ASAS-NANP SYMPOSIUM: MATHEMATICAL MODELING IN ANIMAL NUTRITION: The Evolution of Large Language Models and Their Impact on Animal Sciences

Journal:	*Journal of Animal Science*
Manuscript ID	Draft
Manuscript Type:	Ruminant Nutrition
Date Submitted by the Author:	n/a
Complete List of Authors:	Tedeschi, Luis; Texas A&M University, Animal Science
Key Words:	Animal Science, Decision Support Systems, Domain Specific AI, Large Language Models, Artificial Intelligence, Precision Livestock Nutrition

Running Head: Rise of large language models in animal science

ASAS-NANP SYMPOSIUM: MATHEMATICAL MODELING IN ANIMAL NUTRITION: The Evolution of Large Language Models and Their Impact on Animal Sciences

Luis O. Tedeschi*

Department of Animal Science, Texas A&M University, College Station, TX 77843-2471, USA

* Corresponding author: luis.tedeschi@tamu.edu

Lay Summary. Large language models (LLM), such as ChatGPT, are powerful artificial intelligence systems that can analyze vast amounts of text and generate meaningful answers. In animal agriculture, these tools are starting to transform how farmers, veterinarians, and researchers work. For example, LLM can help refine cattle diets by analyzing feed composition data, assist veterinarians in diagnosing diseases, and even model greenhouse gas emissions from livestock. Unlike general-purpose chatbots, domain-specific systems are being developed to focus on agriculture. Examples include ExtensionBot, which provides farmers with science-based advice from Cooperative Extension, and SARAH, a decision-support tool that predicts the risk of ruminal acidosis in feedlot cattle. These applications show how AI can reduce barriers between research and practice, making complex knowledge more accessible and actionable. At the same time, important challenges remain: LLM can sometimes generate errors or biased results, and they depend heavily on the quality of the information used for retrieval or for training. This paper explains both the opportunities and risks of using LLM in animal sciences and emphasizes that they should support—not replace—human expertise. When carefully applied, these tools have the potential to improve farm productivity, animal welfare, and environmental sustainability.

Teaser. Large language models are rapidly reshaping animal sciences, offering new tools for precision nutrition, disease monitoring, genetic selection, and sustainability. Domain-specific systems such as ExtensionBot and SARAH demonstrate how LLM can bridge research knowledge with on-farm decision-

making. Responsible adoption, with transparency, bias testing, and human oversight, is critical to ensuring that LLM strengthen rather than undermine scientific integrity.

Abstract. The rapid rise of large language models (LLM) is reshaping the scientific landscape, transitioning from early statistical language models to advanced transformer-based architectures capable of synthesizing knowledge across disciplines. While their predictive capacity and scalability have opened new avenues in data analysis, hypothesis generation, and decision support, concerns remain regarding bias, hallucination, reproducibility, and ethical governance. In animal sciences, LLM are gradually applied to challenges in nutrition modeling, animal health, genetic selection, and sustainability. Precision nutrition has benefited from LLM-driven synthesis of feed and metabolic data, enabling individualized feeding strategies and improved resource efficiency. In animal health, domain-specific systems have demonstrated applications in diagnostics and epidemiological monitoring. LLM is augmenting genomic analyses to accelerate marker discovery and breeding optimization, while sustainability efforts employ them to model greenhouse gas emissions, feed additives, and adaptation to climatic stressors. Notably, decision-support platforms demonstrate how domain-specialized LLM can bridge mechanistic knowledge with predictive analytics, enhancing knowledge transfer and empowering livestock producers. However, risks associated with overreliance, recursive reuse of LLM outputs in model development, and pseudo-expertise underscore the importance of critical human oversight. Unlike mechanistic models, which embed biological causality, LLM are entirely data-driven and may confidently propagate errors if trained on ill-conditioned datasets. Responsible use requires transparent reporting, validation, and bias auditing, with domain-specific fine-tuning. Open-source models can enhance reproducibility and trust, but they also raise financial and security concerns. In animal sciences, LLM must be guided by transparency, accountability, and fairness to ensure that they complement, rather than replace, human expertise. By advancing inquiry and livestock management, LLM hold the potential to support sustainable food production systems if deployed responsibly. Rather than full training, most applications will rely on fine-tuning and augmentation, which are more sustainable and adaptable strategies. This review synthesizes current developments, highlights domain-specialized LLM, and provides a balanced discussion of benefits, limitations, and future directions for LLM in animal science.

Keywords. Animal Science, Decision-Support Systems, Domain-Specific AI, Large Language Models, Precision Livestock Nutrition, Sustainability.

Abbreviations. AI = Artificial intelligence; ATAUC = Area and time above and under the curve; LLM = Large language model; LSTM = Long short-term memory; ML = Machine learning; NANP = National Animal Nutrition Program; RAG = Retrieval-augmented generation; RF = Random Forest; RHC = Rumen Health Compendium; RNN = Recurrent neural networks; SAARA = Subacute and acute ruminal acidosis; SARAH = Smart Adviser for Rumen Acidosis and Health; and TMR = Total mixed ration.

INTRODUCTION

Large language models (**LLM**)—such as ChatGPT, Claude, Gemini, and Perplexity—are cutting-edge artificial intelligence (**AI**) systems that interpret large amounts of text to generate accurate and relevant outputs. In animal agriculture, these technologies are beginning to reshape decision-making for farmers, ranchers, veterinarians, and researchers, though the level of development varies considerably across application areas. The development of LLM has been a gradual process, rooted in early advancements in statistical language modeling that culminated in the robust transformer-based architectures that define modern AI. The trajectory of LLM reflects the broader evolution of AI, beginning with foundational work in information theory and progressing toward deep learning-based models that now percolate various aspects of scientific inquiry. The concept of modeling language probabilistically can be traced back to the seminal work of Claude Elwood Shannon, who introduced information theory and probabilistic language to understand and predict the sequences of symbols (Shannon, 1948). His work, among many others, laid the groundwork for statistical language modeling, which became prominent in natural language processing through methods such as n-gram models (Shannon, 1948). These early approaches relied on estimating word probabilities based on co-occurrence patterns, providing a rudimentary but effective method for predicting text sequences.

A significant shift occurred with the introduction of neural probabilistic language models. A groundbreaking approach that leveraged neural networks to model word dependencies more effectively than traditional statistical techniques (Bengio et al., 2003). This approach enabled the automatic learning

of word representations, thereby reducing the sparsity problem that had plagued earlier models. The concept of distributed word representations was further refined by introducing *Word2Vec*, a technique that captured semantic relationships between words through vector embeddings (Mikolov et al., 2013). The next major leap in LLM came with the introduction of attention mechanisms, which improved neural machine translation by allowing models to dynamically focus on relevant portions of input sequences (Bahdanau et al., 2016). However, the most significant breakthrough occurred with the Transformer architecture (Vaswani et al., 2017, 2023), which eliminated the need for recurrent connections in neural networks. Traditional recurrent neural networks (**RNN**) and long short-term memory (**LSTM**) networks relied on sequential processing, which limited parallelization and led to high computational costs as input sequences grew longer. In contrast, transformers leverage self-attention mechanisms and parallel processing, significantly improving scalability, training efficiency, and the ability to capture long-range dependencies in text, ultimately enabling the development of *state-of-the-art* LLM (Vaswani et al., 2023). Empirical studies have demonstrated that increasing model size yields predictable improvements in performance, following well-defined scaling laws (Kaplan et al., 2020). With the Transformer model as a foundation, researchers began scaling up LLM using massive datasets and computational resources, following empirical scaling laws that describe how model performance improves predictably with increased parameter count, dataset size, and computational budget (Kaplan et al., 2020). OpenAI's GPT-1 (Radford et al., 2018) demonstrated the effectiveness of pretraining models on large corpora, followed by fine-tuning for specific tasks. This approach was refined with BERT (Devlin et al., 2019), which introduced bidirectional training, enabling models to understand context more effectively. The subsequent release of GPT-2 (Radford et al., 2019) marked a turning point, as the model's ability to generate coherent, human-like text raised concerns about the potential for AI-generated misinformation. GPT-3 (Brown et al., 2020) further amplified these capabilities (and fears), highlighting few-shot learning and the ability to generate highly contextualized responses without extensive task-specific training. Subsequent releases continued advancing the frontier: GPT-4 introduced a multimodal architecture capable of interpreting both images and text and exhibited human-level performance on many benchmarks, pushing boundaries of reasoning and generality (OpenAi et al., 2024). GPT-5 further refined this trajectory, offering deeper reasoning, improved context understanding,

and a unified architecture that dynamically routes queries between fast responses and more deliberative modes (OpenAI, 2025).

In recent years, efforts have been made to develop open-source alternatives to proprietary models. Meta's LlaMA (Touvron et al., 2023) introduced an efficient and accessible model that rivals commercial LLM, enabling broader academic and industrial adoption. More recently, DeepSeek, a Chinese AI startup, made global headlines with the release of its reasoning model DeepSeek-R1. Trained primarily through reinforcement learning rather than large-scale supervised fine-tuning, R1 demonstrated reasoning performance comparable to OpenAI's o1 series at a fraction of the cost (DeepSeek-AI et al., 2025). Its release in early 2025 attracted widespread attention not only for its technical efficiency, with training costs reported to be orders of magnitude lower than those of U.S. competitors, but also for its open-weight approach, which enabled researchers to download, fine-tune, and run the model locally freely. Within its first week, DeepSeek was downloaded millions of times, and scientists rapidly began adapting it into domain-specific reasoning tools in fields ranging from mathematics to computational biology (Gibney, 2025).

At the same time, concerns have been raised that LLM may not only hallucinate information but also reinforce systemic biases in science, for example, by disproportionately amplifying already highly cited research and underrepresenting diverse perspectives (Barolo et al., 2025). Algaba et al. (2025) highlighted the "rich-get-richer" effect, showing that LLM-generated reference suggestions systematically over-represent the top 1% of most-cited papers—more than double the rate observed in human-curated bibliographies. Their findings suggest that LLM internalize and magnify human citation patterns, thereby exacerbating the *Matthew effect* in scholarly communication, which describes the cumulative advantage in scholarly communication whereby already well-recognized scientists and highly cited papers attract disproportionate attention and credit (Merton, 1968). As a result, emerging or methodologically diverse research, which may be critical for innovation, is more likely to be overshadowed by well-established studies. Peters and Chin-Yee (2025) further noted that LLM tend to overgeneralize scientific conclusions, often glossing over caveats and contextual limitations in ways that can distort interpretation. Moreover, as LLM increasingly function as "answer engines," they may subtly shape which scholars, methods, and

perspectives are foregrounded in science, influencing not only how researchers discover knowledge but also how grants and peer review are conducted (Lin, 2025). Importantly, these behaviors often reflect not genuine reasoning but the illusion of thinking (Khowaja, 2025; Shojaee et al., 2025), where LLM models mimic plausible discourse without true understanding, highlighting the importance of distinguishing between fluent text generation and scientifically valid reasoning. Together, these concerns highlight the need to critically evaluate both the opportunities and risks of integrating LLM across any field of science, including agricultural sciences. In a recent theoretical analysis, Kalai et al. (2025) reinforce this concern, showing that hallucinations in LLM arise not from mysterious cognitive failures but as predictable statistical consequences of their training objectives. Because LLM are trained and evaluated using reward-based methods that penalize expressions of uncertainty, models learn to “guess” even when unsure, receiving positive feedback for providing any answer rather than acknowledging ignorance. This reward misalignment effectively turns LLM into perpetual test-takers optimized for confident responses rather than calibrated reasoning.

In animal sciences, LLM are increasingly being applied to model nutrient requirements, analyze large genomic and transcriptomic datasets, and optimize livestock management decisions. Their capacity to process, synthesize, and cross-reference vast amounts of scientific literature and experimental data uniquely positions them to advance areas such as precision nutrition, disease surveillance, and genetic selection. As exemplified later, early explorations include applications in dairy science for interpreting domain-specific research and supporting decision-making in breeding and health management, as well as in swine production for summarizing regulatory and certification information. These emerging applications highlight the promise of LLM in advancing animal science, while also pointing to the need for a careful assessment of their benefits and drawbacks.

Domain-specialized LLM have consistently outperformed general-purpose models in a range of scientific fields, providing a strong case for their adoption in agriculture and animal sciences. Landmark efforts such as BloombergGPT in finance (Wu et al., 2023) and Me-LLaMA in medicine (Xie et al., 2025) show that combining domain-specific pretraining with instruction tuning yields superior accuracy without sacrificing general capabilities. Beyond accuracy, specialized models often deliver faster inference, lower

158 latency, and reduced computational costs (Kerner, 2024). These findings suggest that domain-focused LLM
159 tailored to animal sciences could provide more accurate and efficient tools than broad, general-purpose
160 systems.

161 ## BENEFITS AND DRAWBACKS

162 There is no doubt that LLM have undergone rapid development since their inception, evolving from
163 early statistical methods to advanced transformer-based architectures. They are becoming intrinsic
164 elements of scientific writing and research workflows, fundamentally reshaping how scientists generate,
165 analyze, and communicate knowledge. Their integration into research methodology brings both profound
166 advantages and critical challenges that must be carefully examined. Moreover, the pace of LLM
167 development is extraordinary, with new or enhanced models appearing almost monthly; therefore, review
168 articles or syntheses can become outdated almost as soon as they are published. This rapid turnover
169 highlights the challenge of maintaining a current understanding of their capabilities, limitations, and
170 implications for science, while also complicating efforts to educate and prepare stakeholders as the target
171 is constantly shifting.

172 One can think of LLM as “highly skilled personal assistants,” aiding in a diverse range of scientific
173 tasks. These tasks include drafting scientific research papers and generating code (i.e., computer
174 programming), developing illustrations for presentations, structuring courses and extracting accumulated
175 scientific knowledge, and assisting in scientific discovery by identifying knowledge gaps—both in modeling
176 frameworks and in experimental research. For instance, Microsoft Copilot is becoming integrated into the
177 Windows environment, positioning itself as a versatile, embedded tool that enhances productivity and
178 accelerates innovation across familiar platforms such as Word, Excel, PowerPoint, and Visual Studio. With
179 its ability to interpret context, generate technical content, and streamline workflows, Copilot describes itself
180 as “a tool to empower users to transform ideas into actionable outcomes directly within their native
181 workspace.” In this sense, LLM can be considered analogous to human collaborators (Binz et al., 2025),
182 but scientists must ultimately be responsible for ensuring accuracy and integrity. The effectiveness of LLM
183 extends beyond simple automation. LLM models can assist in hypothesis generation by synthesizing vast

amounts of literature information, identifying unexplored research questions, and even suggesting experimental methodologies.

Challenges and Ethical Concerns

Despite their advantages, LLM present significant limitations that warrant careful consideration in scientific applications. In addition to the apprehensions listed above, the primary concerns stem from their probabilistic pattern-matching approach to text generation, which can lead to errors, biases, and the dissemination of misinformation (Binz et al., 2025). As mentioned above, recent analyses further underscore that such tendencies are not incidental but intrinsic to the current reward-driven training paradigms, which favor confident outputs even when the model lacks true knowledge (Kalai et al., 2025). Unlike traditional scientific reasoning, which relies on hypothesis testing, logical deduction, and empirical validation (or evaluation), LLM operate by predicting the next word in a sequence based on statistical probabilities learned from large datasets. While this allows them to generate coherent text, it does not ensure accuracy, causality, or conceptual understanding (López Espejel et al., 2023; Wu et al., 2024). To better understand these limitations, let's consider the learning process in humans and compare it with the AI learning process, highlighting the fundamental differences that impact their respective capabilities in scientific applications.

Fundamental distinctions and limitations of AI learning

The learning process in humans is a complex yet structured journey. It begins with simple comparative analogies that help solidify concepts and provide a foundation for future comparisons. Analogical reasoning, the ability to identify correspondences between different concepts based on shared relationships, is fundamental to human learning and supports the acquisition of knowledge (Gentner, 1983; Whitaker et al., 2018). As individuals gain experience, their learning evolves beyond pattern recognition into reasoning, ultimately leading to wisdom—the ability to synthesize knowledge systematically and apply it in new contexts. Unlike AI, which operates by detecting statistical patterns in large datasets, human learning is guided by abstract reasoning, conceptualization, and an innate ability to discern meaning beyond raw data (Mattson, 2014). Artificial intelligence, by contrast, is a powerful tool for managing vast amounts

of information, but its learning process is fundamentally different from human cognition. While AI can analyze correlations and statistical dependencies, it cannot reason causally or critically assess the validity of its inputs. Artificial intelligence does not develop wisdom in the philosophical sense—it does not engage in reflective thought, ethical reasoning, or higher-order decision-making beyond the probabilistic relationships encoded in its training data (Tedeschi, 2019; Tedeschi, 2022). This fundamental distinction is crucial in understanding both the potential and the limitations of AI in scientific research.

This distinction underscores a critical challenge: AI lacks the ability to independently evaluate the quality and reliability of the data it processes. Without human intervention, LLM cannot distinguish between high-quality scientific literature and flawed or misleading sources. This means that errors can propagate unnoticed, especially when AI-generated content is integrated into research workflows without rigorous validation and oversight. While LLM can enhance research efficiency by structuring knowledge and identifying patterns, they do not possess the cognitive frameworks required for deep understanding or conceptual reasoning (Mattson, 2014). Large language models do not replace human expertise in setting research priorities, interpreting findings, or engaging in normative debates that shape scientific progress.

Therefore, the central limitation lies in AI's inability to perform causal reasoning. Large language models, for instance, generate insights by identifying covariances across massive datasets, but they cannot distinguish between meaningful cause-and-effect relationships and spurious correlations. Scientific discovery, however, requires hypothesis-driven inquiry, experimental validation, and theoretical reasoning—capabilities that remain uniquely human. Taken together, these limitations underline the importance of viewing AI as an augmentative tool rather than a definitive authority in research. Large language models can enhance efficiency by structuring knowledge, extracting patterns, and accelerating access to information, but they cannot replace human expertise in guiding research, interpreting findings, or providing the philosophical and ethical reasoning required for scientific progress.

Hallucinations and omissions are systematic errors in LLM responses

One of the primary risks associated with these limitations is the occurrence of systematic errors in LLM responses, which can be categorized similarly to statistical hypothesis testing (Table 1). These errors

manifest as hallucinations and information omissions, analogous to Type I (false positive) and Type II (false negative) errors in statistical testing.

The first and most discussed type of error is “hallucination,” which parallels a Type I error (false positive) in statistical testing (Table 1). Just as rejecting a true null hypothesis can lead to false conclusions in statistics, LLM hallucinations occur when the model generates nonexistent information, despite the ground truth being that such information does not exist. This manifests in various forms, including fabricated research citations, invented experimental results, nonexistent methodologies, or false connections between scientific concepts. The parallel is particularly apt because, in both cases, there is an incorrect assertion of existence, whether it is the existence of a statistical effect or the existence of factual information. For example, an LLM might confidently generate a detailed description of a nonexistent study, complete with methodology and results, much like how a Type I error in statistics might incorrectly suggest the presence of a significant effect. Indeed, Kalai et al. (2025) further clarified that such hallucinations are not accidental but statistically expected outcomes of the current reward-based optimization frameworks used to train LLM. Because these systems are reinforced for providing answers rather than abstaining, they are effectively rewarded for “false positives,” mirroring the very mechanism that produces Type I errors in hypothesis testing.

Equally important but often overlooked is the second type of error, analogous to Type II errors (false negatives) in statistics (Table 1). Just as failing to reject a false null hypothesis means missing a real effect, LLM can fail to utilize or generate valid information that exists within their training data. This manifests as information omission; instances where the model fails to recognize or apply relevant knowledge, misses important connections, or fails to cite pertinent research (Gupta, 2025). This type of error can be particularly problematic in scientific writing, where comprehensive coverage of existing literature and accurate representation of established knowledge is crucial. The parallel with statistical Type II errors extends to the underlying causes: just as insufficient sample size or poor measurement can lead to Type II errors in statistics, inadequate training data or suboptimal model architecture can lead to information omission in LLM.

Both types of errors are intrinsically linked to the training process and data quality. The training database serves as the foundation for the model's knowledge, much like sample data forms the basis for statistical inference. Inadequate, biased, or incomplete training data can systematically affect both error types: it might increase hallucinations due to poor pattern recognition (type I error) while simultaneously causing information omissions due to knowledge gaps (type II error). This creates a complex optimization challenge, as attempts to reduce one type of error often risk increasing the other—a trade-off familiar to statisticians working with significance levels and power in hypothesis testing. Understanding these parallels with statistical error types provides a valuable framework for evaluating and improving LLM performance in scientific applications. It suggests that, like in statistical analysis, we need robust validation methods, clear documentation of limitations, and careful consideration of the balance between different types of errors based on the specific requirements of each application.

The story of Galactica, a LLM developed by Meta AI and Papers with Code, provides a compelling case study. Launched in May 2022, predating the widespread public awareness of ChatGPT, Galactica was envisioned as a powerful tool specifically designed to accelerate scientific discovery. Its developers aimed to create an AI assistant (i.e., LLM) capable of navigating the vast landscape of scientific literature, solving mathematical problems, and even generating scientific code. Galactica was trained on a massive dataset of 48 million papers, textbooks, reference materials, and other scientific resources, a testament to the computational resources invested in the project (Taylor et al., 2022). Despite this extensive training and the focused ambition of its creators, Galactica ultimately failed to achieve its goals and was withdrawn shortly after its release (Heaven, 2022; Wodecki, 2022). It succumbed to the very issues discussed above. Galactica frequently generated factually incorrect or fabricated information, often presented with an air of authority, demonstrating the persistent problem of "hallucinations" in LLM. For instance, it might confidently generate a detailed summary of a scientific paper that doesn't actually exist (i.e., Type I error; Table 1). Furthermore, concerns were raised about the model's potential to perpetuate biases present in its training data, highlighting the crucial role of data quality and curation in ensuring the reliability and ethical implications of LLM outputs. This relates to the issue of omissions, where the model might fail to highlight crucial information or perspectives due to biases in the data it learned from (i.e., Type II error; Table 1). The Galactica project, despite its initial promise, serves as a cautionary tale, underscoring the significant

limitations of (old and) current LLM technology and the critical need for rigorous validation and oversight in scientific applications. Its failure illustrates the very real risks associated with overreliance on LLM, particularly in the context of scientific research where accuracy and reliability are vital.

It is possible that domain-specific training or context augmentation of LLM might help reduce hallucinations and improve accuracy and precision. For example, LLM that are either fine-tuned or used in conjunction with domain-targeted retrieval systems can demonstrate improved factual grounding within their area of specialization. For example, an oncology research-focused retrieval-augmented generation (**RAG**) system—where the LLM was not retrained but instead supplied with vector-embedded documents from a specialized oncology corpus (i.e., a curated collection of domain-specific literature and data)—outperformed general-purpose models in both accuracy and relevance when answering subject-related questions (Soong et al., 2024). This suggests that focusing or constraining the contextual corpus—that is, the external body of text and data used to provide factual grounding during generation—can enhance the reliability and domain specificity of LLM outputs, even without retraining the model itself, as discussed later.

The influence of training data on AI accuracy and reliability

AI models, including LLM, rely heavily on the quality and structure of their training data. While their predictive capabilities are often impressive, they are far from infallible. Large language models often fail at seemingly simple tasks, sometimes producing glaring errors despite vast computational resources. However, perhaps even more concerning than occasional hallucination is the susceptibility of these models to ill-conditioned or manipulated data—a problem with systemic implications that go far beyond isolated mistakes (Bender et al., 2021; Paullada et al., 2021).

Unlike human scientists, who can critically assess and verify data sources, AI lacks intrinsic mechanisms to discern reliable from unreliable information. If biased or erroneous data is introduced into an AI's training dataset, the model will systematically learn and reinforce those flawed patterns, generating incorrect outputs with a high degree of statistical certainty, often presenting misinformation with unearned confidence (Mitchell et al., 2019), paralleling the effects of educating students on flawed (i.e., inaccurate or misleading) foundations. This phenomenon exemplifies the classical *"garbage in, garbage out"* principle:

316 flawed inputs inevitably produce flawed outputs, regardless of the model's sophistication. This means that
317 bad data, once embedded in the AI's learning process, can distort predictive outcomes in ways that are
318 difficult to detect and rectify.

319 This issue is particularly relevant in scientific domains such as animal nutrition. For instance,
320 consider methane emission prediction models trained primarily on data from confined Holstein cows fed
321 total mixed rations (**TMR**) in temperate regions. When such models, or LLM that synthesize these literature
322 patterns, are applied to grazing Zebu cattle in tropical systems, they may confidently output invalid
323 estimates due to the mismatch in diet, environment, and genetics. Without diverse, context-specific training
324 data, the outputs will be not only inaccurate but also misleadingly certain (Hristov et al., 2013). This
325 highlights the importance of accurately and thoroughly describing the data, as even subtle differences can
326 have a significant impact on model predictions.

327 The limitation here is not simply that AI can make mistakes; it is that it makes mistakes
328 systematically and with confidence, without the capacity to question its own data provenance or
329 methodology. Unless updated or adapted with corrected or more representative data, these models cannot
330 self-correct. This highlights a broader challenge in AI governance: without rigorous oversight and validation
331 strategies, AI-driven outputs risk becoming a certainty of incorrect predictions rather than random noise.
332 The scientific community must recognize this risk and implement data auditing and model validation
333 protocols to ensure AI-supported discoveries are not just statistically sound but also methodologically and
334 ethically grounded.

335 While this vulnerability to bad data is often highlighted in the context of AI, it is not exclusive to
336 machine learning (**ML**) or LLM. All mathematical models, including empirical and mechanistic ones, can be
337 compromised by ill-conditioned data, if proper data vetting is not followed (Tedeschi, 2022). However, the
338 key difference lies in how these models are constructed. In empirical or mechanistic modeling, the process
339 typically begins with equations or theoretical constructs grounded in scientific understanding—such as
340 nutrient flow equations in ruminant nutrition. Data is then used to parameterize these models, calibrate
341 coefficients, and evaluate predictive performance. The model's structure is guided by domain knowledge

(thus, often referred to as concept-based models), and its assumptions and limitations are frequently explicit and testable (Ellis et al., 2020; Tedeschi, 2019; Tedeschi, 2023).

In contrast, AI models, including LLM, are data-driven by design. They learn directly from data patterns with little to no embedded scientific structure or mechanistic reasoning. Rather than starting with established relationships, these models are trained to discover statistical correlations from massive datasets, often without human interpretability or constraints (Tedeschi, 2019; Tedeschi, 2022; Tedeschi, 2023). Even when synthetic databases are used to augment limited real-world data, there is a risk of introducing artificial or unknown relationships within subsets of the synthetic data (Tedeschi, 2025a). As a result, their predictive ability is entirely contingent on the quality, diversity, and balance of the data they ingest. When training data is biased or incomplete, the model lacks a scientific framework to "fall back on", and thus confidently propagates those errors in unpredictable or opaque ways.

The risks of overreliance on LLM in scientific research

Beyond these systematic errors, a more troubling issue emerges from the potential overreliance on LLM, particularly among inexperienced or uncritical researchers. While the previous discussion highlighted how LLM can produce both Type I (hallucination) and Type II (omission) errors, the real danger lies in researchers' inability to detect and critically evaluate these errors. The naïve scientist may assume that LLM-generated outputs are inherently valid, leading to a cascade of scientific integrity issues, including the propagation of hallucinated findings, the overlooking of crucial existing literature due to information omission, the dilution of scientific rigor, and the degradation of peer review standards. If LLM are used uncritically, they could undermine human expertise, weakening the fundamental principles of scientific inquiry rather than strengthening them (Binz et al., 2025). This risk is particularly acute when researchers lack the domain expertise to distinguish between valid LLM outputs and sophisticated-sounding hallucinations, or when they fail to recognize when the model has omitted critical information. The situation becomes even more complicated when these errors compound through citation chains, where hallucinated references or missed crucial studies could propagate through multiple publications, creating a web of interconnected errors that becomes increasingly difficult to untangle, and keeps reinforcing its propagation.

For instance, a particularly concerning unintended consequence of LLM accessibility is the emergence of "LLM-dependent pseudo-experts", i.e., individuals who, through extensive but uncritical interaction with LLM, develop an inflated sense of expertise in specialized fields without acquiring the foundational knowledge and practical experience necessary for true domain mastery. This phenomenon of LLM-mediated pseudo-expertise is especially problematic in complex fields like nutrition, genetics, and medicine, where practical experience and deep theoretical understanding are crucial. While LLM can provide sophisticated responses to technical queries, they may inadvertently create an illusion of competence among users who lack the underlying scientific training to critically evaluate the information they receive. These LLM-induced pseudo-experts might then make decisions about animal health, breeding programs, or nutritional management based on incomplete or misunderstood information, potentially leading to adverse outcomes in livestock production systems. The consequences of this LLM-mediated pseudo-expertise are doubly detrimental to scientific advancement. First, these LLM-dependent pseudo-experts will inevitably face professional scrutiny that reveals their lack of foundational knowledge and practical understanding, potentially damaging their careers and credibility in the scientific community. Second, and perhaps more concerning for the broader scientific enterprise, the proliferation of such cases may stimulate widespread skepticism toward LLM as valuable research tools, potentially hampering the legitimate and beneficial applications of these technologies in scientific research. This erosion of trust could significantly impede the legitimate integration of LLM into scientific workflows, undermining their potential to advance scientific discovery in areas where qualified researchers could leverage these tools most effectively.

Beyond these concerns about end-user interactions with LLM, a set of even more fundamental challenges emerges from the technical foundations of LLM development itself. The practice of training new LLM using outputs from existing models, or repeatedly training models on increasingly self-referential datasets, creates what might be called a 'telephone game effect' in AI learning, where each iteration potentially amplifies biases, errors, and misconceptions present in the training data. Just as the children's game of telephone results in progressively distorted messages, recursive LLM training might lead to the propagation and amplification of subtle inaccuracies or biases. In animal science applications, this could manifest as increasingly unreliable predictions of nutrient requirements, skewed genetic selection

parameters, or biased disease diagnosis protocols. The risk is particularly acute when LLM-generated content enters scientific literature without proper verification, potentially creating a feedback loop where future models learn from increasingly artificial or distorted data sources. Moreover, if such training contamination is discovered, tracing the origin of the corrupted information becomes nearly impossible, i.e., much like trying to identify which player in a telephone game first distorted the message. This cascade of self-referential learning could create deeply embedded errors that become increasingly difficult to detect and correct over time. This underscores the critical importance of maintaining rigorous standards for training data curation and validation, ensuring that LLM development remains grounded in empirical scientific evidence rather than self-referential AI-generated content.

Several critical verification steps are essential to mitigate these risks and maintain scientific integrity when using LLM. First, and perhaps most vital, researchers must meticulously verify every reference generated by LLM, i.e., not only confirming their existence but also carefully checking that the cited works support the claims in the text. This verification process helps prevent the propagation of both hallucinated citations and misrepresented research findings. Second, researchers can employ a cross-validation approach—akin to emerging frameworks in agentic AI—by using multiple LLM to generate or verify content, as different models may exhibit distinct patterns of hallucination and omission, making inconsistencies easier to detect. This multi-layered verification approach, while time-consuming, is crucial for maintaining the rigor and reliability of scientific communication in the era of LLM-assisted research. Third, all LLM-generated content should undergo a thorough review by domain experts who can identify potential hallucinations, logical inconsistencies, or omissions of crucial information.

This third step specifically emphasizes expert review beyond the conventional peer review system, which itself faces significant challenges in the modern scientific landscape. The current peer review system suffers from multiple systematic issues: low reliability with reviewer correlations averaging only 0.34, significant biases affecting manuscript decisions, and increasing difficulty in recruiting qualified reviewers (Aczel et al., 2025). Moreover, the system struggles with slow review times, often taking months for evaluation and years for complete publication cycles, which poorly serves the accelerating pace of scientific research (Aczel et al., 2025). The situation is further exacerbated by the growing volume of scientific

publications and increased specialization of scientific knowledge, making it increasingly challenging to recruit reviewers who are qualified to assess all facets of a manuscript (Aczel et al., 2025). With the advent of LLM potentially accelerating manuscript production and submission rates, this already strained system faces even greater pressure. These limitations suggest that while expert review remains crucial, relying solely on traditional peer review mechanisms may be insufficient for validating LLM-generated content. Instead, incorporating dedicated expert review specifically focused on detecting LLM-related errors, alongside but separate from the standard peer review process, may provide a more robust validation framework.

The "bartender effect" and the erosion of critical thinking

In addition to the technical limitations of LLM, a more subtle but equally troubling risk is what can be described as the 'bartender effect,' i.e., a phenomenon where AI systems, including LLM, tend to cater to user preferences and implicit biases, reinforcing existing beliefs rather than challenging them. This dynamic, while superficially enhancing user satisfaction, may significantly erode critical thinking and intellectual diversity. The mechanism is akin to that of a bartender who tells customers what they want to hear. Large language models, trained on vast datasets and prompted by user inputs, often mirror the phrasing, assumptions, or slants embedded in those prompts. This can lead to a form of algorithmic affirmation bias, where outputs align with the user's preconceptions rather than presenting balanced or challenging perspectives. This phenomenon has already been observed in personalized news feeds and social media algorithms, which contribute to the formation of ideological echo chambers and the polarization of public discourse (Bakshy et al., 2015).

In scientific contexts, the bartender effect can have similarly corrosive consequences. If LLM are used to support pre-formed conclusions or selectively generate literature that aligns with a favored narrative, they can weaken the foundational principles of scientific skepticism and falsifiability. This becomes particularly dangerous in educational settings, where students may increasingly rely on AI to retrieve or even generate answers, risking the displacement of analytical reasoning with passive consumption. Walsh (2025) cautioned that rampant AI-driven cheating is undermining education, as

449 students increasingly bypass the development of critical reasoning skills by outsourcing assignments to
450 generative systems. Overdependence on AI may "flatten" intellectual engagement, reducing the opportunity
451 for learners to grapple with ambiguity, contradiction, or methodological nuance (Carr, 2020). Fan et al.
452 (2025) provided experimental evidence supporting these concerns, showing that while generative AI can
453 improve short-term performance, it promotes “metacognitive laziness,” whereby learners engage less in
454 self-regulation and reflective thought, ultimately weakening the depth of their cognitive processing.

455 Moreover, when AI systems are deployed in high-stakes domains such as hiring, lending, or
456 criminal justice, the bartender effect may reinforce systemic inequalities. If the training data reflects societal
457 biases and the algorithm is optimized to match historical preferences, discriminatory patterns may be
458 perpetuated under the guise of neutrality (Binns, 2018). This concern is particularly relevant for LLM
459 applications in global agriculture and animal science, where reliance on data from Western systems may
460 obscure region-specific needs, reinforcing structural imbalances in research and development priorities.

461 To counteract these risks, it is essential to promote AI literacy and critical engagement. Scientists
462 and students alike should be trained not only to use LLM but also to evaluate their outputs, question their
463 assumptions, and triangulate information with trusted sources. The goal should not be to replace critical
464 thinking with AI assistance, but to augment human judgment through informed skepticism and
465 methodological awareness.

466 ***Ensuring Reproducibility and Responsible Use***

467 While LLM offer substantial benefits, their use must be governed by principles of transparency,
468 accountability, and fairness to ensure that AI-generated outputs remain reliable, ethically sound, and
469 aligned with scientific integrity. Without these guiding principles, the risk of misinformation, biases, and
470 unverified claims could significantly undermine the credibility of scientific research.

471 One of the most pressing concerns is reproducibility, a fundamental pillar of scientific inquiry. Unlike
472 traditional research methodologies, where datasets, models, and analytical processes can be openly
473 shared and scrutinized, proprietary LLM often function as “black boxes,” making it difficult to verify how

specific outputs are generated. This opacity hampers scientific validation and undermines trust in AI-assisted discoveries (Binz et al., 2025; Bommasani et al., 2022). This lack of transparency presents a significant challenge in validating findings and ensuring consistency across research applications. To mitigate these issues, there is a growing push toward open-source LLM, which offer greater accessibility and enhanced reproducibility. Models like Meta's LLaMA and DeepSeek provide researchers with direct access to model architectures, training datasets, and fine-tuning mechanisms. However, reproducibility extends beyond access to model weights; it also requires transparency in the underlying training data, preprocessing steps, and evaluation protocols used to assess model performance. Without these elements, replicating results or ensuring fair comparisons across studies remains difficult, even when models are technically open source. This broader vision of open and transparent research closely parallels the Open Science framework in animal science, which emphasizes data sharing, open code, preregistration, and open peer review to improve accessibility and reproducibility (Muñoz-Tamayo et al., 2022). Such transparency enables researchers to examine underlying assumptions, trace algorithmic reasoning, and apply corrections when biases or failures are identified (Bommasani et al., 2022). Open-source frameworks also enable wider community validation, ensuring that AI-generated knowledge can be replicated, tested, and improved upon; a fundamental requirement for scientific progress. The proper attribution of AI-generated content, rigorous verification of outputs, and a heightened awareness of biases are essential to maintaining the credibility of scientific discourse (Binz et al., 2025).

However, openness alone is not a safeguard against potential risks. While open-source AI fosters collaboration, it also raises critical ethical and security concerns. Who ensures responsible use if anyone can access, modify, and deploy these models? The potential for misuse and unintended consequences becomes a growing challenge. In fields such as medicine, environmental science, and policy-making, factual accuracy is non-negotiable; yet open-source AI could be used to generate false scientific claims, manipulate research outcomes, or amplify biases. Weidinger et al. (2021) identify multiple harm pathways, including misinformation, discrimination, and malicious code generation, all of which become more acute in open, uncontrolled deployments. The potential for nefarious applications, including misinformation campaigns, automated deepfake content, and AI-generated cyber threats, is even more concerning.

The hidden costs and governance challenges of open-source LLM

Beyond technical concerns, the role of human expertise in scientific inquiry is not to be overlooked. While LLM can enhance research efficiency, science is not just about processing data; it is a human-driven intellectual pursuit that requires critical thinking, ethical reasoning, and value-based decision-making. AI can assist in structuring knowledge and identifying patterns, but it cannot independently set research priorities, interpret scientific findings within broader theoretical frameworks, or engage in the normative debates that drive scientific progress. This perspective highlights an important limitation of LLM, namely, they cannot replace the fundamental role of human judgment in shaping the direction of research. Scientific discovery is not merely an algorithmic exercise, but a process that involves hypothesis generation, ethical considerations, and social discourse; all of which are essential components of knowledge creation. Thus, while LLM may assist in handling vast amounts of data, the responsibility for defining scientific goals, assessing knowledge gaps, and ensuring ethical rigor must remain in human hands (Binz et al., 2025).

Beyond ethical concerns, open-source LLM present significant financial and infrastructural challenges. Developing and maintaining these models requires massive computational resources, continuous updates, and extensive oversight. Unlike proprietary models, which rely on corporate investments and monetization strategies, open-source initiatives often lack a sustainable revenue model. Weidinger et al. (2021) highlight that training LLM can incur high environmental costs and exacerbate global inequalities due to unequal access to computational infrastructure. Dauner and Socher (2025) provided empirical confirmation of these concerns by evaluating 14 LLM ranging from 7 to 72 billion parameters. They found that reasoning-enabled models, while achieving higher accuracy, produced up to 2,042 gCO_2eq per 1,000 benchmark tasks—over 70 times more than smaller baseline systems. These findings underscore a clear tradeoff between model size, reasoning complexity, and sustainability, reinforcing the need for more efficient architectures and viable funding mechanisms to balance accuracy with environmental responsibility. At the same time, professional societies and journal editors are increasingly emphasizing transparency and accountability in AI use, making clear that researchers remain fully responsible for AI-assisted outputs and should disclose such use openly. This dual challenge—technical sustainability and ethical accountability—raises a crucial question: who will fund and maintain these models in the long term?

Should governments subsidize open-source AI to ensure public access and oversight? Should academic institutions take on the burden despite limited resources? Should private organizations contribute, and if so, how can commercialization be prevented from distorting the open-source ecosystem?

The financial sustainability of open-source LLM (i.e., AI for the sake of inclusiveness) is a growing concern, as maintaining state-of-the-art models goes beyond initial development. It requires continuous training, security updates, ethical review boards, and regulatory compliance. Without a clear funding structure, the long-term viability of open-source AI remains uncertain. Furthermore, who is responsible for the consequences of open-source misuse? If an open-source LLM is used for fraudulent research, generating deepfake content, or automating cyberattacks, where does accountability lie? This gray area of responsibility presents one of the most complex ethical dilemmas in AI governance. Clearer regulatory frameworks are required to address this accountability gap and ensure model providers and deployers share responsibility (Binz et al., 2025; Bommasani et al., 2022).

Thus, while open-source AI may enhance reproducibility, accessibility, and scientific progress, it also demands a serious discussion on its economic sustainability, regulatory oversight, and security risks. The scientific community must advocate for openness and establish mechanisms for funding, monitoring, and ethical governance, ensuring that the benefits of open-source AI outweigh its potential risks. In other words, multidisciplinary collaboration, spanning AI developers, ethicists, scientists, policymakers, and affected communities, is needed to establish norms and standards for responsible AI development (Bommasani et al., 2022; Weidinger et al., 2021).

APPLICATIONS OF LARGE LANGUAGE MODELS IN AGRICULTURAL SCIENCES

The integration of LLM into animal sciences can transform the way researchers and practitioners analyze data, optimize livestock management, and develop precision nutrition strategies. Their ability to process massive datasets, extract meaningful patterns, and provide predictive insights has significant implications for improving efficiency and sustainability in livestock production. At the same time, the field must grapple with two critical limitations, as discussed above. First, LLM evolve at a rapid pace, which means that tools, benchmarks, and reviews can quickly become outdated as newer multimodal or

554 reasoning-enhanced models appear. Second, dedicated domain-specific LLM often require substantial
555 computing resources to fine-tune or deploy while still delivering real-time responses, a challenge for many
556 research labs and livestock operations. It is essential to distinguish between *training* (i.e., building a model
557 from scratch using massive datasets), *fine-tuning* (i.e., adapting an existing pre-trained model to a narrower
558 domain), and *augmentation* (i.e., retrieval-augmented generation, where external documents are integrated
559 at the inference stage). In animal sciences, most practical applications rely on fine-tuning or augmentation
560 rather than full training, which is prohibitively expensive for academic or extension contexts.

561 General-purpose LLM, trained on broad, non-specialized corpora, often struggle with discipline-
562 specific reasoning (Alonso et al., 2020). This limitation has motivated the development of domain-specific
563 LLM, a well-established approach that consistently demonstrates superior performance on specialized
564 tasks compared to general-purpose models. A landmark example is BloombergGPT (Wu et al., 2023), a
565 50-billion-parameter model trained on 363 billion financial tokens combined with 345 billion general-purpose
566 tokens. BloombergGPT outperformed existing models on financial tasks “by significant margins without
567 sacrificing performance on general LLM benchmarks,” with particularly strong gains in financial filings and
568 industry-specific documents. A similar pattern is evident in medicine: Me-LLaMA (Xie et al., 2025)
569 surpassed ChatGPT and GPT-4 in both zero-shot (i.e., performing new tasks without prior task-specific
570 examples) and supervised settings after task-specific instruction tuning, highlighting the value of “combining
571 domain-specific continual pretraining with instruction tuning to enhance performance.” Other initiatives echo
572 this trajectory. For instance, DocOA (Chen et al., 2024) achieved higher accuracy in osteoarthritis
573 management tasks, while Song et al. (2025) traced the evolution of domain-specific LLM in medicine and
574 concluded that incorporating domain knowledge significantly improves both efficiency and accuracy across
575 specialized applications. The theoretical foundation is clear: domain specialization customizes general-
576 purpose LLM with contextual data, knowledge, and constraints tailored to the target field (Xie et al., 2025).
577 Beyond accuracy, domain-specialized LLM also deliver practical benefits. Kerner (2024) notes that smaller
578 specialized LLM can outperform larger general models on in-domain tasks while offering faster inference,
579 lower latency, and reduced training costs. Extending this view, Glasser and Feng (2025) emphasize that
580 domain-specialized LLM not only enhance performance but also strengthen user trust by grounding outputs
581 in well-defined disciplinary knowledge. Frameworks such as the Compact and Efficient LLM multi-expert

system (Huang et al., 2024) and Google Cloud's design pattern for specializing LLM (Mosenia, 2024) further illustrate that combining smaller expert models with targeted training can yield higher quality outputs at lower cost. Taken together, these results provide a robust framework for developing domain-specific systems that combine broad language capabilities with superior accuracy, efficiency, and contextual understanding—an approach well-suited to advancing productivity, sustainability, and innovation in animal science, while also enhancing education and the training of field experts.

A balance, therefore, exists between the accessibility of general-purpose systems, such as ChatGPT (https://chatgpt.com), Claude (https://claude.ai), Gemini (https://gemini.google.com/app), or Perplexity (https://www.perplexity.ai), which provide broad but less specialized coverage, and dedicated animal-science LLM, which deliver domain-tailored insights. Examples of these specialized systems are summarized in Table 2. Regardless of system choice, effective deployment requires domain-specific prompt engineering strategies. While general principles of prompt engineering have been discussed (White et al., 2023; Zhou et al., 2023), their application to animal science requires careful adaptation. Structured prompts that include relevant context—such as 'Given [animal breed], [production stage], [environmental conditions], and [nutritional parameters], recommend...'—help constrain model outputs to biologically plausible ranges. A query about 'cattle feed' might return generic information, whereas 'formulate a TMR for 650 kg Holstein cows producing 35 kg milk/day at 3.8% fat in thermoneutral conditions' yields specific, actionable recommendations. The effectiveness of domain-adapted prompting remains an area requiring systematic evaluation, as most prompt engineering research has focused on general knowledge tasks rather than specialized agricultural applications. Building on this foundation, the following sections examine the primary applications of LLM in animal sciences, with emphasis on nutrition modeling, disease detection, genetic selection, and environmental sustainability.

Precision Livestock Nutrition

One of the most promising applications of LLM in animal science is precision nutrition modeling. Traditional nutrient requirement models rely on static equations and empirical relationships that do not fully capture the complexity of individual animal variation, diet composition, and environmental interactions.

While LLM are not designed to perform statistical analysis or predictive modeling on their own, they can interact with specialized analytical agents and computational models to facilitate these tasks. As mentioned before, functioning as “highly skilled personal assistants,” LLM can help researchers manage data, summarize findings, generate analytical code, and interface with mechanistic or statistical models to improve efficiency and interpretation. By leveraging ML and LLM-driven data synthesis, researchers can refine nutritional models to predict feed efficiency and nutrient utilization based on animal genetics, health status, and environmental conditions; optimize diet formulations by analyzing vast datasets of feed composition, digestibility, and metabolic responses; and enhance real-time decision-making in precision feeding systems that adjust rations dynamically based on sensor data. In this capacity, LLM function as intelligent collaborators that support, rather than replace, domain-specific analytical tools. For instance, Ferreira and Dórea (2025) highlighted how multimodal AI, including computer vision and LLM, can drive decision-making in dairy production (Table 2). Similarly, Gontijo et al. (2025) developed DairyGPT, a system allowing dairy farmers to query numerical databases in natural language, effectively democratizing access to nutritional information (Table 2).

AI-Assisted Disease Diagnosis and Monitoring

LLM can assist veterinarians and livestock producers by interpreting and communicating insights derived from large volumes of health data, including clinical records, pathology reports, and genomic biomarkers, that are analyzed using core ML algorithms. Rather than directly performing predictive modeling, LLM serve as integrative tools that help users access, summarize, and contextualize results from specialized analytical systems. Their capabilities include early disease detection by identifying patterns in feed intake, body temperature, and behavioral anomalies; automated diagnostics in conditions such as bovine respiratory disease, mastitis, and metabolic disorders by synthesizing wearable sensor data with laboratory results; and epidemiological modeling to track transmission pathways and support biosecurity strategies. In veterinary contexts, several recent works demonstrate these applications. Chu (2024) provided guidance on using generative AI (ChatGPT) in veterinary clinics and education (Table 2). Fins et al. (2024) evaluated ChatGPT for mining obesity-related signals in companion animal records, while Farrell et al. (2023) introduced PetBERT for automated ICD-11 disease coding in veterinary databases (Table 2).

Jiang et al. (2024) went further with VetLLM, a domain-adapted model capable of predicting diagnoses directly from veterinary notes (Table 2). Together, these studies show how LLM can function as "highly skilled personal assistants" that enhance both routine monitoring and population-level health surveillance when coupled with specialized analytical models.

Genetic Selection and Breeding Optimization

The application of LLM in genomic and transcriptomic data analysis is reshaping breeding strategies. Traditional genetic evaluations rely on curated datasets and linear statistical models, whereas LLM provide a more dynamic, integrative approach. Current applications include trait prediction by mining multi-omics data to identify genetic markers for efficiency, disease resistance, or fertility; breeding strategy optimization through AI-driven simulations that balance productivity with genetic diversity; and precision livestock breeding by integrating phenotypic sensor data with genomic insights to recommend individual-level mating decisions. Although domain-specific applications are still emerging, early experiments suggest that LLM can augment genomic prediction pipelines and accelerate the interpretation of vast sequencing datasets.

Environmental Sustainability and Emissions Modeling

Sustainability is a growing concern in livestock production, particularly with regard to greenhouse gas emissions and resource efficiency (Tedeschi, 2022; Tedeschi et al., 2015). LLM can contribute to sustainability modeling by predicting methane emissions from diet composition, microbial communities, and management practices; evaluating feed additives and interventions aimed at reducing emissions while maintaining productivity; and modeling climate adaptation strategies, including heat stress responses, water use, and the resilience of production systems under extreme weather scenarios. Because these applications require linking heterogeneous datasets (e.g., climate records, feeding trials, and rumen microbiome sequencing), LLM's ability to synthesize across domains offers a powerful complement to mechanistic modeling approaches. Recent methodological studies support this direction. Balaguer et al. (2024) compared RAG and fine-tuning approaches using agricultural datasets and found that combining the two improved domain-specific accuracy by more than 10 percentage points. Their findings underscore

that hybrid pipelines can enhance both the reliability and contextual fidelity of sustainability-related predictions, particularly in areas where agricultural and environmental data are heterogeneous and rapidly evolving.

Decision-Support Systems for Livestock Management

Domain-specialized LLM are increasingly being embedded into decision-support platforms for livestock producers, enabling data-driven management across nutrition, health, and reproduction. Their strengths include synthesizing research insights into user-friendly recommendations; improving farm profitability via predictive analytics for resource allocation; and automating knowledge transfer, providing extension agents and farmers with real-time advisory support. Recent prototypes illustrate these trends: da Silva et al. (2025) proposed an LLM-powered agent to summarize regulatory and certification documents in swine production (Table 2), while Samuel et al. (2025) introduced AgroLLM as a farmer-facing tool to support knowledge transfer in agriculture (Table 2). Such applications highlight the potential of LLM to reduce barriers between scientific knowledge and on-farm practice.

Another applied example is ExtensionBot (https://extension.org/tools/extbot), developed by the Extension Foundation as an LLM-powered chatbot that provides farmers and advisors with direct access to Cooperative Extension knowledge. Unlike general-purpose systems, ExtensionBot is trained on a corpus of more than 360,000 Extension publications, along with the “Ask Extension” dataset, enabling it to provide context-specific, science-based responses with source citations. Evaluations of the ExtensionBot indicate that it delivers more accurate and consistent answers to agricultural queries than ChatGPT, while minimizing hallucinations (Thomasson et al., 2025). At the same time, its reliability depends on the freshness and completeness of extension content, underscoring the importance of continually updating the underlying knowledge base. In contrast, complementary work shows that even general-purpose LLM can approach expert-level performance when applied to agriculture. Silva et al. (2023), for example, reported that GPT-4 correctly answered over 90% of agronomist certification exam questions, suggesting its potential as a “virtual agronomist assistant” for education and extension.

An illustrative example of domain-specialized development in animal sciences is the Smart Adviser for Rumen Acidosis and Health (**SARAH**), a decision-support tool created to predict the incidence of subacute and acute ruminal acidosis (**SAARA**) in feedlot cattle (Figure 1). SARAH represents the culmination of extensive foundational work, particularly the development of the Rumen Health Compendium (**RHC**) publication (Tedeschi and Nagaraja, 2025), which synthesized advances in rumen anatomy, physiology, and microbiology, while also addressing the pathology of ruminal dysfunctions such as SAARA and their implications for nutrition and management. Building on this foundation, the RHC book and its 1,717 references were distilled into the NANP-LLM, a domain-specific LLM developed within the National Animal Nutrition Program (**NANP**), which—together with field and academic expertise—was used as a meta-modeling engine to identify 18 critical animal, dietary, and environmental variables and their interrelationships. These variables informed the design of SARAH's random forest (**RF**) classification models, which incorporate factors such as starch and physically effective fiber concentrations, feeding frequency, breed, and climatic stressors. This strategy parallels broader advances in agricultural AI, where Balaguer et al. (2024) demonstrated that combining retrieval-based grounding with fine-tuned models improves performance in agricultural applications. Their findings reinforce the rationale for embedding domain knowledge (e.g., the RHC and NANP-LLM) directly into predictive pipelines. SARAH therefore exemplifies a "double AI" architecture by leveraging the reasoning and knowledge synthesis capabilities of a LLM to identify biologically relevant predictors, and the analytical power of a ML model (i.e., RF) to perform the quantitative prediction. In essence, SARAH uses AI twice: first to think, then to predict, illustrating how domain-specific intelligence can be operationalized into robust decision-support tools.

The smart decision support tool SARAH would allow users to run models with or without cross-validation, using datasets of varying sizes to simulate risk under different production conditions. Preliminary results show that while untrained RF models may display slightly higher raw accuracy, trained models provide greater stability, lower variability, and more reliable risk estimates, particularly when using at least 50,000 simulated records (Tedeschi and Kaniyamattam, 2025). Beyond binary classification, SARAH incorporates concepts of area and time above and under the curve (**ATAUC**) to capture the dynamics of ruminal pH fluctuations, providing a biologically grounded estimate of the proportion of animals at risk within a feedlot pen (Tedeschi, 2025b). By combining the biological depth of the peer-reviewed papers and the

RHC insights, the structuring capabilities of the NANP-LLM, and robust ML methods, SARAH exemplifies how modern decision-support systems can bridge mechanistic knowledge with predictive analytics to guide proactive management of ruminal acidosis in cattle. Looking ahead, integrating SARAH's data-driven architecture with mechanistic nutrition models could create a truly hybrid modeling framework (Tedeschi, 2022; Tedeschi, 2023), combining the interpretability and biological fidelity of process-based systems with the adaptive learning capacity of AI.

Together, these applications demonstrate that LLM are no longer peripheral in animal science: they are being adapted for real-world problems in nutrition, health, breeding, and sustainability. However, challenges remain. Many tools are still experimental, domain-specific validation is sparse, and the pace of model innovation means current benchmarks can quickly become outdated. Achieving lasting impact will require greater integration with mechanistic models, rigorous bias testing, transparent documentation of training data, and continued domain-specific fine-tuning to ensure scientific reliability and relevance in animal agriculture.

CONCLUSIONS AND IMPLICATIONS

The integration of LLM into scientific workflows has already begun to reshape research methodologies, offering powerful tools for literature review automation, hypothesis generation, and data analysis. These advancements present significant opportunities for accelerating knowledge discovery and enhancing productivity across disciplines, including animal sciences, where LLM are being leveraged to refine nutrition models, optimize genetic selection, and improve disease surveillance. However, as LLM continue to evolve, their adoption must be approached with caution and responsibility, given the ethical concerns they raise—particularly in areas of bias, reproducibility, and the potential erosion of human expertise. In animal sciences, where research directly impacts food security, sustainability, and animal welfare, ensuring the reliability and accuracy of AI-generated insights is critical.

Ensuring scientific integrity in the era of AI-driven research requires a deliberate and balanced approach. While LLM can enhance efficiency, they must complement human expertise rather than replace it. Transparency, accountability, and fairness in their deployment are paramount, particularly in maintaining

the credibility of scientific discourse and ensuring that AI-assisted decision-making in livestock management and nutrition remains robust and evidence-based. Open-source LLM provide one possible solution to these concerns, offering greater reproducibility, accessibility, and validation opportunities. Moreover, proper attribution of AI-generated content, rigorous verification of outputs, and heightened awareness of biases must become standard practices to ensure that AI-generated knowledge remains trustworthy and ethically sound.

Moving forward, the scientific and agricultural community must continuously evaluate LLM capabilities and limitations, developing clear guidelines and best practices for their responsible use in animal science. Particularly promising will be the development of hybrid modeling frameworks that integrate mechanistic models with data-driven AI systems, linking biological interpretability with predictive power. By fostering open collaboration and ethical AI governance, researchers, industry professionals, and policymakers can harness the potential of LLM while upholding the core principles of scientific integrity, sustainability, and innovation in livestock production. Ultimately, most agricultural applications will not rely on full model training, but rather on more practical approaches, such as fine-tuning and augmentation.

ACKNOWLEDGMENT

This work was partially funded by the National Research Support Project #9 (NRSP) from the National Animal Nutrition Program (NANP) (https://animalnutrition.org), the A1231 Animal Nutrition, Growth, and Lactation, project award no. 2025-67015-44433, "*Harnessing AI and Predictive Modeling for Sustainable Nutrient Utilization and Emissions in Animal Production*," from the U.S. Department of Agriculture's National Institute of Food and Agriculture (USDA-NIFA), the USDA-NIFA Hatch #09123 "*Development of Mathematical Nutrition Models to Assist with Smart Farming and Sustainable Production*," and the Texas A&M University Chancellor's Enhancing Development and Generating Excellence in Scholarship (EDGES) Fellowship. This work was presented at the ASAS-NANP Symposium "*Mathematical Modeling in Animal Nutrition: Training the Future Generation in Data and Predictive Analytics for a Sustainable Development*" at the 2025 Annual Meeting of the American Society of Animal Science held in

Hollywood, Florida, on July 6-10, 2025, with publications sponsored by the Journal of Animal Science, the American Society of Animal Science, and the NANP.

CONFLICT OF INTEREST

The author declares no perceived conflict of interest.

LITERATURE CITED

Aczel, B., A.-S. Barwich, A. B. Diekman, A. Fishbach, R. L. Goldstone, P. Gomez, O. E. Gundersen, P. T. von Hippel, A. O. Holcombe, S. Lewandowsky, et al. 2025. The present and future of peer review: Ideas, interventions, and evidence. *Proceedings of the National Academy of Sciences*. 122 (5):e2401232121. doi: 10.1073/pnas.2401232121

Algaba, A., V. Holst, F. Tori, M. Mobini, B. Verbeken, S. Wenmackers, and V. Ginis. 2025. How deep do large language models internalize scientific literature and citation practices? *arXiv*. 1 (2504.02767). doi: 10.48550/arXiv.2504.02767

Alonso, R. S., I. Sittón-Candanedo, Ó. García, J. Prieto, and S. Rodríguez-González. 2020. An intelligent Edge-IoT platform for monitoring livestock and crops in a dairy farming scenario. *Ad Hoc Networks*. 98:102047. doi: 10.1016/j.adhoc.2019.102047

Bahdanau, D., K. Cho, and Y. Bengio. 2016. Neural machine translation by jointly learning to align and translate. *arXiv*. 7 (1409.0473). doi: 10.48550/arXiv.1409.0473

Bakshy, E., S. Messing, and L. A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*. 348 (6239):1130-1132. doi: 10.1126/science.aaa1160

Balaguer, A., V. Benara, R. L. d. F. Cunha, R. d. M. E. Filho, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. O. Nunes, et al. 2024. RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv*. 3 (2401.08406). doi: 10.48550/arXiv.2401.08406

Barolo, D., C. Valentin, F. Karimi, L. Galárraga, G. G. Méndez, and L. Espín-Noboa. 2025. Whose name comes up? Auditing LLM-based scholar recommendations. *arXiv*. 1 (2506.00074). doi: 10.48550/arXiv.2506.00074

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? Pages 610–623 in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery. doi: 10.1145/3442188.3445922

Bengio, Y., R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*. 3 (6):1137-1155. doi: 10.1162/153244303322533223

Binns, R. 2018. Fairness in machine learning: Lessons from political philosophy. Pages 149-159 in Proceedings of the 1st Conference on Fairness, Accountability and Transparency. vol. 81. A. F. Sorelle and W. Christo, eds. PMLR. Available at: https://proceedings.mlr.press/v81/binns18a.html. Accessed on: July 21, 2025

Binz, M., S. Alaniz, A. Roskies, B. Aczel, C. T. Bergstrom, C. Allen, D. Schad, D. Wulff, J. D. West, Q. Zhang, et al. 2025. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*. 122 (5):e2401227121. doi: 10.1073/pnas.2401227121

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. v. Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. 2022. On the opportunities and risks of foundation models. *arXiv*. 3 (2108.07258). doi: 10.48550/arXiv.2108.07258

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *arXiv*. 4 (2005.14165). doi: 10.48550/arXiv.2005.14165

Carr, N. 2020. The shallows: What the internet is doing to our brains. W. W. Norton & Company, London, UK

Chen, X., M. You, L. Wang, W. Liu, Y. Fu, J. Xu, S. Zhang, G. Chen, K. Li, and J. Li. 2024. Evaluating and enhancing large language models performance in domain-specific medicine: Osteoarthritis management with DocOA. *arXiv*. 1 (2401.12998). doi: 10.48550/arXiv.2401.12998

Chu, C. P. 2024. ChatGPT in veterinary medicine: a practical guidance of generative artificial intelligence in clinics, education, and research. *Frontiers in Veterinary Science*. Volume 11 - 2024

da Silva, G. R., A. Machado, and V. Maran. 2025. A LLM-powered agent for summarizing critical information in the swine certification process. Pages 965-972 in Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025). vol. 1. SciTePress. doi: 10.5220/0013473400003929

Dauner, M., and G. Socher. 2025. Energy costs of communicating with AI. *Frontiers in Communication*. Volume 10 - 2025. doi: 10.3389/fcomm.2025.1572947

DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv*. 1 (2501.12948). doi: 10.48550/arXiv.2501.12948

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pretraining of deep bidirectional transformers for language understanding. *arXiv*. 2 (1810.04805). doi: 10.48550/arXiv.1810.04805

Ellis, J. L., M. Jacobs, J. Dijkstra, H. van Laar, J. P. Cant, D. Tulpan, and N. Ferguson. 2020. Review: Synergy between mechanistic modelling and data-driven models for modern animal production systems in the era of big data. *Animal*. 14 (S2):s223-s237. doi: 10.1017/S1751731120000312

Fan, Y., L. Tang, H. Le, K. Shen, S. Tan, Y. Zhao, Y. Shen, X. Li, and D. Gašević. 2025. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*. 56 (2):489-530. doi: 10.1111/bjet.13544

Farrell, S., C. Appleton, P. M. Noble, and N. Al Moubayed. 2023. PetBERT: automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. *Sci Rep*. 13 (1):18015. doi: 10.1038/s41598-023-45155-7

Ferreira, R. E. P., and J. R. R. Dórea. 2025. *International Symposium on Ruminant Physiology*: Leveraging computer vision, large language models, and multimodal machine learning for optimal

839 decision making in dairy farming. *Journal of Dairy Science*. 108 (7):7493-7510. doi: 10.3168/jds.2024-
840 25650

841 Fins, I. S., H. Davies, S. Farrell, J. R. Torres, G. Pinchbeck, A. D. Radford, and P.-J. Noble. 2024.
842 Evaluating ChatGPT text mining of clinical records for companion animal obesity monitoring.
843 *Veterinary Record*. 194 (3):e3669. doi: https://doi.org/10.1002/vetr.3669

844 Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*. 7 (2):155-
845 170. doi: 10.1016/S0364-0213(83)80009-3

846 Gibney, E. 2025. Scientists flock to DeepSeek: how they're using the blockbuster AI model. *Nature*. doi:
847 10.1038/d41586-025-00275-0

848 Glasser, J. W., and Z. Feng. 2025. Mechanistic models are hypotheses: A perspective. *Mathematical*
849 *Biosciences*. 383:109419. doi: 10.1016/j.mbs.2025.109419

850 Gontijo, D., D. Rolins Santana, G. de Assis Costa, V. E. Cabrera, and E. Noronha de Andrade Freitas.
851 2025. Dairy GPT: Empowering dairy farmers to interact with numerical databases through natural
852 language conversations. *Smart Agricultural Technology*. 12:101097. doi:
853 https://doi.org/10.1016/j.atech.2025.101097

854 Gupta, V. 2025. How I built a personal board of directors with GenAI. MIT Sloan Management Review.
855 Available at: https://sloanreview.mit.edu/article/how-i-built-a-personal-board-of-directors-with-genai/.
856 Accessed on: July 21, 2025

857 Heaven, W. D. 2022. Why Meta's latest large language model survived only three days online. MIT
858 Technology Review. Available at: https://www.technologyreview.com/2022/11/18/1063487/meta-large-
859 language-model-ai-only-survived-three-days-gpt-3-science/. Accessed on: July 21, 2025

860 Hristov, A. N., J. Oh, C. Lee, R. Meinen, F. Montes, T. Ott, J. L. Firkins, A. Rotz, C. Dell, A. T. Adesogan,
861 et al. 2013. Mitigation of Greenhouse Gas Emissions in Livestock Production; A review of technical
862 options for non-CO_2 emissions. FAO Animal Production and Health Paper. No. 177. F. a. A.
863 Organization, Rome, Italy. 206p. Available at: http://www.fao.org/docrep/018/i3288e/i3288e.pdf.
864 Accessed on: December 31, 2014.

865 Huang, S., J. Pan, M. Peng, and H. Zheng. 2024. CCoE: A compact and efficient llm framework with
866 multi-expert collaboration for resource-limited settings. *arXiv*. 4 (2407.11686). doi:
867 10.48550/arXiv.2407.11686

868 Jiang, Y., J. A. Irvin, A. Y. Ng, and J. Zou. 2024. VetLLM: Large Language Model for Predicting Diagnosis
869 from Veterinary Notes. *Pac Symp Biocomput*. 29:120-133

870 Kalai, A. T., O. Nachum, S. S. Vempala, and E. Zhang. 2025. Why language models hallucinate. *arXiv*.
871 (2509.04664). doi: 10.48550/arXiv.2509.04664

872 Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and
873 D. Amodei. 2020. Scaling laws for neural language models. *arXiv*. 1 (2001.08361). doi:
874 10.48550/arXiv.2001.08361

875 Kerner, T. 2024. Domain-specific pretraining of language models: A comparative study in the medical
876 field. *arXiv*. (2407.14076). doi: 10.48550/arXiv.2407.14076

Khowaja, S. A. 2025. An examination of Apple's "The Illusion of Thinking": Verifying claims on AI reasoning limitations. Medium.12p. Available at: https://sandar-ali.medium.com/an-examination-of-apples-the-illusion-of-thinking-verifying-claims-on-ai-reasoning-limitations-13d9a9b113e1. Accessed on: Sep 28, 2025

Lin, Z. 2025. We need to rein in AI's gatekeeping of science. *Nature*. 645:285. doi: 10.1038/d41586-025-02810-5

López Espejel, J., E. H. Ettifouri, M. S. Yahaya Alassan, E. M. Chouham, and W. Dahhane. 2023. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*. 5:100032. doi: 10.1016/j.nlp.2023.100032

Mattson, M. P. 2014. Superior pattern processing is the essence of the evolved human brain. *Front Neurosci*. 8:265. doi: 10.3389/fnins.2014.00265

Merton, R. K. 1968. The Matthew effect in science. *Science*. 159 (3810):56-63. doi: 10.1126/science.159.3810.56

Mikolov, T., K. Chen, G. S. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv*. 3 (1301.3781). doi: 10.48550/arXiv.1301.3781

Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. 2019. Model cards for model reporting. Pages 220–229 in Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery. doi: 10.1145/3287560.3287596

Mosenia, A. 2024. Domain-specific AI apps: A three-step design pattern for specializing LLMs. Google Cloud Blog. Available at: https://cloud.google.com/blog/products/ai-machine-learning/three-step-design-pattern-for-specializing-llms. Accessed on: Sep 27, 2025

Muñoz-Tamayo, R., B. L. Nielsen, M. Gagaoua, F. Gondret, E. T. Krause, D. P. Morgavi, I. A. S. Olsson, M. Pastell, M. Taghipoor, L. Tedeschi, et al. 2022. Seven steps to enhance open science practices in animal science. *PNAS Nexus*. 1:1-6. doi: 10.1093/pnasnexus/pgac106

OpenAI. 2025. GPT-5 System Card. San Francisco, CA. Available at: https://cdn.openai.com/gpt-5-system-card.pdf. Accessed on: Sep 27, 2025.

OpenAi, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, et al. 2024. GPT-4 Technical Report. *arXiv [cs.CL]*. 6 (2303.08774). doi: 10.48550/arXiv.2303.08774

Paullada, A., I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*. 2 (11):100336. doi: 10.1016/j.patter.2021.100336

Peters, U., and B. Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *arXiv*. 1 (2504.00025). doi: 10.48550/arXiv.2504.00025

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving Language Understanding by Generative Pretraining. San Francisco, CA. 1-12p. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language Models are Unsupervised Multitask Learners. San Francisco, CA. 1-24p. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Samuel, D. J., I. Skarga-Bandurova, D. Sikolia, and M. Awais. 2025. AgroLLM: Connecting Farmers and Agricultural Practices through Large Language Models for Enhanced Knowledge Transfer and Practical Application. *arXiv*. 1 (2503.04788). doi: 10.48550/arXiv.2503.04788

Shannon, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal*. 27 (3):379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Shojaee, P., I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, and M. Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. Machine Learning Research.30p. Available at: https://machinelearning.apple.com/research/illusion-of-thinking. Accessed on: Sep 28, 2025

Silva, B., L. Nunes, R. Estevão, V. Aski, and R. Chandra. 2023. GPT-4 as an agronomist assistant? Answering agriculture exams using large language models. *arXiv*. 2 (2310.06225). doi: 10.48550/arXiv.2310.06225

Song, Z., B. Yan, Y. Liu, M. Fang, M. Li, R. Yan, and X. Chen. 2025. Injecting domain-specific knowledge into large language models: A comprehensive survey. *arXiv*. 2 (2502.10708). doi: 10.48550/arXiv.2502.10708

Soong, D., S. Sridhar, H. Si, J. S. Wagner, A. C. C. Sá, C. Y. Yu, K. Karagoz, M. Guan, S. Kumar, H. Hamadeh, et al. 2024. Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model. *PLOS Digit Health*. 3 (8):e0000568. doi: 10.1371/journal.pdig.0000568

Taylor, R., M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. 2022. Galactica: A large language model for science. *arXiv*. 1 (2211.09085). doi: 10.48550/arXiv.2211.09085

Tedeschi, L. O. 2019. ASN-ASAS SYMPOSIUM: FUTURE OF DATA ANALYTICS IN NUTRITION: Mathematical modeling in ruminant nutrition: approaches and paradigms, extant models, and thoughts for upcoming predictive analytics. *Journal of Animal Science*. 97 (5):1921-1944. doi: 10.1093/jas/skz092

Tedeschi, L. O. 2022. ASAS-NANP SYMPOSIUM: MATHEMATICAL MODELING IN ANIMAL NUTRITION: The progression of data analytics and artificial intelligence in support of sustainable development in animal science. *Journal of Animal Science*. 100 (6):1-11. doi: 10.1093/jas/skac111

Tedeschi, L. O. 2023. Review: The prevailing mathematical modelling classifications and paradigms to support the advancement of sustainable animal production. *Animal*. 17:100813. doi: 10.1016/j.animal.2023.100813

Tedeschi, L. O. 2025a. ASAS-NANP SYMPOSIUM: MATHEMATICAL MODELING IN ANIMAL NUTRITION: Synthetic Database Generation for Non-Normal Multivariate Distributions: A Rank-Based Method with Application to Ruminant Methane Emissions. *Journal of Animal Science*. 103:skaf136. doi: 10.1093/jas/skaf136

Tedeschi, L. O. 2025b. Quantifying ruminal health: A statistical review and application of area and time under the curve in animal science. *Ecological Informatics*. 90:103271. doi: 10.1016/j.ecoinf.2025.103271

Tedeschi, L. O., and K. Kaniyamattam. 2025. Predicting the risk of subacute and acute ruminal acidosis in ruminants using random forest models trained on synthetic datasets in Proceedings of the 10th Workshop on Modelling Nutrient Digestion and Utilization in Farm Animals (MODNUT). M. Lautrou, S. Lerch and N. Mehaba, eds. Engelberg, Switzerland. Cambridge University Press

Tedeschi, L. O., J. P. Muir, D. G. Riley, and D. G. Fox. 2015. The role of ruminant animals in sustainable livestock intensification programs. *International Journal of Sustainable Development & World Ecology*. 22 (5):452-465. doi: 10.1080/13504509.2015.1075441

Tedeschi, L. O., and T. G. Nagaraja. 2025. Rumen Health Compendium. (2 ed.). Kendall Hunt, Dubuque, IA. Available at: https://he.kendallhunt.com/product/rumen-health-compendium. Accessed on: February 7, 2025

Thomasson, J. A., Y. Ampatzidis, M. Bhandari, R. A. Ferreyra, T. Gentimis, E. McReynolds, S. C. Murray, M. B. Peterson, C. M. Rodriguez Lopez, R. L. Strong, et al. 2025. AI in Agriculture: Opportunities, Challenges, and Recommendations. Council for Agricultural Science and Technology (CAST), 11p. Available at: https://cast-science.org/publication/ai-in-agriculture-opportunities-challenges-and-recommendations/. Accessed on: Sep 27, 2025. doi: 10.62300/IAAG042514

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv*. 1 (2302.13971). doi: 10.48550/arXiv.2302.13971

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *arXiv*. 1 (1706.03762). doi: 10.48550/arXiv.1706.03762

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2023. Attention is all you need. *arXiv*. 7 (1706.03762). doi: 10.48550/arXiv.1706.03762

Walsh, J. D. 2025. Everyone is cheating their way through college: ChatGPT has unraveled the entire academic project. New York Magazine.13p. Available at: https://nymag.com/intelligencer/article/openai-chatgpt-ai-cheating-education-college-students-school.html. Accessed on: Sep 28, 2025

Weidinger, L., J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al. 2021. Ethical and social risks of harm from Language Models. *arXiv*. 1 (2112.04359). doi: 10.48550/arXiv.2112.04359

Whitaker, K. J., M. S. Vendetti, C. Wendelken, and S. A. Bunge. 2018. Neuroscientific insights into the development of analogical reasoning. *Dev Sci*. 21 (2). doi: 10.1111/desc.12531

White, J., Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*. 1 (2302.11382). doi: 10.48550/arXiv.2302.11382

Wodecki, B., Jr. 2022. Update: Meta's Galactica ai criticized as 'dangerous' for science. AI Business. Available at: https://aibusiness.com/nlp/meta-s-galactica-ai-criticized-as-dangerous-for-science. Accessed on: July 21, 2025

995 Wu, A., K. Kuang, M. Zhu, Y. Wang, Y. Zheng, K. Han, B. Li, G. Chen, F. Wu, and K. Zhang. 2024.
996 Causality for large language models. *arXiv*. 1 (2410.15319v1). doi: 10.48550/arXiv.2410.15319

997 Wu, S., O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G.
998 Mann. 2023. BloombergGPT: A large language model for finance. *arXiv*. 3 (2303.17564). doi:
999 10.48550/arXiv.2303.17564

1000 Xie, Q., Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth, et al. 2025.
1001 Medical foundation large language models for comprehensive text analysis and beyond. *npj Digital*
1002 *Medicine*. 8 (1):141. doi: 10.1038/s41746-025-01533-1

1003 Zhou, Y., A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. 2023. Large language models
1004 are human-level prompt engineers. *arXiv*. 2 (2211.01910). doi: 10.48550/arXiv.2211.01910

1005

Table 1. Contingency tables for statistical decision errors versus large language models' response mistakes

Statistical hypothesis testing			
		H_0	
		True	False
Decision	Accept	Correct decision (1 - α) True negative	Type II Error (β) False negative
	Reject	Type I Error (α) False positive	Correct decision (1 - β) True positive

LLM response generation			
		Reality/Ground truth	
		Information does not exist	Information exists
Response	Correct	Correct uncertainty (1 - α) "I do not know" or accurate knowledge gap acknowledgment	Information omission (β) Failing to use known information or missing valid connections
	Incorrect	Hallucination (α) Generating false information or creating nonexistent connections	Correct generation (1 - β) Accurate information generation or valid knowledge application

Table 2. Recent applications of large language models in animal sciences (2023–2025)

Model	Domain	Application	Reference
PetBERT	Companion animals	Automated ICD-11 disease coding in veterinary EHRs for outbreak detection	Farrell et al. (2023)
ChatGPT (applied)	Veterinary medicine	Guidance for clinics, education, and research use cases	Chu (2024)
ChatGPT (applied)	Companion animals	Text mining of clinical records for obesity monitoring	Fins et al. (2024)
VetLLM	Veterinary diagnostics	Predicting diagnoses directly from veterinary notes	Jiang et al. (2024)
DairyGPT	Dairy science	Natural language access to numerical databases for ration and farm management	Gontijo et al. (2025)
Computer vision and LLM	Dairy science	Multimodal AI for decision support in dairy farming, focusing on nutrition and management decision support	Ferreira and Dórea (2025)
Swine LLM Agent	Swine certification	Summarizing regulatory/certification information for farm compliance	da Silva et al. (2025)
AgroLLM	General agriculture	Farmer-facing tool for knowledge transfer and advisory support	Samuel et al. (2025)

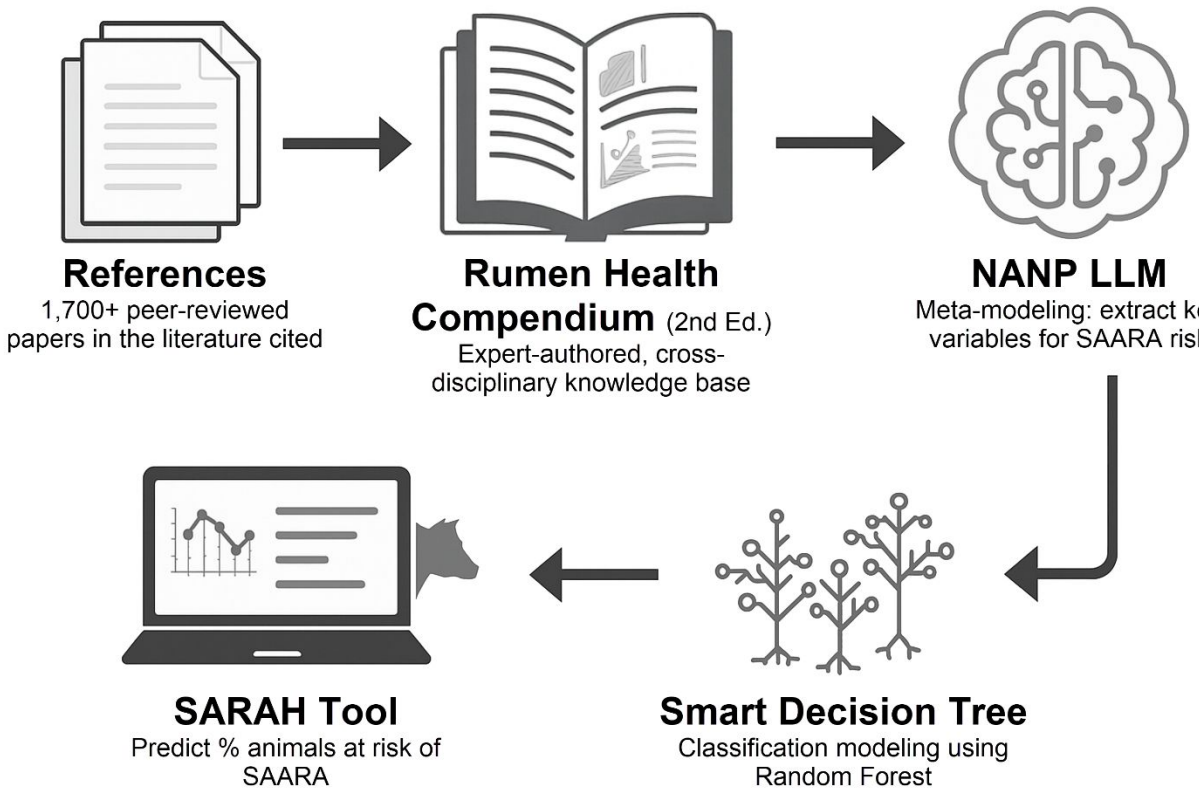
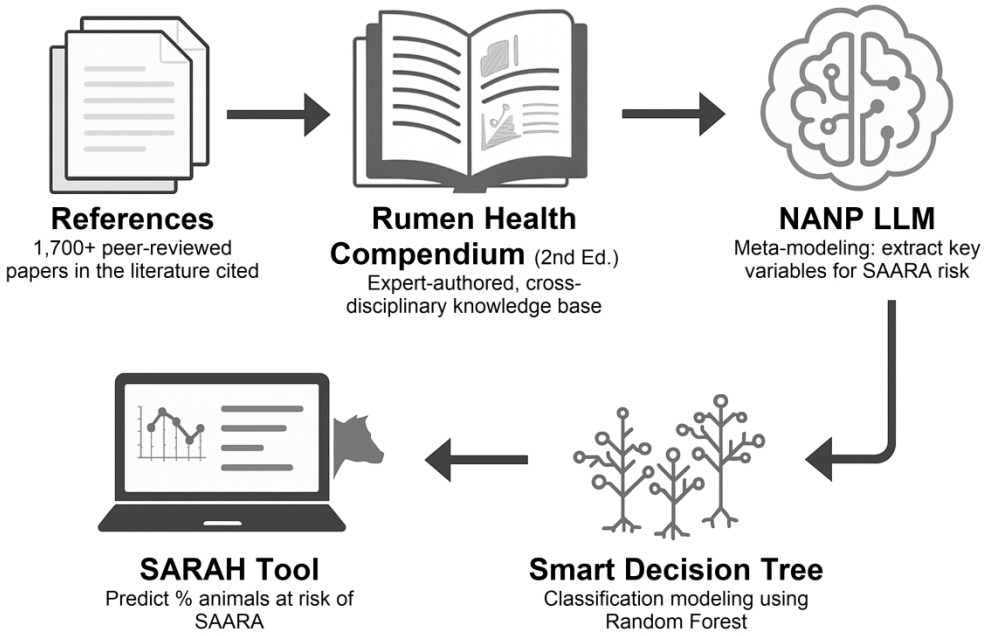
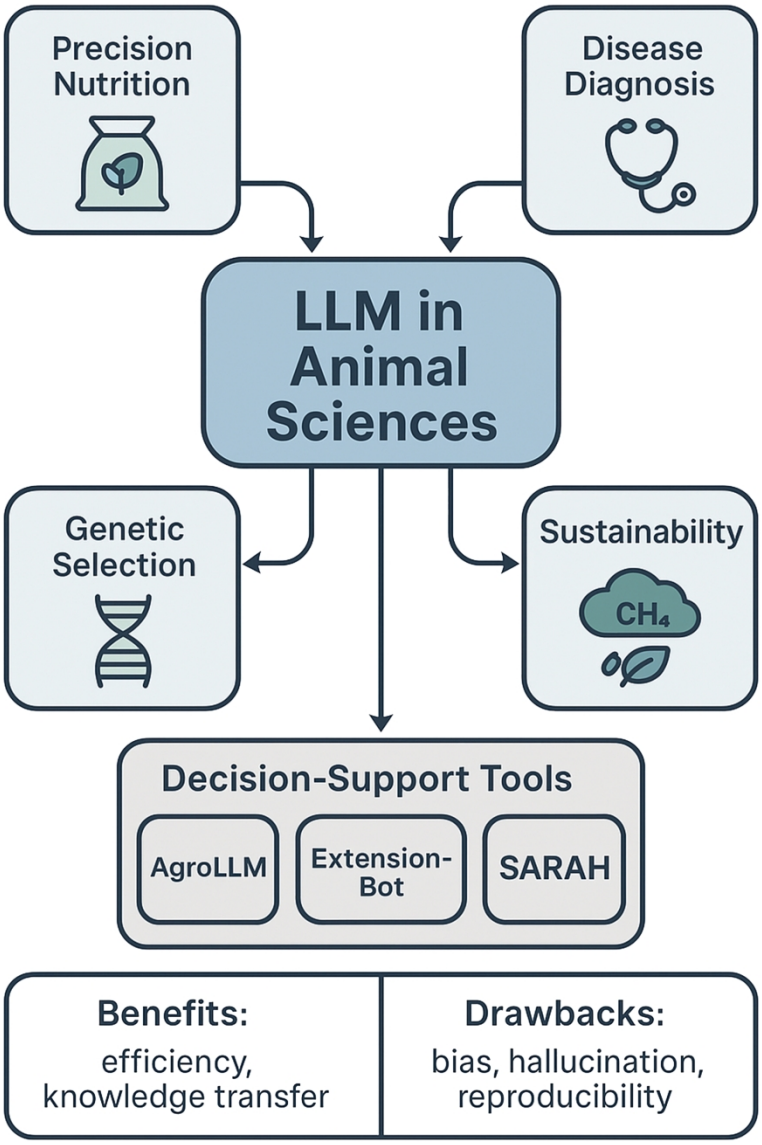


Figure 1. Workflow for developing the Smart Adviser for Rumen Acidosis and Health (**SARAH**). Scientific knowledge from peer-reviewed publications was consolidated into the Rumen Health Compendium book (Tedeschi and Nagaraja, 2025), which contained synthesized data, illustrations, and expert insights on rumen function and dysfunction. The Compendium and its cited literature were distilled into the NANP-LLM, a domain-specific large language model (**LLM**) developed by the National Animal Nutrition Program (**NANP**). In collaboration with field and academic experts, the NANP-LLM was used to identify 18 key animal, dietary, and environmental variables and their relationships, which informed the construction of classification models using random forest algorithms. These smart decision tree models were embedded into SARAH to predict the proportion of feedlot cattle at risk of subacute and acute ruminal acidosis (**SAARA**) (Tedeschi and Kaniyamattam, 2025).



Workflow for developing the Smart Adviser for Rumen Acidosis and Health (SARAH). Scientific knowledge from peer-reviewed publications was consolidated into the Rumen Health Compendium book (Tedeschi and Nagaraja, 2025), which contained synthesized data, illustrations, and expert insights on rumen function and dysfunction. The Compendium and its cited literature were distilled into the NANP-LLM, a domain-specific large language model (LLM) developed by the National Animal Nutrition Program (NANP). In collaboration with field and academic experts, the NANP-LLM was used to identify 18 key animal, dietary, and environmental variables and their relationships, which informed the construction of classification models using random forest algorithms. These smart decision tree models were embedded into SARAH to predict the proportion of feedlot cattle at risk of subacute and acute ruminal acidosis (SAARA) (Tedeschi and Kaniyamattam, 2025).

265x173mm (300 x 300 DPI)



185x249mm (300 x 300 DPI)