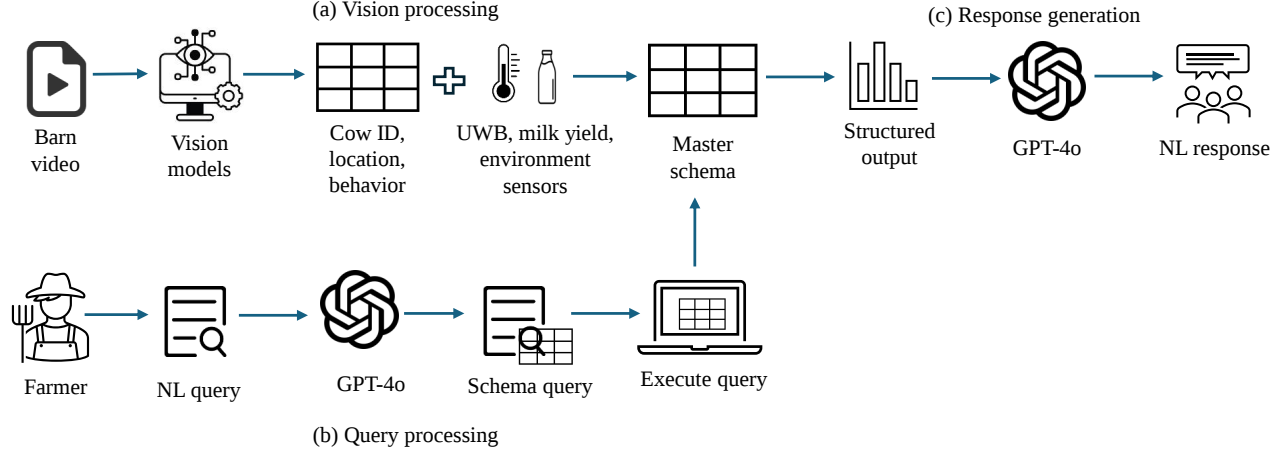


Figure 2. MooBot architecture and data processing pipelines.

3.1. Vision Processing

As shown in Fig. 2(a), input video streams are processed using a sequence of computer vision model. Initially, it detects individual cows in each frame, followed by identification through a cow ID recognition model. Subsequently, it classifies the behavior of each detected cow into predefined activity categories. This pipeline produces structured, frame-level annotations which associate each cow with a unique ID, inferred behavior, and spatial location across time.

These annotations are aggregated into a *master schema* Tab. 2, a unified representation of video-derived insights. The schema stores time-stamped records for each detected cow, including:

- Identification data: Predicted cow ID along with the corresponding confidence score.
- Behavioral data: Predicted activity of the cow such as lying, feeding, or walking.
- Location data: Frame-level projected coordinates and when available, 3D positions from UWB sensors.
- Health and productivity data: Visual health indicators and milk yield.
- Environmental context: Temperature, humidity, and other sensor readings.

This structured schema serves as MooBot’s internal representation for querying and response generation.

Table 2. Example entries from MooBot’s master schema.

Timestamp	Cow 01		Cow 02		Cow 03 Milk yield
	Location	Behavior	Location	CBT	
00:00:01	barn	standing	ventilation	normal	-
00:00:02	feeding	lying	feeding	high	-
00:00:03	barn	feeding	barn	high	-
00:00:04	milking	milking	milking	normal	-
00:00:05	pasture	standing	milking	normal	6.5 li
00:00:06	...	...	...	...	...

Motivation for schema-based querying. Traditional systems often perform computer vision inference after receiving the query, leading to high latency and substantial computational overhead from computer vision models’ inference [11]. In contrast, MooBot adopts a *preprocessing-then-querying* paradigm, where videos are processed in advance to extract and store relevant insights—such as cow identity, behavior, and location—in a structured schema. This decoupling of video processing from query execution enables efficient, low-latency responses without the need to re-run expensive vision models after receiving the query. This design offers several practical advantages:

- Low-latency queries: Once video inference is complete, MooBot can answer queries by directly retrieving from the schema.
- Edge-friendly deployment: Decoupling of AI inference and query execution enables offloading of large AI models to remote servers.
- Consistency: Queries are executed against a fixed representation of the video, ensuring that the same query always returns the same result, as vision model inference is performed only once.
- Transparency: Confidence scores are preserved throughout the pipeline, enabling MooBot to flag uncertain or low-confidence results.

This approach is well-suited to the constraints of dairy farm environments, where computational resources are limited, while timely, trustworthy insights are essential.

3.2. Query Processing

Farmer-submitted natural language queries are routed to this module, where they are interpreted and converted to structured schema-compatible queries. As shown in Fig. 2(b), MooBot uses an LLM, GPT-4o to interpret farmer

questions such as “Which cows were least active today?” or “Did Cow 07 show signs of heat stress?” and convert them into structured, schema compatible queries.

The LLM is used as-is via an API, with no additional fine-tuning. The system prompt given to the LLM contains the schema definition and instructions to convert user queries into structured outputs such as SQL-like or DataFrame logic.

The system prompt provided to the LLM includes the necessary schema context to be able to translate the queries within its scope. The LLM parses the user query and identifies key entities and temporal references. It translates the parsed intent into a structured query (e.g., SQL-like or DataFrame operations). Then, the structured query is executed over the master schema to retrieve the relevant data.

3.3. Response Generation

The query results retrieved from the master schema go into the next block as a structured output. In this final stage (Fig. 2(c)), MooBot translates query results into natural language responses and visual summaries.

As in the query stage, GPT-4o is used without fine-tuning. A second system prompt instructs the LLM to generate domain-appropriate answers that explain structured outputs clearly for farmers.

MooBot summarizes queried behavior and ranks results by confidence. Each response is annotated with model confidence (e.g., Cow 07 was detected feeding with 92% confidence). A second LLM stage transforms structured output into fluent, domain-appropriate responses. By propagating and exposing confidence throughout the pipeline, MooBot enables trustworthy and interpretable insights for non-technical users.

4. Implementation and Experiments

We evaluate MooBot’s effectiveness in processing barn video data and returning accurate, interpretable responses to natural language queries. Our experiments focus on the performance of the vision model, schema-based query translation, and the quality of final responses.

4.1. Dataset

We use the MmCows dataset [21], a multimodal collection from a real-world dairy farm comprising 14 days of continuous barn video from multiple viewpoints, 16 uniquely tagged cows, frame-level annotations for detection, identity, and behavior, and synchronized sensor streams including 3D UWB location, IMU acceleration, environmental conditions, and milk yield.

In addition to raw data, we construct a QnA dataset consisting of 150 natural language queries and gold-standard answers. These were generated using GPT-4o over the master schema and validated by dairy science researchers. The QnA pairs span a range of reasoning types, including

identity, behavior patterns, activity comparison, and spatio-temporal context. We provided GPT with a system prompt describing the schema and task, then manually curated and corrected outputs. We ensure temporal split separation between training, validation, and test splits to prevent identity leakage and evaluate generalization. For more details on MmCows and the system prompt used for QnA pairs generation, please refer to the supplementary material.

4.2. Evaluation Metrics

MooBot uses vision models pre-trained on an annotated portion of MmCows [21] in the vision pre-processing step. This pipeline achieves robust results across all vision tasks under real-world farm conditions, which we rely on to evaluate the following standard metrics to evaluate MooBot’s core modules:

- Cow detection: YOLOv8 model fine-tuned on a portion of the annotated part of the MmCows dataset detects cows in video frames, evaluated in mAP@0.5 and mAP@0.5:0.95.
- Cow identification: A 16-class classifier matches cows to known Cow IDs, evaluated in accuracy and per-class F1 score.
- Behavior classification: EfficientNet fine-tuned on MmCows categorizes behavior into 8 classes, evaluated in accuracy and per-class F1 score.
- Query translation: Match accuracy for fields, conditions, joins, and full queries by comparing LLM-generated queries to queries written and verified by expert annotators.
- Response quality: GPT-4o as a judge rating the responses on a scale 1–5 for relevance (correctness of response), groundedness (factual alignment with input), helpfulness (clarity and usefulness), and confidence transparency (clear indication of model certainty). To guide GPT-4o’s subjective scoring, we adopt few-shot prompting with examples of responses rated 1 and 5.

This setup allows quantitative assessment of MooBot’s interpretability and reliability in real-world use cases. For more details about the model training and inference, please refer to the supplementary.

4.3. Results

Vision model performance (Table S1 in Supplementary)

MooBot uses vision models pre-trained on an annotated portion of MmCows [21] in the vision processing step. In dairy barns, challenges are encountered with fine-grained cattle localization across multiple camera angles, lighting conditions and occlusions. This pipeline achieves robust results across all vision tasks under real-world farm conditions. For more details about the vision model training and results, please refer to the supplementary.

Query translation accuracy (Table 3): We evaluate MooBot’s ability to convert natural language questions into

executable schema queries. Results are measured against expert-verified ground truth for 150 queries. We observed that MooBot often confused feeding and licking behaviors in cows lying down, indicating a need for integrating UWB or accelerometer data.

Table 3. Query translation accuracy across subtasks.

Subtask	Accuracy
Field selection	94.0 $\pm$ 2.37%
Condition parsing	81.9 $\pm$ 3.85%
Join mapping	98.3 $\pm$ 1.01%
Full query equivalence	77.6 $\pm$ 4.16%

Response quality (Table 4): Responses were generated from 150 farmer-style natural language queries. GPT-4o was prompted with the question, MooBot’s response, and the schema context to rate four criteria. The responses of MooBot are scored by GPT-4o as a judge on a scale of 1 to 5.

Table 4. GPT-4o as judge evaluation of MooBot’s responses.

Metric	Average score (1–5)
Relevance	4.8 $\pm$ 0.1
Groundedness	4.4 $\pm$ 0.4
Helpfulness	3.7 $\pm$ 0.5
Confidence transparency	3.5 $\pm$ 0.8

4.4. Limitations and Future Work

Our current evaluation is limited to offline video processing and curated QnA pairs, with subjective scoring using GPT-4o. While we improve the consistency of these ratings via few-shot prompting (4.2), this remains an internal metric. Future work will include external validation through user studies involving farmers and livestock researchers, and comparison against standard analytics dashboards.

Without the unified master schema, many complex multi-modal queries would be computationally expensive or even infeasible to answer reliably. For example, aligning milk yield drops with spatial patterns and visual behaviors across asynchronous logs (e.g., video, accelerometer, milk events) requires precomputed temporal and semantic joins. Our system pre-processes and stores these relationships to support queries such as “Which cows stood still near the gate before a heat event?”—queries that would otherwise require slow and error-prone on-demand joins.

Future work includes supporting more complex temporal queries, enriching sensor integration, and deploying MooBot in live farm environments to assess real-world decision impact.

While MooBot is designed to support daily barn monitoring, any critical predictions (e.g. health alert) should be reviewed by veterinary or dairy science experts before

action. MooBot is an assistive tool, not a replacement for professional oversight.

While MooBot is developed for dairy barns, its core architecture—LLM-based interfaces over vision-derived structured data—can generalize to other tracking domains, including human activity monitoring. However, such applications raise important ethical concerns around privacy, surveillance, and misuse. Future research must consider safeguards to ensure responsible deployment, particularly in contexts involving humans.

5. Conclusion

MooBot presents a novel integration of computer vision and natural language interfaces for precision livestock farming. By applying vision models to barn videos and transforming the results into a queryable schema, MooBot enables farmers to interact with complex visual data through natural language while ensuring transparency in confidence levels. The system demonstrates how computer vision can play a crucial role in extracting valuable information from farm videos, while the natural language interface makes this complex data accessible to users without technical expertise.

A key contribution of MooBot lies in its modular pipeline that separates vision inference from query translation, enabling independent improvements and better error traceability. This design, along with the propagation of confidence scores throughout the pipeline, represents a key contribution to transparent AI systems in dairy farming. Future work will focus on improving vision model accuracy, expanding behavior classification capabilities, expanding natural language understanding of dairy cattle health conditions, and developing more sophisticated confidence reporting mechanisms to further enhance the reliability and utility of video-based farm management systems.

Acknowledgment

The authors thank the anonymous reviewers for their valuable feedback. This work was supported by the USDA National Institute of Food and Agriculture grant 2021-67021-34036 and the National Science Foundation grant 2435327.

References

- [1] Bhavika Arora, Dheeraj Singh Chaudhary, Mahima Satsangi, Mahima Yadav, Lotika Singh, and Prem Sewak Sudhish. Agribot: A natural language generative neural networks engine for agricultural applications. In *Proceedings of the International Conference on Contemporary Computing and Applications (IC3A)*, pages 28–33, 2020.
- [2] Said Benaissa, F. A. M. Tuytens, David Plets, Luc Martens, Leen Vandaele, Wout Joseph, and Bart Sonck. Improved cattle behaviour monitoring by combining Ultra-Wideband location and accelerometer data. *Animal*, 17(4):100730, 2023.

- [3] Hanwook Chung, Jingjie Li, Younghyun Kim, Jennifer M. C. Van Os, Sabrina H. Brounts, and Christopher Y. Choi. Using implantable biosensors and wearable scanners to monitor dairy cattle's core body temperature in real-time. *Computers and Electronics in Agriculture*, 174:105453, 2020.
- [4] Aishwary Sanjay Gattani, Shubham Sunil Kasar, Om Chakane, and Pratiksha Patil. Chatbot-based e-system for animal husbandry with e-farming. In *Proceedings of the International Conference on Information, Communication and Computing Technology (ICICCT)*, pages 569–584, 2023.
- [5] Navid Ghassemi, Ali Goldani, Ian Q. Whishaw, and Majid H. Mohajerani. Moo-ving beyond tradition: Revolutionizing cattle behavioural phenotyping with pose estimation techniques. *arXiv preprint arXiv:2408.06336*, 2024.
- [6] Zephyr Global. AgriAssist: AI in agriculture. <https://www.zephyrglobal.co/ai-agriculture>, 2024. (Accessed: 5/22/25).
- [7] Kevin Herrera, Juan Miranda, and David Mauricio. Milchbot: app to support the process of feeding and caring for dairy cows in Peru. *AGRIS on-line Papers in Economics and Informatics*, 14(4):27–37, 2022.
- [8] Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. ChatDB: Augmenting LLMs with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*, 2023.
- [9] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [10] Summaira Jabeen, Xi Li, Muhammad Shoib Amin, Omar Bourahla, Songyuan Li, and Abdul Jabbar. A review on methods and applications in multimodal deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):1–41, 2023.
- [11] Daniel Kang, Peter Bailis, and Matei Zaharia. BlazeIt: Optimizing declarative aggregation and limit queries for neural network-based video analytics. *arXiv preprint arXiv:1805.01046*, 2018.
- [12] Adrien Kroese, Moudud Alam, Elin Hernlund, David Berthet, Lena-Mari Tamminen, Nils Fall, and Niclas Högberg. 3-dimensional pose estimation to detect posture transition in freestall-housed dairy cows. *Journal of Dairy Science*, 107(9):6878–6887, 2024.
- [13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9459–9474, 2020.
- [14] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. FinGPT: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*, 2023.
- [15] Roua Mkadmi, Rabaa Youssef-Douss, and Amel Benazza-Benyahia. Cow monitoring system based on deep learning models for multiple objects detection and tracking. In *Proceedings of the International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8, 2024.
- [16] Yongliang Qiao, Yangyang Guo, Keping Yu, and Dongjian He. C3D-ConvLSTM based cow behaviour classification using video data for precision livestock farming. *Computers and Electronics in Agriculture*, 193:106650, 2022.
- [17] Namita Singh, Jacqueline Wang'ombe, Nereah Okanga, Tetyana Zelenska, Jona Repishti, Sanjeev Mishra, Rajsekar Manokaran, Vineet Singh, Mohammed Irfan Rafiq, Rikin Gandhi, et al. Farmer.Chat: Scaling AI-powered agricultural services for smallholder farmers. *arXiv preprint arXiv:2409.08916*, 2024.
- [18] Moniek Smink, Haotian Liu, Dörte Döpfer, and Yong Jae Lee. Computer vision on the edge: Individual cattle identification in real-time with ReadMyCow system. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 7056–7065, 2024.
- [19] E. Tullo, I. Fontana, D. Gottardo, K. H. Sloth, and M. Guarino. Validation of a commercial system for the continuous and automated monitoring of dairy cow activity. *Journal of Dairy Science*, 99(9):7489–7494, 2016.
- [20] Hien Vu, Hanwook Chung, Christopher Choi, and Younghyun Kim. eTag: An energy-neutral ear tag for real-time body temperature monitoring of dairy cattle. In *Proceedings of the International Conference on Mobile Computing and Networking (MobiCom)*, pages 706–720, 2023.
- [21] Hien Vu, Omkar Prabhune, Unmesh Raskar, Dimuth Panditharatne, Hanwook Chung, Christopher Y. Choi, and Younghyun Kim. MmCows: A multimodal dataset for dairy cattle monitoring. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 59451–59467, 2024.
- [22] Siqiao Xue, Caigao Jiang, Wenhui Shi, Fangyin Cheng, Keting Chen, Hongjun Yang, Zhiping Zhang, Jianshan He, Hongyang Zhang, Ganglin Wei, et al. DB-GPT: Empowering database interactions with private large language models. *arXiv preprint arXiv:2312.17449*, 2023.
- [23] Zhihao Zhang, Carrie-Ann Wilson, Rachel Hay, Yvette Everingham, and Usman Naseem. BeefBot: Harnessing advanced LLM and RAG techniques for providing scientific and technology solutions to beef producers. In *Proceedings of International Conference on Computational Linguistics: System Demonstrations (COLING)*, pages 54–62, 2025.
- [24] Xiaohu Zhu, Qian Li, Lizhen Cui, and Yongkang Liu. Large language model enhanced text-to-SQL generation: A survey. *arXiv preprint arXiv:2410.06011*, 2024.