A Decision Support AI-Copilot for Poultry Farming: Leveraging Retrieval-Augmented LLMs and Paraconsistent Annotated Evidential Logic E

Marcus Vinicius Leite *, Jair Minoro Abe, Irenilza de Alencar Nääs, Marcos Leandro Hoffmann Souza

Posted Date: 16 December 2025

doi: 10.20944/preprints202512.1284.v1

Keywords: poultry production; poultry farming; decision support system; LLM large language models; RAG retrieval augmented generation; paraconsistent annotated evidential logic E_T; smart farming

 doi:10.20944/preprints202512.1284.v1

Article

A Decision Support AI-Copilot for Poultry Farming: Leveraging Retrieval-Augmented LLMs and Paraconsistent Annotated Evidential Logic Eτ to Enhance Operational Decisions

Marcus Vinicius Leite [1,*], Jair Minoro Abe [1], Irenilza de Alencar Nääs [1] and Marcos Leandro Hoffmann Souza [2]

[1] Graduate Program in Production Engineering, Paulista University – UNIP, Brazil

[2] Computer Science, Universidade do Vale do Rio dos Sinos – UNISINOS, Brazil

* Correspondence: marcus.leite@gmail.com

Abstract

Driven by the global rise in animal protein demand, poultry farming has evolved into a highly intensive and technically complex sector. According to FAO, animal protein production increased by about 16% in the past decade, with poultry alone expanding 27% and becoming the leading source of animal protein. This intensification requires rapid, complex decisions across multiple aspects of production under uncertainty and strict time constraints. This study presents the development and evaluation of a conversational decision support system (DSS) designed to support decision-making to assist poultry producers in addressing technical queries across five key domains: environmental control, nutrition, health, husbandry, and animal welfare. The system combines a large language model (LLM) with retrieval-based generation (RAG) to ground responses in a curated corpus of scientific and technical literature. Additionally, it adds a reasoning component using Paraconsistent Annotated Evidential Logic Eτ, a non-classical logic designed to handle contradictory or incomplete information. Evaluation was conducted by comparing system responses with expert reference answers using semantic similarity (cosine similarity with SBERT embeddings). Results indicate that the system successfully retrieves and composes relevant content, while the paraconsistent inference layer makes results easier to interpret and more reliable in the presence of conflicting or insufficient evidence. These findings suggest that the proposed architecture provides a viable foundation for explainable and reliable decision support in modern poultry production, achieving consistent reasoning under contradictory or incomplete information where conventional RAG chatbots would fail.

Keywords: poultry production; poultry farming; decision support system; LLM large language models; RAG retrieval augmented generation; paraconsistent annotated evidential logic Eτ; smart farming

1. Introduction

Poultry production has become the most widely consumed source of animal protein worldwide, driven by rising global demand, rapid urbanization, and the intensification of livestock systems [1–4]. According to FAO, animal protein production increased by about 16% in the past decade, with poultry alone expanding 27% and becoming the leading source of animal protein [1]. As production scales grow, poultry farmers are increasingly required to make rapid and complex decisions involving environmental control, nutrition, health, animal welfare, and husbandry, often under conditions of uncertainty, time pressure, and conflicting information [4,5].

 doi:10.20944/preprints202512.1284.v1

To cope with the growing decision complexity of intensive poultry systems, a variety of farm management platforms integrating decision-support tools have been introduced [36,38,39]. Examples include eFarm, a precision agriculture application that integrates health, feed, and production metrics for dairy and poultry operations [39]; farmOS, a community-driven open-source platform for planning and record keeping [40]; and the Poultry Farming Management System, which automates data collection for inventory, production, sales, and expenses [41]. While these systems improve data organization and reporting providing valuable functionalities for record keeping, planning, and health or production tracking, they primarily serve as dashboards or recordkeeping applications. Their embedded DSS modules, although useful for routine monitoring, are not designed to cope with uncertainty, contradictory inputs, or overlapping decision domains—challenges that are common in intensive poultry farming. As a result, similar to these platforms and their decision-support modules, most existing tools remain narrow in scope, focused on isolated domains, with limited integration across technical areas and little resilience to contradictory or incomplete information [4,5,36,38,39].

In practice, Decision Support Systems (DSS) in poultry farming often take the form of deterministic rule-based or AI-based controllers, IoT monitoring platforms, big data solutions, and statistical dashboards that track environmental conditions, animal health indicators, and production metrics [5,29,30,36]. Although these technologies provide valuable data, they typically operate under significant limitations—such as infrastructure demands, expertise gaps, and cost-related constraints—and are frequently based on fixed thresholds or rigid decision rules, lacking mechanisms for context-aware inference or adaptive reasoning [17,29]. Consequently, current systems struggle to accommodate uncertainty, conflicting signals, and the need for multi-domain integration in real-world decision-making scenarios [34,35].

These constraints have motivated the exploration of knowledge-based approaches that incorporate structured reasoning and domain expertise to enhance decision robustness [8,36]. In this context, recent advances in Large Language Models (LLMs) offer promising capabilities for contextual understanding, flexible inference, and semantic generalization, particularly when enriched with Retrieval-Augmented Generation (RAG) mechanisms that ground responses in external content [6–10,36]. However, despite the potential of LLMs in extracting, composing, and synthesizing complex technical knowledge from unstructured sources, these models still struggle when faced with conflicting or incomplete information [8,11–13]. Moreover, there is a significant knowledge gap in the application of LLMs to livestock production, particularly regarding the challenges of poultry farming processes, which opens opportunities for further research and technological advances [38]. This gap highlights the need for a framework that not only leverages LLM+RAG but also introduces an evidential reasoning layer capable of contradiction-tolerant inference. In this sense, standard RAG-based models collapse under contradictory signals, whereas paraconsistent reasoning explicitly tolerates and structures such conflicts.

To address these challenges, this study examines the integration of LLMs and RAG with Paraconsistent Annotated Evidential Logic Eτ (Logic Eτ). This non-classical framework enables reasoning under contradictory, insufficient, or ambiguous evidence. While LLMs provide linguistic generalization and RAG ensures factual grounding through external sources, Logic Eτ adds an inferential layer that explicitly handles conflicting or incomplete evidence, providing transparency and robustness in decision-making processes [8,14–16].

The objective of this proof-of-concept study is to develop and evaluate a conversational knowledge-based DSS for poultry production, structured as a conversational agent—the Decision Support AI-Copilot for Poultry Farming—that answers domain-specific queries using LLMs, content retrieved via RAG, and paraconsistent inference based on Logic Eτ. The system supports five critical areas of poultry production: environmental management, animal nutrition, health monitoring, husbandry, and animal welfare. Its configuration was defined through controlled experiments designed to evaluate both the quality of semantic retrieval and the strength of the reasoning, optimizing generative behavior and logical consistency. Performance was then assessed through comparison with expert-curated references using semantic similarity metrics (cosine similarity with

 doi:10.20944/preprints202512.1284.v1

SBERT embeddings) and evidential assessments. While the Discussion briefly contrasts the proposed framework with recent LLM+RAG approaches, a comprehensive comparison with other DSS approaches lies beyond the scope, as the focus here is on feasibility and methodological contribution.

2. Materials and Methods

This study adopts an applied and experimental methodology to design and evaluate a knowledge-based decision support system for poultry production, combining theoretical modeling, computational implementation, and empirical evaluation. All materials, algorithms, and procedures are described in detail to ensure reproducibility and enable replication by future research.

2.1. Methodological Framework Overview

The methodological framework integrates three complementary components:

1. Theoretical modeling with Logic Eτ, which provides the inferential foundation for reasoning under uncertainty and contradiction, supporting key decision points in the system workflow.
2. Experimental validation through Design of Experiments (DoE), conducted as proof-of-concept trials to tune system-level parameters affecting semantic retrieval, preprocessing, and generative behavior, rather than as large-scale validation.
3. System implementation of the Decision Support AI-Copilot for Poutry Farming, developed as a modular RAG-based architecture that integrates LLMs with evidential reasoning mechanisms based on Logic Eτ.

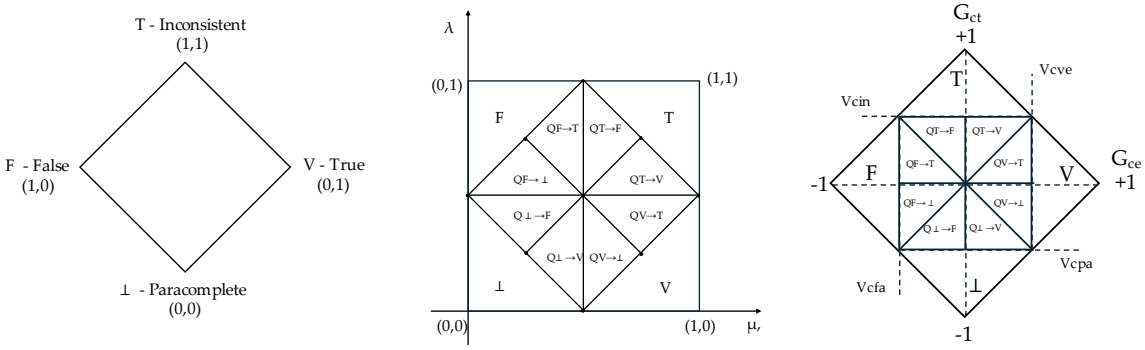
The following subsections present each component in sequence, ensuring a coherent integration between theoretical modeling, experimental validation, and system implementation.

2.2. Evidential Inference with Logic Eτ

Conventional LLM-based systems struggle when confronted with imprecise, incomplete, or contradictory inputs, a critical limitation in technical decision-support scenarios [6,11]. To address these challenges, the proposed system incorporates Logic Eτ as a complementary inference mechanism for handling evidential uncertainty and inconsistency in a mathematically tractable manner [14–16,32].

As a non-classical logical system, Logic Eτ is designed to support reasoning under uncertainty, contradictory, and incomplete information. Its expressive capability stems from the use of dual evidence degrees to express knowledge about a proposition enabling a granular representation of evidential states [14,18].

Logic Eτ assigns to each proposition p an evidential annotation (μ, λ), where μ and λ denote degrees of favorable and unfavorable evidence respectively. This dual-valued representation prevents trivialization in inference, even when μ and λ simultaneously assume high values, a condition under which Classical Logic becomes inconsistent and deductively trivial [14,32]. These evidential annotations are formally interpreted within three conceptual spaces [14,15,32], depicted in Figures 1a, 1b, 1c, each capturing a specific aspect of paraconsistent reasoning:



 doi:10.20944/preprints202512.1284.v1

(**a**) (**b**) (**c**)

Figure 1. Key concepts about visual decision states in Logic Eτ, adapted from [14,15]. (a) The Lattice τ with partial order, where classical logical states: True, False, Inconsistent, and Paracomplete, correspond to extremal vertices. (b) The USCP (Unit Square of the Cartesian Plane) provides a geometric representation of evidential states (μ, λ), highlighting both extreme and non-extreme (quasi) logical regions. (c) The logical lattice τ results from a nonlinear transformation T (μ, λ) = (μ − λ, μ + λ − 1), mapping evidential inputs into a plane where the horizontal axis encodes certainty (Gce) and the vertical axis uncertainty (Gct). This transformed space enables graded reasoning across nuanced logical states.

1. Lattice τ with Partial Order: This structure defines a complete lattice over the unit square $[0,1]^2$, where each pair (μ, λ) encodes the degrees of favorable and unfavorable evidence about a proposition. A partial order is defined by:

$$(\mu 1, \lambda 1) \leq (\mu 2, \lambda 2) \Leftrightarrow \mu 1 \leq \mu 2 \text{ and } \lambda 1 \geq \lambda 2$$

This order reflects evidential dominance and enables lattice-theoretic operations (infimum, supremum, neutral elements). A canonical negation operator ∼ (μ, λ) = (λ, μ) supports dual reasoning and contradiction handling. The evidential lattice serves as the operational substrate for all inference processes in Logic Eτ-based systems [14,15].

2. USCP (Unit Square of the Cartesian Plane), from a geometric standpoint, the evidential lattice can be visualized as a unit square of the Cartesian plane. Each evidential pair (μ, λ) corresponds to a point in this 2D unit square (Figure 1b), allowing for an intuitive representation of the underlying information state. While the lattice defines logical and computational operations through ordering, the USCP offers a descriptive and analytic space for visualizing evidential distributions and for mapping them onto the logical plane [14–16,32].
3. Diagram of Certainty and Contradiction Degrees: A nonlinear transformation maps USCP into the logical diagram, where inference operates in Figure 1c. The transformation defines two axes: the certainty degree (Gce), and the contradiction degree (Gct).

$$T(\mu, \lambda) = (G_{ce}(\mu, \lambda), G_{ct}(\mu, \lambda)) = (\mu - \lambda, \mu + \lambda - 1)$$

Extreme logical states (True, False, Inconsistent, Paracomplete) correspond to the four lattice extremities (1,0) → true, (−1,0) → false, (0,1) → inconsistent, (0, −1) → paracomplete (incomplete). Intermediate regions correspond to non-extreme states such as quasi-true, quasi-false, quasi-inconsistent and quasi-paracomplete and their respective tendencies, allowing graded reasoning, a crucial asset in non-deterministic conversational contexts as shown in Table 1 [14,15,32].

Table 1. Symbolic representation of extreme and non-extreme logical states in Logic Eτ, including quasi-states and transitional tendencies.

Symbol	State
V	True
QV→T	Quasi-true, tending to inconsistent;
QV→⊥	Quasi-true, tending to paracomplete
F	False
QF→T	Quasi-false, tending to inconsistent
QF→⊥	Quasi-false, tending to paracomplete
T	Inconsistent
QT→V	Quasi-inconsistent, tending to true
QT→F	Quasi-inconsistent, tending to false
⊥	Paracomplete or Indeterminate
Q⊥→V	Quasi-paracomplete, tending to true
Q⊥→F	Quasi-paracomplete, tending to false

Adapted from [15].

 doi:10.20944/preprints202512.1284.v1

The annotations support the deduction of both extreme and non-extreme logical states, including quasi-states and directional trends. Each of these logical outcomes serves as a semantic signal that guides the system's behavior, prompting clarification requests, refining domain classification, or flagging inadequate answers. This evidential logic framework introduces interpretability and resilience, avoiding reliance on brittle heuristics or handcrafted rules.

In Logic Eτ, the degree of certainty (Gce = $\mu - \lambda$) expresses the balance between supporting and opposing evidence, while the degree of uncertainty (Gco = $\mu + \lambda - 1$) indicates the extent to which such evidence is simultaneously conflicting.

Logic Eτ was applied as a reasoning component to handle contradictory and incomplete information. Rather than functioning as a predictive or statistical model, it operated as a non-classical framework that qualified evidential states and guided interpretation within the decision workflow.

As detailed in later sections, Logic Eτ underpins the system's core inferential mechanisms by enabling control decisions associated with propositions such as "The user question is clear", "The user question belongs to one of poultry production domains", or "The generated answer is adequate".

2.3. Design of Experiments for System-Level Parameter Tuning

In decision-oriented systems that demand precision, traceability, and trust, it is critical to address the limitations of large language models, particularly their non-deterministic behavior and susceptibility to hallucinations [8,11–13,19]. This study applied a Design of Experiments (DoE) approach to conduct a series of controlled tests, aiming to investigate how variations in system-level configurations affect the reliability, interpretability, and semantic accuracy of responses generated by the DSS architecture.

A controlled subset of the domain-specific knowledge base served as the foundation for the experiments. This corpus enabled the development of a fixed set of predefined queries; each paired with a gold-standard curated answer used as reference in the evaluation process. Two complementary metrics were analyzed. The first assessed system performance by measuring the semantic similarity (cosine similarity with SBERT embeddings) between the retrieved content and the reference answer, serving as a proxy for content fidelity and practical utility. The second examined the semantic alignment between the retrieved content and the original query, reflecting contextual coherence. While informative, this second metric does not guarantee factual correctness and may overvalue responses that are lexically aligned but semantically inaccurate or incomplete.

All experiments shared the same computational setup, including preprocessing libraries, LLM access, and vector-based retrieval infrastructure. The experimental dataset consisted of synthetic question–answer pairs generated from a curated knowledge base. This knowledge base was built from scientific literature, technical manuals, and extension materials covering poultry nutrition, welfare, housing, and husbandry. Inputs to the system are therefore natural-language queries, not raw sensor data or farm records. Full implementation details and software versions are provided in Section 2.4 (Reproducibility and Software Environment).

Five experiments investigated the chunking strategy, input preprocessing, and generation parameters:

1. Chunk Size and Overlap: In the RAG pipeline, chunk size refers to the number of tokens in each embedded segment, while overlap specifies the number of tokens repeated between adjacent chunks, directly affecting contextual continuity and information density. The interaction between these parameters affects retrieval precision, semantic cohesion, and computational efficiency [20].

The experiment utilized set of predefined question–answer pairs adopted across the other experiments and assessed both semantic alignment with the reference answer and contextual relevance to the original query. Three chunk sizes were tested: 128 tokens (high semantic precision, suitable for fine-grained reasoning), 256 tokens (practical optimum in most RAG pipelines), and 512 tokens (which maximizes cohesion in technical paragraphs). Overlap values included 32 tokens

 doi:10.20944/preprints202512.1284.v1

(minimal redundancy, avoiding abrupt cuts), 64 tokens (standard default, balances coherence and cost), and 128 tokens (high redundancy, beneficial for larger chunks but computationally heavier) [20,21]. A complete factorial design (3 × 3) was employed to investigate the combined effects of chunk size and overlap.

The objective was to identify optimal trade-offs between granularity and cohesion, determine points of diminishing semantic returns, and establish thresholds beyond which overlap increases computational cost without improving retrieval quality.

2. Lemmatization: This preprocessing step reduces inflected or derived words to their base form (lemma), preserving grammatical context and semantic identity. By mapping morphological variants to a unified lexical representation, it may reduce embedding dispersion and improve retrieval alignment [22,23].

Lemmatization was evaluated as a binary configuration: either applied or omitted symmetrically to both the indexed corpus and the user question–answer pairs. This experiment employed the chunking configuration identified in Experiment 1 and used the same set of predefined question–answer pairs, along with the evaluation criteria previously established.

The objective was to determine whether the inclusion of lemmatization improves semantic similarity to the reference answer and enhances contextual alignment with the original query.

3. Normalization: This preprocessing step standardizes both the domain-specific corpus and the question–answer pairs by reducing superficial variability that does not affect meaning. It directly influences lexical alignment, improves embedding consistency, and enhances semantic matching, particularly in architectures where token-level similarity governs access to relevant content [24].

Normalization was evaluated before vectorization as a binary configuration: either applied or omitted symmetrically to both the indexed corpus and the question–answer pairs used for evaluation. A complete 2^4 factorial design was used to test all possible combinations of four operations: lowercasing, punctuation removal, diacritic stripping, and whitespace collapsing.

The objective was to determine whether these steps, individually or in combination, enhanced retrieval quality in terms of semantic similarity and contextual relevance.

Synonym Expansion: This preprocessing strategy enriches the indexed corpus and the question–answer pairs by appending or substituting terms with semantically equivalent alternatives. It aims to mitigate vocabulary mismatches and improve alignment between the user formulation and the stored knowledge base [25]. Following established evidence in information retrieval [45], synonym expansion was applied to reduce vocabulary mismatch and increase recall. This was particularly effective in poultry-related contexts: for instance, queries with *'feed formulation'* improved retrieval when expanded with *'broiler diet'*, and *'temperature control'* benefited from the inclusion of *'thermal regulation'*. While this strategy increased coverage, it also introduced a small number of false positives (e.g., *'lighting program'* matched with *'lightweight'*), which we acknowledge as a trade-off in retrieval precision.

Synonym expansion was evaluated as a binary configuration: either applied or omitted symmetrically to both the indexed corpus and the question–answer pairs. Lexical resources, including the semantic lexicon WordNet and its multilingual extension OMW, were used to identify synonym candidates prior to vectorization.

The objective was to assess whether this strategy enhances retrieval performance, particularly in terms of semantic similarity to the reference answer, in scenarios where lexical variation might otherwise reduce retrieval effectiveness.

4. Temperature and Top-p: The foundation model parameters regulate the stochastic behavior of the language model during response generation. Temperature controls the entropy of the output distribution, modulating the balance between determinism and exploration [26,27]. Top-p (nucleus sampling) constrains the sampling space to the smallest set of tokens whose cumulative probability exceeds a given threshold, shaping the diversity and unpredictability of the generated text [26,27].

 doi:10.20944/preprints202512.1284.v1

This experiment employed the chunking configuration identified in Experiment 1 and utilized the same set of predefined question–answer pairs, along with the evaluation criteria previously established. This model generated responses across a parameter space that ranged from factual and deterministic completions to controlled interpretative outputs and exploratory generations. The tested values were temperature ∈ {0.0, 0.3, 0.6, 0.9} and top-p ∈ {0.8, 0.9, 1.0}. A complete 4 × 3 factorial design was employed to isolate the interaction effects of parameters within a realistic retrieval-augmented generation workflow. The tested ranges for temperature and top-p were informed by previous research on LLM generation parameters, which demonstrated that very low temperature values tend to produce deterministic and repetitive outputs, while very high values increase incoherence [11,46–48]. Similarly, top-p values between 0.6 and 1.0 have been widely adopted in foundational work to balance output diversity with factual reliability [46,48] (Brown et al., 2020). These ranges therefore represent established practice in controlled experiments with large language models.

The objective was to evaluate how different sampling configurations impact semantic fidelity to the reference answer and contextual relevance to the original query, while maintaining generation stability of generation and interpretability.

Collectively, the experiments provided the empirical foundation for configuring the conversational agent. The selected parameters were directly incorporated into the final architecture, ensuring that the system strikes a balance between semantic precision, contextual relevance, and computational efficiency under realistic decision-making conditions. All procedures described here were executed within a controlled and reproducible software environment (see Section 2.4 for details).

2.4. System Architecture

The Decision Support AI-Copilot for Poultry Farming is implemented as a functional decision-support system composed of integrated modules structured around a RAG architecture (Figure 2). The first module, the Knowledge Base Construction Pipeline (KB-CP), performs the collection, preprocessing, segmentation, embedding, and indexing of curated scientific and technical sources. The resulting repository, the Domain-Specific Knowledge Base (DS-KB), organizes poultry production knowledge by thematic domains to enable precise and semantically guided retrieval.

The second module, the Conversational Decision-Support Agent (C-DSS-A), operates as the reasoning core of the system. It manages query interpretation, evidence retrieval, answer generation, and logical evaluation by integrating a large language model (GPT-4o) with the DS-KB through the RAG pipeline and the Logic Eτ. The agent can be accessed through a variety of client applications—such as web chat interfaces, mobile apps, dashboards, or voice-assistant integrations—that instantiate the conversational layer of the DSS and serve as its user-facing interface. Together, these modules enable the system to provide explainable and evidence-qualified responses under conditions of informational incompleteness or contradiction.

The architecture integrates LLMs and RAG techniques with Logic Eτ to support decision-making across multiple technical domains in poultry production. Its structure allows for independent evaluation and fine-tuning of semantic retrieval, language generation, and paraconsistent reasoning.

The system employs GPT-4o as its core language model. GPT-4o was selected for its semantic precision, low latency, and cost-efficiency, which make it particularly suitable for domain-specific RAG applications [28]. At the time of implementation, it was the most recent publicly available model in the GPT-4-turbo family. Its extended context window (up to 128k tokens) enables the integration of long retrieved passages while maintaining stable performance, an essential requirement for logic-grounded decision support. No training or fine-tuning of GPT-4o was performed in this study.

The operational parameters, such as chunking configurations, preprocessing routines, and generation settings, were empirically defined through the controlled experiments described in Section 2.1.2. These tests guided the selection of configurations that optimize trade-offs between granularity and cohesion, improve semantic similarity (cosine similarity with SBERT embeddings)

 doi:10.20944/preprints202512.1284.v1

between generated responses and the knowledge base, and enhance retrieval quality in terms of both content fidelity and contextual relevance. The system was also tuned to enhance robustness under lexical variability, strengthen alignment with the indexed content, and ensure generation stability and interpretability of generation across decision-making scenarios.

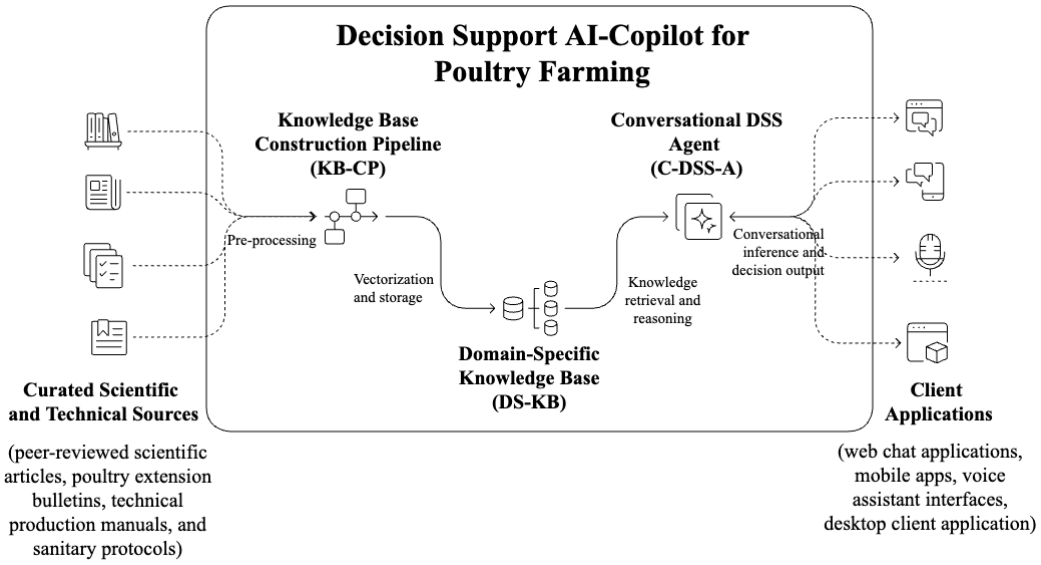


Figure 2. Modular architecture of the Decision Support AI-Copilot. The KB-CP preprocesses and embeds domain-specific content into the DS-KB. The C-DSS-A accesses this indexed repository to interpret user queries, retrieve relevant content, and generate logic-informed responses.

Semantic search is powered by FAISS (Facebook AI Similarity Search), selected for its scalability, support for both CPU and GPU backends, and proven efficiency in dense retrieval pipelines. The system utilized OpenAI's text-embedding-ada-002 model to encode knowledge base segments, and computes similarity via inner product (dot product), consistent with the model's scoring logic.

Vector indexing adopts the IndexFlatIP structure, a non-quantized flat index based on inner product similarity. This configuration ensures exact search results, which is crucial given the moderate scale of the dataset (fewer than 10,000 vectors) and the need for precise retrieval. The system performs retrieval via k-nearest neighbor (k-NN) search with a setting that balances contextual diversity with semantic relevance. Since latency is not a limiting factor in this application, exact k-NN was preferred to ensure retrieval fidelity and grounding quality in all downstream generations.

This architectural foundation supports the system's core functionalities and establishes the baseline over which configuration-level experiments (Section 2.1) were conducted to optimize performance and interpretability. Full details on code availability, software versions, and reproducibility protocols are provided in Section 2.4.

2.4.1. Knowledge Base Construction Pipeline (KB-CP)

To support domain-grounded retrieval and ensure high semantic precision during generation, the system relies on a knowledge base specifically constructed for poultry production decision-making. This repository was built through a structured pipeline comprising five main stages:

1. Document Collection: A sample of 48 technical documents were curated from authoritative sources, including peer-reviewed scientific articles, poultry extension bulletins, technical production manuals, and sanitary protocols. The selection prioritized content with high informational density, practical relevance, and clear domain affiliation. Documents were collected through targeted searches in scientific databases, institutional repositories, and validated extension services.

 doi:10.20944/preprints202512.1284.v1

Domain Classification: Each document was manually assigned to one of five predefined poultry production domains: (i) Housing and Environmental Control, (ii) Animal Nutrition, (iii) Poultry Health, (iv) Husbandry Practices, and (v) Animal Welfare.

These domains reflect core areas of technical decision-making in intensive poultry systems and are grounded in established animal welfare frameworks. The FAO's work on poultry welfare identifies health, nutrition, environmental comfort, and welfare as core aspects of assessment [29,30,42–44]. Classification was performed based on thematic focus, terminology patterns, and stated objectives of the material. In cases of overlap, domain assignment favored the dominant technical axis addressed by the document.

Preprocessing: All documents were converted to plain text and segmented into overlapping chunks, preserving local semantic cohesion. Chunk size and overlap were defined according to the optimal configuration identified in Experiment 1 (Section 2.1.2), which balances retrieval granularity with contextual integrity. This preprocessing step ensured that segment boundaries did not compromise sentence-level coherence, thereby improving embedding stability.

Vectorization: Each chunk was embedded using OpenAI's text-embedding-ada-002 model, producing dense vector representations in a high-dimensional semantic space. These embeddings captured contextual relationships at the subparagraph level, enabling fine-grained semantic retrieval aligned with user queries.

Domain-Based Indexing: For each knowledge domain, a separate FAISS index was created using the IndexFlatIP configuration (inner product similarity). This design supports fast and exact k-nearest neighbor (k-NN) search within each semantic repository. The use of independent indexes per domain facilitates targeted retrieval and minimizes semantic noise during generation.

The complete dataset, including raw documents, processed embeddings, and the full indexing pipeline, is publicly available via GitHub at [33].

2.4.2. Reasoning Workflow of the Conversational DSS Agent (C-DSS-A)

The C-DSS-A operates through a structured reasoning cycle (Figure 3) that integrates language comprehension, evidential assessment, semantic retrieval, and logical consistency checks.

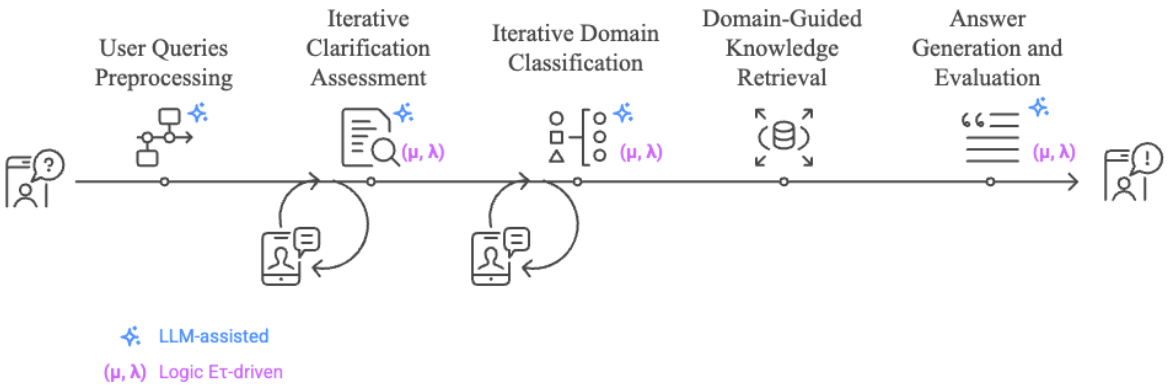


Figure 3. Workflow of the Conversational DSS Agent (C-DSS-A), detailing five sequential stages that combine LLM-based understanding with paraconsistent logic operations for query refinement, domain inference, knowledge retrieval, and evidence-grounded response generation.

Each decision stage is governed by a logic-based proposition evaluated under Logic Eτ. The complete workflow is composed of the following stages:

1. User Queries Preprocessing: User queries were preprocessed before both vector-based retrieval and language model inference. The adopted preprocessing configuration reflected the outcomes of controlled experiments. Synonym expansion was enabled as the only non-trivial

 doi:10.20944/preprints202512.1284.v1

transformation, selected for its capacity to bridge lexical gaps between user queries and indexed content. Lemmatization and punctuation removal were also applied, given their low computational cost and consistent contribution to lexical normalization. Conversely, diacritic stripping and whitespace collapsing were turned off by default, as their empirical impact on retrieval effectiveness proved negligible.

Iterative Clarification Assessment: Upon receiving a preprocessed user query, the system initiates an iterative process to evaluate and refine the clarity of the input. This is framed as the proposition:

$$P_1(\mu, \lambda)\text{: ``The user question is clear.''}$$

The annotation relies on a structured prompting protocol that infers evidential values directly from the LLM. Two specialized prompts quantify distinct epistemic dimensions: Clarity, defined as technical specificity and semantic coherence, and Vagueness, defined as conceptual ambiguity or logical imprecision. Both values are returned on a continuous scale from 0 to 1 and respectively, correspond to (μ, λ).

The Gce(μ, λ) determines whether the system has sufficient confidence to proceed. Following prior applications of Logic Eτ in expert systems, a conservative threshold of Gce ≥ 0.75 was adopted to prevent unstable classifications in quasi-state borderline regions of the USCP [14–16,18]. If Gce(μ, λ) < 0.75, the query is considered underdetermined. In such cases, the model generates a clarification prompt, which is appended to the conversational context. The revised input is re-evaluated using the same procedure, forming an iterative loop that continues until the certainty threshold is met (Gce(μ, λ) ≥ 0.75). At that point, the system proceeds to domain classification.

Iterative Domain Classification: Once the question is considered clear, the system prompts the LLM to classify it into one of five predefined poultry production domains: (i) housing and environmental control, (ii) animal nutrition, (iii) poultry health, (iv) husbandry practices, or (v) animal welfare. The classification is formalized as an annotated proposition:

$$P_2(\mu, \lambda) = \text{``The question pertains to [identified domain]''}.$$

As in the previous step, the evidential values μ and λ are inferred by the LLM through guided prompting and interpreted under Logic Eτ. If the resulting Gce(μ, λ) falls below 0.75, the system generates a meta-question to validate the classification (e.g., "Does your question relate to [suggested domain]?"). If the user confirms the domain, the classification is accepted and the system proceeds. If the user rejects it, the domain is removed from the candidate list, and the LLM is prompted again using the updated domain set. This loop continues until a confident domain assignment is achieved, enabling the system to advance to semantic evidence retrieval.

Domain-Guided Knowledge Retrieval: With a clarified question and an identified domain, the system proceeds to semantic retrieval. The input query is embedded using OpenAI's text-embedding-ada-002 model, and a k-NN search (k = 5) is performed in a FAISS vector index (IndexFlatIP with dot-product similarity) to retrieve the most relevant content chunks. Each passage is linked to its original source and metadata.

Answer Generation and Evaluation: The retrieved passages are concatenated with the clarified user query and submitted as the prompt context to GPT-4o (via OpenAI API). The model then generates a draft response. In parallel, it evaluates the annotated proposition:

$$P_3(\mu, \lambda) = \text{``The generated answer appropriately addresses the user's question.''}$$

As in previous stages, the values μ and λ are inferred through guided prompting and interpreted under Logic Eτ. The resulting Gce(μ, λ) reflects the system's internal confidence in the adequacy of the response. If Gce(μ, λ) < 0.75, the response is flagged as potentially unreliable and may be revised or explicitly marked with a disclaimer to inform the user of evidential insufficiency or contradiction. The evidential outputs produced in this stage are then passed to the logical evaluation module, detailed in the following section.

Section 2.4 provides a detailed account of the software stack, experimental environment, and reproducibility measures employed in this work.

 doi:10.20944/preprints202512.1284.v1

2.4.3. Reasoning Support with Logic Eτ

Each proposition formalizes a key decision point in the conversational reasoning cycle. Evidential values μ and λ are interpreted under Logic Eτ, and the resulting Gce(μ, λ) determines whether the system proceeds, flags the interaction, or initiates an iterative refinement. Thresholds and corresponding actions are defined to ensure interpretability, domain alignment, and response adequacy (Table 2)

Table 2. Annotated Propositions and Evidential Control Logic.

Proposition ID	Evaluated Statement	Purpose in System	Threshold (Gce)*	Action if Gce< Threshold	System Interaction Type
$P_1(\mu, \lambda)$	"The user question is clear."	Assess linguistic clarity; ensure interpretability	0.75	Trigger clarification question; append user response	Iterative clarification loop
$P_2(\mu, \lambda)$	"The question pertains to [identified domain]."	Classify query into production domain	0.75	Pose meta-question to user; eliminate rejected domain	Iterative domain pruning
$P_3(\mu, \lambda)$	"The generated answer appropriately addresses the user's question."	Assess adequacy and relevance of generated response	0.75	Flag response as uncertain; optionally trigger regeneration	Response flagging or regeneration

* *Threshold adopted to prevent quasi-state borderline regions of the USCP [14–16,18].*

2.5. Evaluation Protocol

The evaluation protocol focused on unit-level assessment of each reasoning stage within the C-DSS-A architecture and comprised three sets of tests, each designed to isolate and validate the behavior of individual components under controlled conditions. This approach enabled precise attribution of strengths and limitations at each stage of the decision workflow.

4. The test of the Iterative Clarification Assessment stage used a synthetic dataset of 130 user questions, generated from the DS-KB and labeled as Clear or Unclear. Each label encompassed a gradient of linguistic phenomena, including ambiguous phrasing, underspecified referents, non-technical constructions, and malformed syntax. Manual validation ensured internal consistency and class balance. The objective was to assess the system's ability to evaluate the proposition $P_1(\mu, \lambda)$: "The user question is clear", by discriminating underdetermined inputs based on evidential clarity rather than surface features. System performance was measured by its ability to converge to the correct classification through iterative reformulation, with convergence defined as Gce(μ, λ) ≥ 0.75 for proposition P_1.
5. The second test targeted the Iterative Domain Classification stage, using a new set of 100 synthetically generated questions, randomly assigned to one of the five defined domains, Housing and Environment, Animal Nutrition, Poultry Health, Husbandry Practices, Animal Welfare, or to no domain at all. This randomized distribution simulated open-query conditions. The dataset also included domainless questions to test rejection behavior under semantic uncertainty. The objective was to evaluate the system's ability to assess the proposition $P_2(\mu, \lambda)$: "The question belongs to [domain]", identifying the most appropriate category without forcing classification when evidential support was lacking. Classification was accepted only when the certainty degree satisfied Gce(μ, λ)≥0.75, ensuring evidential convergence before domain attribution.
6. The third test focused on the stages of Domain-Guided Knowledge Retrieval and Answer Generation and Evaluation, using 100 synthetically generated question–answer pairs curated

 doi:10.20944/preprints202512.1284.v1

from source articles in the DS-KB. Each question was processed under two conditions: first, through direct prompting without retrieval or evidential evaluation, and second, through the whole system workflow, which includes retrieval from the DS-KB, generative response, and Logic Eτ-based self-assessment. In the second condition, the system instructed the model to evaluate the adequacy of its answer using Logic Eτ, producing an evidential annotation for the proposition $P_3(\mu, \lambda)$: "The generated answer is adequate". This annotation served as a meta-level judgment of response quality. For both conditions, the generated answers were compared to gold-standard references using semantic similarity metrics (cosine similarity with SBERT embeddings). The objective was to assess whether the evidential reasoning introduced by Logic Eτ improves the system's capacity to generate semantically valid responses and enhances the interpretability and trustworthiness of the final output.

2.6. Reproducibility and Software Environment

All system developments, experimental procedures, and evaluation workflows were implemented and executed in a reproducible Python environment (v3.9.6) using Visual Studio Code. The implementation leveraged a modular architecture composed of tools for retrieval orchestration, language generation, vector search, preprocessing, and evaluation. The main libraries and frameworks include:

- Language modeling and embedding: Openai 1.95.1 (for GPT-4o and text-embedding-ada-002), tiktoken 0.9.0 (for token counting and window control).
- Retrieval and orchestration: faiss-cpu 1.11.0 (for dense vector search using IndexFlatIP), langchain 0.3.25 and related packages (langchain-core, langchain-openai, langchain-community, langchain-text-splitters, langchain-xai, langsmith) for chaining retrieval, embedding, and generation steps.
- Text preprocessing and NLP: spaCy 3.8.7 (for lemmatization and syntactic analysis), nltk 3.9.1 (for lexical resources and linguistic tagging), including downloads: punkt, wordnet, omw-1.4, averaged_perceptron_tagger, averaged_perceptron_tagger_eng, punkt_tab.
- Data analysis and visualization: Pandas 2.3.1 (for data manipulation), scikit-learn 1.6.1 (for combinatorial evaluation routines), matplotlib 3.10.3 and seaborn 0.13.2 (for results visualization).
- Auxiliary and system tools: python-dotenv 1.1.0, requests 2.32.3, aiohttp 3.12.2, httpx 0.28.1 (for API and system orchestration), tenacity, joblib, threadpoolctl, Django 5.0.4 (web framework), djangorestframework 3.15.1 (APIs RESTful), and tqdm (for robustness, parallelization, and progress monitoring).

All software dependencies are publicly listed and version-pinned in the project's requirements.txt file. The complete codebase, prompts, dataset, and reproducibility pipeline are available via GitHub at [33]. All the versions indicated are up to date and consistent with the releases available until July 2025.

2.7. GenAI Disclosure

Generative AI was employed for development support (integrated into Visual Studio Code), the construction of synthetic test data, and the refinement of analysis statements. Synthetic questions and answers used in unit tests and controlled experiments were generated from real source material, under strict semantic constraints and manually curated for consistency and domain alignment. All experimental evidence and analytical results derive exclusively from real system outputs or curated domain content.

3. Results

This section presents the results of a two-part evaluation. The first part presents controlled experiments assessing how system-level configurations influence response quality, interpretability,

 doi:10.20944/preprints202512.1284.v1

and semantic alignment. The second part analyzes the behavior of the complete system in end-to-end operation, with emphasis on evidential consistency and domain adequacy.

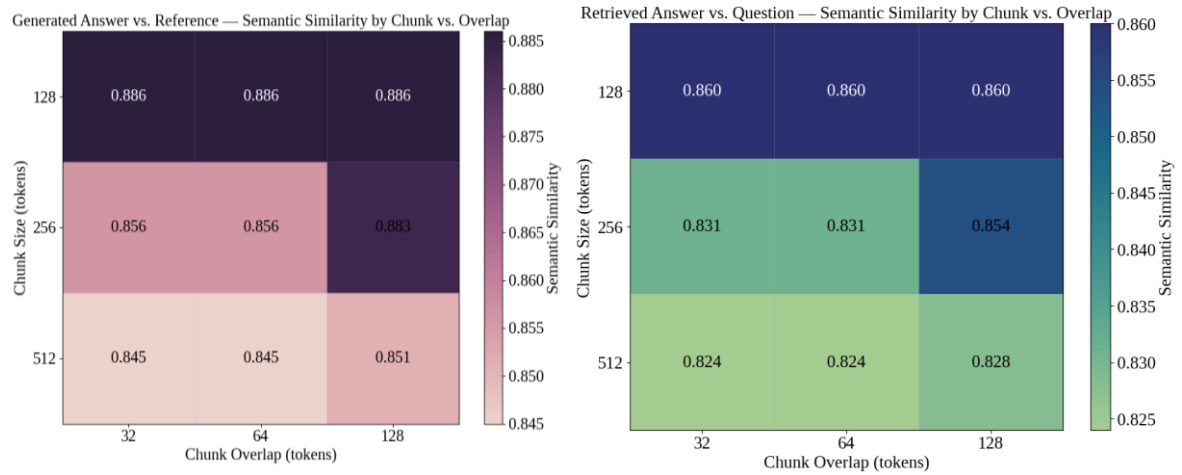
3.1. Experimental Results

The system was systematically tested through a design of experiments approach, comprising multiple controlled evaluations of preprocessing, retrieval, and generation parameters.

The controlled experiments were conducted as proof-of-concept to evaluate the impact of variations in chunking strategy, input preprocessing, and generation parameters impact the quality of retrieval and response formulation. Each test isolates a specific configuration variable, allowing for a precise assessment of its impact on semantic accuracy, contextual relevance, and output stability.

3.1.1. Effects of Chunk Size and Overlap on Retrieval Quality

The experiment evaluated the impact of chunk size and overlap on retrieval quality in the RAG workflow (Figures 4a and 4b). With 128-token chunks, semantic similarity averaged ≈ 0.886 (RA vs. R) and ≈ 0.860 (RA vs. Q) across all overlap levels. For 256-token chunks, overlaps of 32 and 64 tokens resulted in lower similarity values, while an overlap of 128 tokens restored similarity to ≈ 0.883 (RA vs. R) and ≈ 0.854 (RA vs. Q). For 512-token chunks, similarity remained consistently lower across both metrics, even at maximum overlap.



(**a**) (**b**)

Figure 4. Heatmaps showing the average semantic similarity (cosine similarity with SBERT embeddings) between (a) retrieved answers (RA) and gold-standard reference (R) and (b) retrieved answers (RA) and questions (Q), measured across different chunk and overlap configurations. The experiment corresponds to the parameter tuning analysis described. Each cell represents the mean similarity score for a given combination, using a controlled query set and vector retrieval via FAISS. Higher scores indicate a greater alignment between retrieved content and the expert reference.

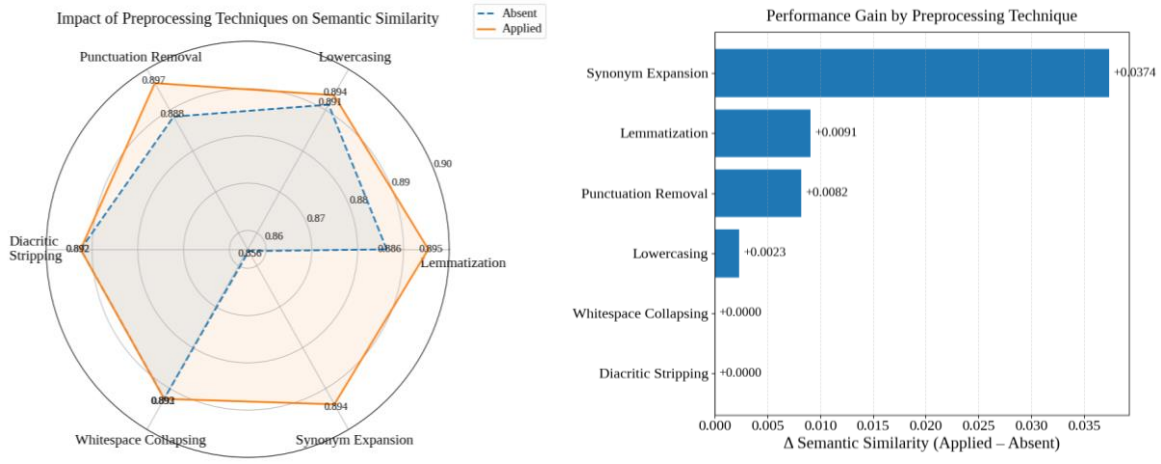
Based on these results, the system was configured with a chunk size of 128 tokens and an overlap of 32 tokens, as this setting provided stable similarity scores with minimal redundancy. The similarity plateau observed (around ~0.865) should be interpreted as an internal performance reference that guided parameter tuning in this proof-of-concept study, rather than as an optimal or benchmark value.

3.1.2. Effects of Preprocessing on Retrieval Quality

The experiment assessed the isolated impact of standard preprocessing techniques on semantic similarity (cosine similarity with SBERT embeddings) between retrieved answers and gold-standard references, while keeping the retrieval architecture constant (Figures 5a and 5b). Synonym expansion produced the highest gain, increasing similarity by +0.0374, the largest delta among all

 doi:10.20944/preprints202512.1284.v1

transformations. Lemmatization and punctuation removal also showed positive contributions (+0.0091 and +0.0082, respectively). Lowercasing and whitespace collapsing yielded marginal improvements (< +0.002), while diacritic stripping showed no measurable effect. None of the tested techniques decreased retrieval quality.



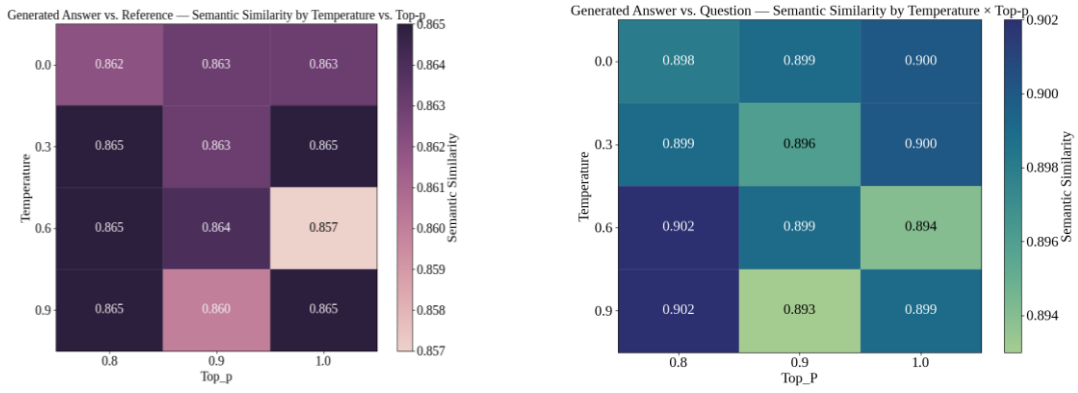
(**a**) (**b**)

Figure 5. (a) Radar chart comparing semantic similarity (cosine similarity with SBERT embeddings) between retrieved answers and reference answers, with and without each preprocessing technique. The orange contour represents performance with preprocessing applied; the dashed blue line corresponds to the baseline (absent). (b) Delta plot showing absolute performance gain (Δ similarity) for each technique, ordered from highest to lowest. Positive deltas indicate increased semantic alignment after applying the corresponding transformation.

Based on these results, the system was configured to enable synonym expansion as the only non-trivial transformation. Lemmatization and punctuation removal were also adopted due to their consistent but lightweight benefits, while diacritic stripping and whitespace collapsing were turned off by default.

3.1.3. Effects of Temperature and Top-p on Response Quality

The experiment examined the impact of sampling temperature and top-p on the semantic similarity (cosine similarity with SBERT embeddings) of generated answers, measured both against gold-standard references and the original user question (Figures 6a and 6b). Across the entire grid, response similarity to the reference (GA vs. R) remained stable, with most configurations converging around ≈ 0.865. The lowest value was observed at (temperature = 0.6, top-p = 1.0), where similarity decreased to 0.857. In contrast, alignment with the original question (GA vs. Q) varied more. The highest similarity (0.902) occurred at both (0.6, 0.8) and (0.9, 0.8). Configurations with top-p = 1.0 showed lower contextual alignment, with scores around 0.894–0.899.



(**a**) (**b**)

 doi:10.20944/preprints202512.1284.v1

Figure 6. Heatmaps showing the average semantic similarity (cosine similarity with SBERT embeddings) between (a) generated answers (GA) and gold-standard reference (R), and (b) generated answers (GA) and original questions (Q), across different combinations of temperature and top-p values. Each cell indicates the mean similarity score for a fixed (temperature, top-p) configuration, based on cosine distance between sentence embeddings. Higher scores represent greater semantic alignment.

Based on these results, the system was configured with temperature = 0.6 and top-p = 0.8, as this setting achieved the best overall contextual alignment (0.902 GA vs. Q) while maintaining high factual similarity (0.865 GA vs. R).

3.2. Evaluation of Conversational DSS Agent Workflow Stages

The results of the stage-specific evaluations of the C-DSS-A enabled a fine-grained analysis of performance, robustness, and evidential behavior across the decision workflow.

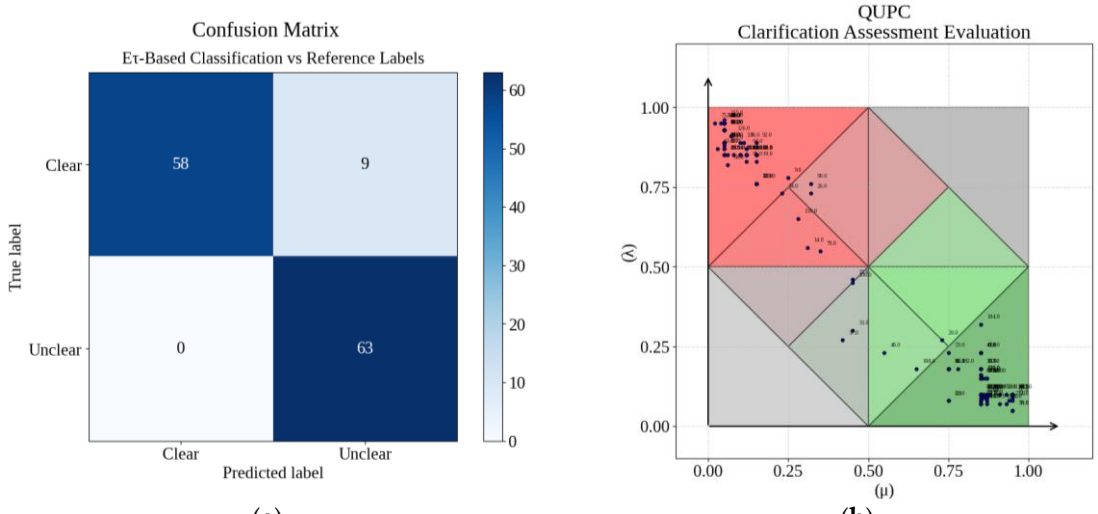
3.2.1. Results for the Iterative Clarification Assessment Stage

The classification of the proposition *"The user question is clear"* achieved high overall performance (Table 3), with accuracy of 0.931, macro-averaged F1 of 0.931, and substantial agreement across both classes. Precision for Clear was 1.000, while recall was 0.866. For Unclear, recall reached 1.000 with precision of 0.875. These values indicate that the system consistently recognized unclear queries, while occasionally misclassifying clear queries as ambiguous.

Table 3. Performance Metrics for Clarity Classification stage (Clear vs. Unclear).

	Precision	Recall	F1-score	Support
Clear	1.000	0.866	0.928	67
Unclear	0.875	1.000	0.933	63
Accuracy			0.931	130
Macro avg	0.938	0.933	0.931	130
Weighted avg	0.939	0.931	0.931	130

The confusion matrix (Figure 7a) shows that most errors occurred when Clear queries were labeled as Unclear (9 instances), while no Unclear queries were misclassified as Clear. The USCP projection (Figure 7b) illustrates that Clear queries concentrated in regions of high μ and low λ, while Unclear queries clustered in areas of higher λ values, particularly near inconsistent or paracomplete regions.



(**a**) (**b**)

 doi:10.20944/preprints202512.1284.v1

Figure 7. (a) Confusion matrix comparing the system's final classification of the proposition "The user question is clear" against the ground truth labels (Clear vs. Unclear). The matrix reflects the asymmetric robustness of the model. The observed misclassifications reveal a conservative tendency, with a stricter evidential threshold for confirming clarity. (b) Two-dimensional representation of the evidential annotations (μ, λ) in the QUPC, illustrating the distribution of outputs from the Iterative Clarification Assessment module. Points are colored according to their final classification and plotted against the paraconsistent decision regions derived from Logic Eτ. The separation between classes and the concentration near τ-lattice diagonals highlight the model's discriminative sensitivity to varying degrees of clarity and vagueness.

Based on these results, the module was retained with its default evidential thresholds (Gce ≥ 0.75), as this setting balanced precision for *Clear* with maximal recall for Unclear, ensuring reliable detection of underdetermined inputs.

3.2.2. Results for the Domain Classification Stage

The classification of the proposition *"The question belongs to [domain]"* showed variation across categories (Table 4). Animal Nutrition achieved precision of 0.818 and recall of 1.000 (F1 = 0.900). Animal Welfare reached precision of 0.600 and recall of 0.947 (F1 = 0.735). Housing and Environment obtained precision of 0.762 and recall of 0.889 (F1 = 0.821). Poultry Health registered precision and recall of 0.813 each (F1 = 0.813).

No instances of the domain "Husbandry Practices" were available in the test set; this class was therefore excluded from Table 4, as no meaningful metrics could be computed. The absence of this domain results from corpus imbalance, since the curated knowledge base and question set contained fewer examples related to husbandry practices compared with other domains such as nutrition, welfare, and housing. This condition was anticipated under the experimental design and does not represent a technical limitation but rather an opportunity for future corpus expansion.

The class labeled "Out-of-scope queries" corresponds to test inputs that could not be associated with any of the predefined domains. This behavior was intentional, as the system was designed to detect and explicitly signal questions outside its modeled scope rather than forcing uncertain assignments. The explicit recognition of out-of-scope queries reinforces the robustness and transparency of the classification process, reflecting realistic operating conditions in which users may submit questions beyond the coverage of the retrieved knowledge context. For this class, precision was 1.000 and recall 0.800 (F1 = 0.889). Overall accuracy was 0.737, with a macro-averaged F1 of 0.693.

Table 4. Performance metrics for domain classification stage.

	Precision	Recall	F1-score	Support
Animal Nutrition	0.818	1.000	0.900	18
Animal Welfare	0.600	0.947	0.735	19
Housing and Environment	0.762	0.889	0.821	18
Poultry Health	0.813	0.813	0.813	16
Out-of-scope queries [2]	1.000	0.800	0.889	10
Accuracy			0.737	99
Macro avg	0.665	0.741	0.693	99
Weighted avg	0.635	0.737	0.675	99

[1] No instances of the domain *Husbandry Practices* were present in the test set; therefore, no metrics were computed for this class. [2] *Out-of-scope queries* denotes test inputs that did not correspond to any predefined domain and were intentionally flagged by the system.

The confusion matrix (Figure 8a) shows that most errors involved confusion between Husbandry Practices and semantically adjacent categories (Animal Welfare and Poultry Health). The USCP projection (Figure 8b) indicates that most accepted classifications occupied regions of high certainty ($\mu > 0.75$, $\lambda < 0.25$), while a few borderline cases appeared near decision boundaries.

 doi:10.20944/preprints202512.1284.v1

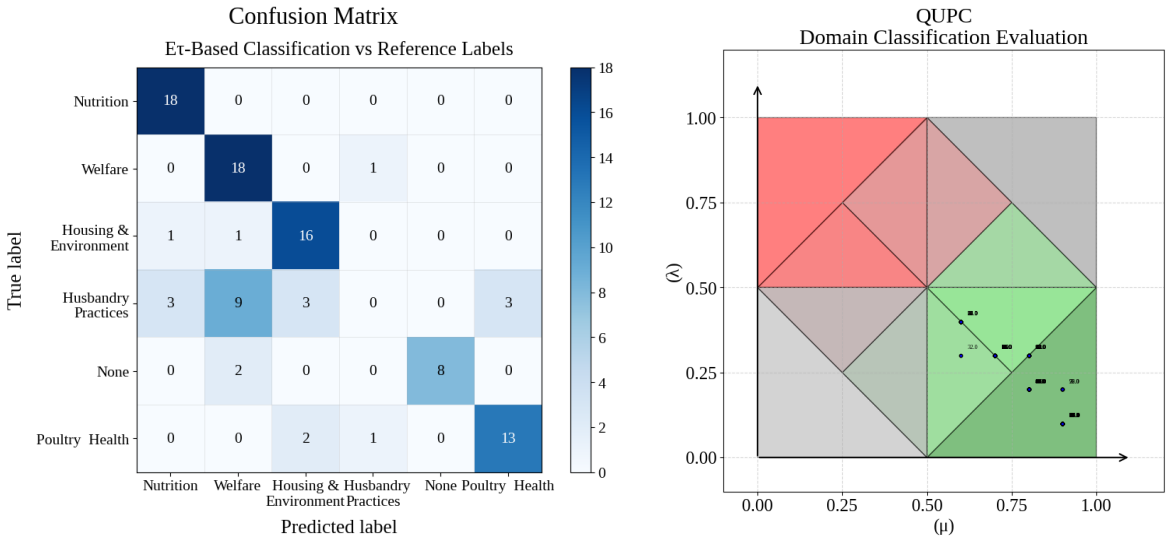


Figure 8. (a) Confusion matrix comparing predicted and reference labels for the proposition "The question belongs to [domain]", across five poultry production domains and an out-of-domain class (None). Misclassifications are concentrated in semantically adjacent categories, particularly in *Husbandry Practices*. Correct abstentions in the *None* class confirm the system's capacity to reject uncertain assignments when evidential support is insufficient. (b) USCP projection of the evidential annotations (μ, λ) associated with domain classification decisions. Most points fall within regions of high certainty and low contradiction, consistent with valid assignments. The sparse activation near decision boundaries highlights borderline cases, suggesting residual ambiguity in specific domain transitions.

Based on these results, the module was retained with the same evidential threshold (Gce ≥ 0.75), as this configuration supported robust rejection of out-of-domain queries while maintaining reliable classification for most domains.

3.2.3. Results for the Domain-Guided Knowledge Retrieval and Answer Generation and Evaluation Stages

The comparative evaluation between direct and system-guided queries showed consistent differences in semantic similarity (cosine similarity with SBERT embeddings) as shown in Figures 9a and 9b. System-mediated responses displayed a higher median similarity and reduced variance compared to direct queries, with fewer low-similarity outliers. The similarity curve (Figure 10) showed that system queries concentrated around higher similarity values, with a sharper peak near 0.97.

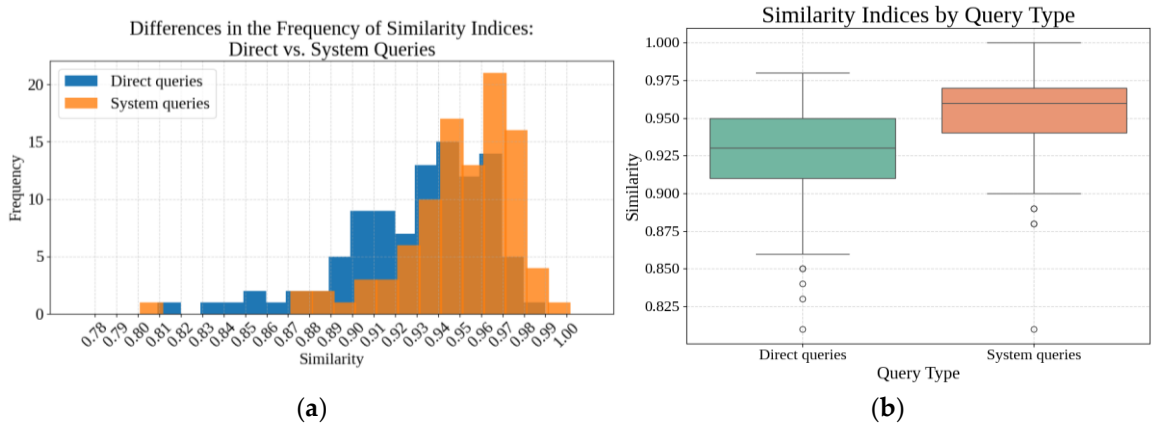


Figure 9. (a) Histogram of similarity scores comparing direct queries and system-mediated queries against the gold-standard answers. System responses exhibit a higher concentration of high-similarity outputs, with a clear

 doi:10.20944/preprints202512.1284.v1

rightward shift in distribution. (b) Boxplot summarizing similarity score distributions for each query type. System-mediated queries show higher median similarity and reduced variance, indicating more consistent semantic alignment.

The similarity distribution and polynomial trend curves (Figures 10 and 11) showed consistent differences across configurations. System-mediated queries concentrated around higher similarity values, with a sharper peak near 0.97, while RAG-only queries also shifted the distribution toward higher alignment compared to the LLM-only baseline. In contrast, the LLM-only setting displayed a flatter and more dispersed distribution, with a larger proportion of low-similarity outputs. The sharper peaks for both RAG and System-mediated configurations indicate a higher concentration of well-aligned responses and reduced variance, highlighting the stabilizing effect of knowledge retrieval.

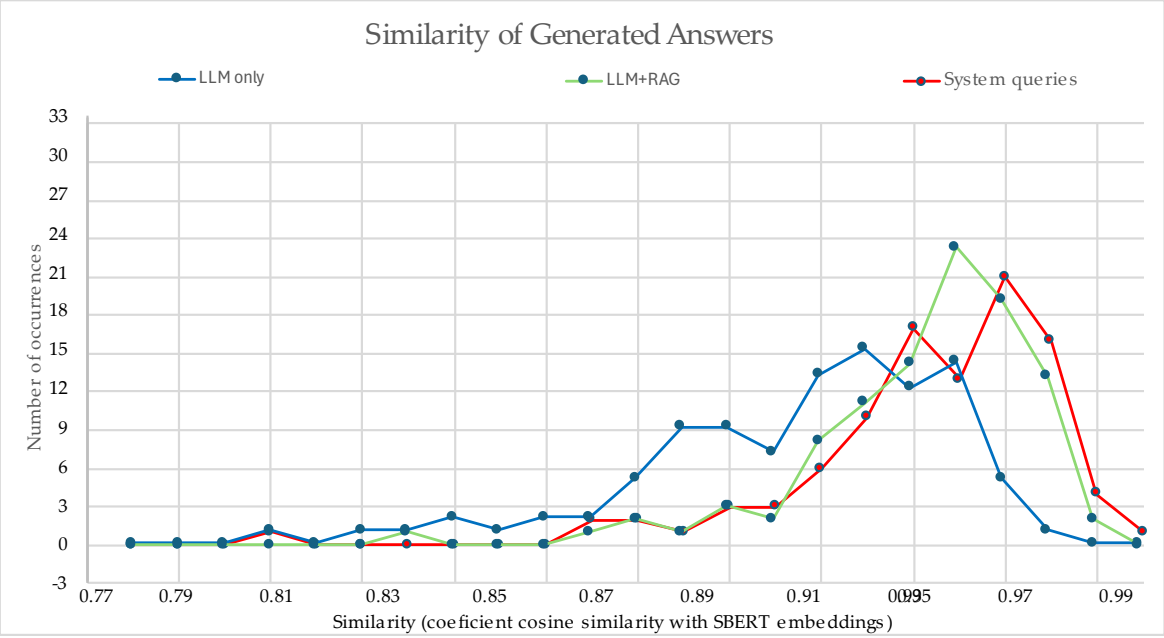


Figure 10. Distribution of semantic similarity coefficients (cosine similarity with SBERT embeddings) between generated answers and gold-standard references, across three configurations: LLM only (blue), LLM+RAG (green), and System-mediated with Logic Eτ (red).

The polynomial curves further emphasize these overall distribution patterns, making the relative gains in similarity concentration and reliability more evident when evidential reasoning is incorporated

Evidential judgments (Figures 12a and 12b) showed that 95% of cases fell in high-certainty regions ($\mu > 0.75$, $\lambda < 0.25$), with most outputs classified as True (V). A minority were assigned to False (2%), Quasi-false tending to Paracomplete (1%), and Quasi-true tending to Paracomplete (2%). Approximately 5% of the outputs appeared near contradictory regions, corresponding to marginally lower similarity scores.

Table 5 provides illustrative cases where the system flagged generated answers as inadequate, showing the corresponding user queries, excerpts of the outputs, evidential judgments (μ, λ, and Gce), and the user-facing messages that communicate uncertainty in a neutral and supportive tone.

Based on these results, the system was configured to operate with retrieval and Logic Eτ-based self-assessment enabled by default, as this combination consistently improved similarity alignment and provided evidential annotations for reliability control.

 doi:10.20944/preprints202512.1284.v1

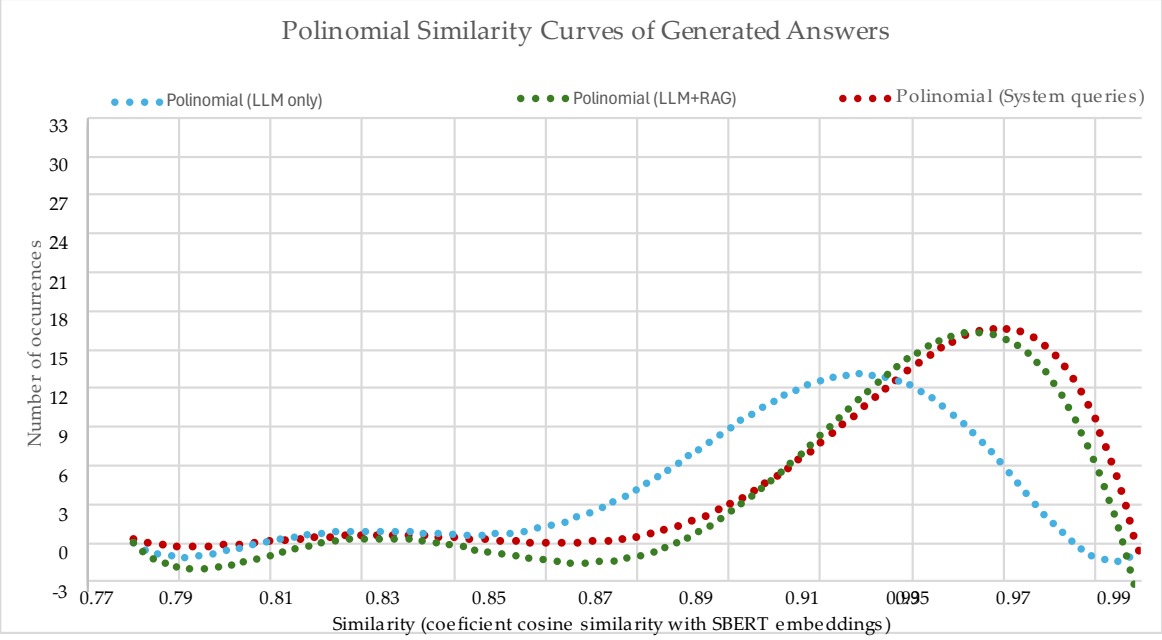


Figure 11. Polynomial trend curves of semantic similarity (cosine similarity with SBERT embeddings) between generated answers and gold-standard references, comparing three configurations: LLM only (blue), LLM+RAG (green), and System-mediated with Logic Eτ (red).

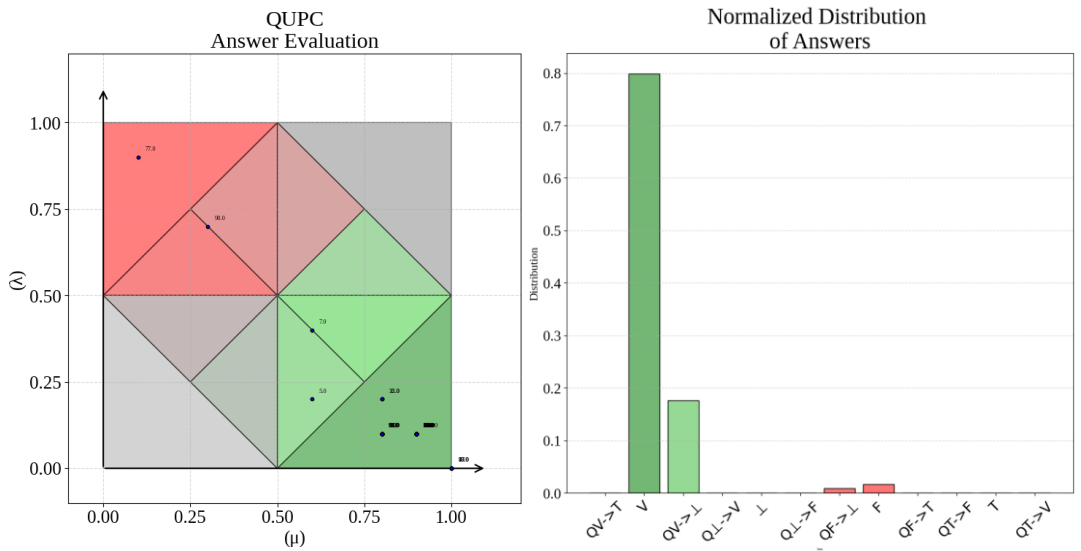


Figure 12. a) USCP projection of the evidential annotations (μ, λ) for the proposition "The generated answer is adequate." Most points concentrate in regions of high certainty and low contradiction, reflecting confident adequacy judgments. Sparse activation in quasi-inconsistent zones highlights borderline or semantically ambiguous responses. (b) Normalized distribution of the resulting logical states, with 95% classified as True (V), and a minority assigned to False (F, 2%), Quasi-false tending to paracomplete (QF→⊥, 1%), and Quasi-true tending to paracomplete (QV→⊥, 2%). The distribution reinforces the predominance of reliable outputs and the system's ability to flag marginal cases with non-extreme logical states.

 doi:10.20944/preprints202512.1284.v1

Table 5. Illustrative Cases of Evidence-Based Inadequacy Judgments.

User Query	Generated Answer (excerpt)	Evidential Judgment	User-Facing Message
What is the recommended broiler diet for heat stress?	Conflicting guidelines were retrieved, some emphasizing increased electrolytes, others focusing on energy adjustment	$\mu = 0.61$, $\lambda = 0.12$ Gce = 0.49 (Inadequate)	"Some retrieved information appears inconsistent. The answer may need clarification. Please consider refining your query or reviewing the supporting evidence."
What is the optimal temperature for broiler housing at 21 days of age?	Broilers at 21 days should be kept at 28 °C; some sources also mention 24–26 °C depending on ventilation	$\mu = 0.64$, $\lambda = 0.39$ Gce = 0.25 (Inadequate)	"Retrieved guidelines vary across sources. Please consider reviewing the suggested ranges or providing more context for your query."
How often should litter be replaced in broiler houses?	Some sources recommend complete replacement each cycle, others suggest partial reuse if treated with drying agents	$\mu = 0.78$, $\lambda = 0.17$ Gce = 0.61 (Inadequate)	"The retrieved evidence shows differing recommendations. The answer may depend on farm conditions—please review the supporting guidelines."
Is vaccination against coccidiosis always required in broilers?	Most sources recommend vaccination for long-cycle broilers; some mention prophylaxis may suffice under high biosecurity.	$\mu = 0.62$, $\lambda = 0.14$ Gce = 0.48 (Inadequate)	"Evidence for this query is partly inconsistent. The system combined both vaccination and prophylaxis approaches—please interpret according to your production context"

4. Discussion

This section examines the findings from two complementary dimensions: system-level tuning experiments that revealed how configuration choices condition retrieval and generation, and stage-wise evaluations that demonstrated how evidential mechanisms regulate reasoning across modules. Considered together, these findings outline an integrative perspective on how technical optimization and modular inference interact in shaping the overall behavior of the DSS.

4.1. Implications of System-Level Parameter Tuning Experiments

The experiments conducted as proof-of-concept trials, provided evidence on how system-level parameters influenced retrieval fidelity, contextual alignment, and the stability of generative outputs. These results revealed trade-offs that guided the technical configuration of the DSS and offered methodological insights for the design of RAG pipelines.

 doi:10.20944/preprints202512.1284.v1

The experiments on chunk size and overlap indicate that smaller segments tend to preserve semantic integrity, reducing the need for redundant overlap. At 256 tokens, additional overlap was required to restore retrieval quality, suggesting the presence of a threshold below which segment boundaries start to fragment contextual cohesion. Very large chunks of 512 tokens, even with high overlap, showed limited benefit, indicating the constraints of relying on size alone to secure retrieval accuracy. These findings highlight that segment configuration is not a neutral choice but a determinant factor for balancing semantic precision, contextual cohesion, and computational cost.

The analysis of preprocessing strategies shows that synonym expansion proved particularly influential in bridging lexical gaps between queries and knowledge base content, strengthening retrieval precision through semantic diversification. Lemmatization and punctuation removal also proved beneficial, though to a lesser extent, by reducing morphological variability and surface noise that can obscure semantic matches. By contrast, lowercasing, whitespace collapsing, and diacritic stripping offered negligible contributions, indicating that common normalization routines may be redundant in embedding spaces that already capture semantic robustness. The overall absence of negative effects suggests that preprocessing choices can be selectively applied, with meaningful gains concentrated in a few targeted transformations rather than in broad, indiscriminate normalization.

The evaluation of generation parameters indicated that factual accuracy remained relatively stable across configurations, while contextual alignment was more sensitive to variation. The performance drop observed at mid-level temperature combined with unfiltered sampling (0.6, 1.0) reflected a weakening of grounding when lexical openness was high. In contrast, settings that paired moderate or high diversity with a constrained nucleus, such as (0.6, 0.8) and (0.9, 0.8), improved relevance to user queries without compromising fidelity to reference answers. Configurations with top-p = 1.0 confirmed the risk of topic drift, as unrestricted sampling introduced variability that diluted semantic focus. These patterns suggest that balanced decoding strategies, exemplified by the adopted profile of temperature = 0.6 and top-p = 0.8, provide a practical equilibrium between determinism and contextual flexibility in domain-constrained tasks.

As proof-of-concept, these experiments suggested key trade-offs and helped establish a preliminary foundation for subsequent system evaluations. They delineate a methodological basis from which broader validations and domain-specific extensions can be pursued, emphasizing that parameter tuning is not a peripheral adjustment but a defining step that conditions the robustness of the DSS.

4.2. Evidential Reasoning in Stage-Wise System Evaluations

Beyond parameter tuning, the stage-wise evaluation of the C-DSS-A showed how evidential reasoning shaped system behavior across successive modules, revealing distinct patterns of performance and uncertainty management. At the same time, the agent as a whole exhibited a coherent operational profile, with all modules displaying controlled behavior under uncertainty. Logic Eτ provided the continuous interpretive structure that supported query clarification, domain attribution, and answer validation across stages [14,32].

The system demonstrated high sensitivity to semantic underdetermination during question clarification. Its conservative bias toward flagging inputs as ‘Unclear’ was not merely a trade-off, but an intentional mechanism for ambiguity control. By embedding classification within a Logic Eτ approach, the system preserved the evidential structure of borderline cases, avoiding premature resolution. This behavior prevented uncertain queries from propagating unchecked into subsequent inference stages, ensuring that downstream modules operated on well-formed and contextually interpretable inputs, thereby reducing the risk of hallucinations and erratic responses [11–14]. However, this imbalance between precision and recall in the Clear class also suggests that some genuine queries may elicit unnecessary clarification prompts. While maximizing recall for Unclear inputs ensures robust detection of ambiguity, it introduces a usability concern: repeated clarification loops could affect user perception of responsiveness, potentially causing frustration when clear questions are unnecessarily flagged. To mitigate this, future work will explore adaptive calibration

 doi:10.20944/preprints202512.1284.v1

of evidential thresholds and dynamic adjustment strategies to minimize superfluous prompts while maintaining reliability in detecting truly ambiguous queries.

In domain classification, the system exhibited variable performance across categories. Domains such as *Animal Nutrition* and *Housing and Environment* were consistently well distinguished. No instances of *Husbandry Practices* were observed during testing, a limitation that reflects corpus imbalance rather than a modeling issue. Consequently, questions from this domain were often absorbed by semantically neighboring categories. This pattern highlights intrinsic challenges in domain categorization and likely reflects both the underrepresentation of *Husbandry Practices* in the corpus and its semantic proximity to related domains such as *Animal Welfare* and *Poultry Health*, resulting in boundary errors. The absence of inter-rater validation may also have introduced annotation noise, underscoring the need for corpus diversification and multi-annotator validation to strengthen domain coverage and reliability. Despite these issues, the system demonstrated robust rejection behavior for queries outside predefined domains, with the evidential threshold effectively blocking premature assignments. The emergence of the "*Out-of-scope queries*" category further illustrates the system's ability to detect and signal questions beyond its modeled knowledge boundaries, ensuring transparency and preventing forced assignments under uncertainty. This selective restraint is particularly relevant for LLM-based architectures [36], which often default to overgeneralization in the face of ambiguity [8,13].

The comparative distributions further illustrate the incremental contribution of retrieval and evidential reasoning to answer quality. The flatter and more dispersed curve of the LLM-only baseline reflects a higher incidence of poorly aligned responses, whereas the inclusion of RAG concentrated outputs closer to the reference. The system-mediated configuration, which integrates Logic Eτ, sharpened this peak, reducing variance and reinforcing reliability. Although these differences are visually clear, their statistical significance was not formally tested, and the trends should therefore be interpreted with caution.

Ultimately, the system demonstrated that evidential inference plays a central role in enhancing response reliability. When retrieval and Logic Eτ-based self-assessment were active in the system workflow, responses shifted toward higher semantic alignment with reference answers and exhibited reduced variance. The analysis showed that most outputs were evaluated as highly adequate under the proposition P_3, while borderline or contradictory cases were correctly flagged through hesitant or inconsistent annotations. This evidential trace, absent in standard generation workflows, introduced a meta-level of interpretability that reinforced the trustworthiness of the final output [8,9,14,15].

Taken together, these results demonstrate that Logic Eτ is not merely an explanatory overlay but a functional mechanism that modulates the system's behavior. It enables abstention, detects vagueness, calibrates domain attribution, and qualifies the generative output, all within a consistent inferential framework. Importantly, it achieves these functions without relying on heuristics or hard-coded decision trees [14–16,32,36,37].

From a practical standpoint, this evidential strictness positions the system for real-world deployment in complex poultry production contexts [36,37]. It is relevant for high-stakes scenarios where interpretive caution, traceable reasoning, and the ability to admit uncertainty are critical requirements [2–5]

These findings reinforce the working hypothesis that integrating Logic Eτ with LLM and RAG architectures enhances reliability and interpretability in decision-support scenarios. By enabling structured self-assessment and evidential control, the system addresses key limitations of previous approaches, particularly their brittleness in the face ambiguity or contradiction [8,11–13]. The observed behaviors align with prior research on paraconsistent reasoning in uncertain environments [32], while extending its application to high-level semantic workflows.

 doi:10.20944/preprints202512.1284.v1

4.3. Integrative Perspective: From Parameter Tuning to System Behavior

Taken together, the results from the tuning experiments and the stage-wise evaluations converge on a complementary view: parameter tuning defined the conditions for stable and precise retrieval and generation, while Logic Eτ qualified and constrained outputs under uncertainty. This integration illustrates that robustness emerges not from isolated components, but from their interaction. It also underscores that technical optimization alone is insufficient; without evidential reasoning, the system would remain vulnerable to contradiction, while Logic Eτ itself depends on a calibrated retrieval pipeline.

In this respect, the proposed framework diverges from recent LLM+RAG approaches. While those systems improve contextual reasoning and factual grounding, they still rely on classical logic assumptions and lack mechanisms to formally manage contradictory or incomplete evidence [8,11–13]. By adding Logic Eτ, the present architecture introduces explicit evidential quantification and structured self-assessment, enabling the system to qualify ambiguous inputs and outputs rather than force binary resolutions [14–16,18]. This contrast clarifies that the contribution of the study is not a replacement of existing RAG pipelines, but their extension with contradiction-tolerant reasoning, oriented toward explainable and trustworthy decision support in poultry production.

4.4. Limitations and Future Work

While the system exhibited coherent and controlled behavior across all inference stages, some important limitations must be acknowledged.

The experimental design relied on a restricted sample size and a limited evaluation scope, which constrains the transferability of the findings. These experiments should therefore be regarded as proof-of-concept trials to test the feasibility of combining LLMs and RAG with Logic Eτ under controlled conditions, rather than as large-scale validation. In particular, the evaluation of chunk size and overlap was restricted to a narrow set of parameter configurations and metrics, without validation across different poultry production domains. Future work will expand the parameter range and validate chunking strategies in multiple domains to improve robustness and generalization.

Another limitation concerns the representativeness of the knowledge base used in the RAG pipeline. The current corpus was intentionally restricted to a small and homogeneous set of documents, which enabled controlled testing but limited the diversity of contexts and production scenarios represented. This restriction may affect retrieval performance and response reliability. In addition, domain classification was conducted by a single annotator without inter-rater validation, and cross-domain materials were reduced to a dominant category. The block segmentation strategy, optimized with synthetic QA data, has also not yet been validated for semantic integrity in long technical documents. Future work should expand the knowledge base with more diverse sources including manuals, scientific literature, regulatory guidelines, and field reports, while ensuring balance across document types, timeframes, and regions. Multi-annotator validation and extended segmentation assessments will also be incorporated to strengthen corpus reliability and semantic fidelity.

The system has not yet been validated by domain experts nor tested in real production environments. The current implementation should therefore be regarded as an early-stage prototype, evaluated under controlled conditions. Next steps should prioritize participatory assessments with poultry specialists and in-situ deployments to assess usability, robustness, and contextual adaptability. From an operational perspective, deployment feasibility is constrained by available computing resources. Performance testing, including stress scenarios, will be required to validate scalability, and farms with limited infrastructure may still require hybrid cloud-based solutions. Large-scale deployment will also require optimization to support thousands of concurrent queries and integration with sensor-driven alarm systems, ensuring timely responses in intensive production environments. Such evaluations are crucial for consolidating operational maturity and refining evidential thresholds in response to practical decision-making demands. In parallel, future work may

 doi:10.20944/preprints202512.1284.v1

explore adaptive calibration strategies for domain boundaries and clarity thresholds, as well as pathways for generalization beyond poultry production.

Finally, although evidential reasoning improved robustness against uncertainty and contradictions, the system has not yet been stress-tested under adversarial, hostile, or multilingual queries. These scenarios represent potential points of failure and should be included in future evaluations to better characterize resilience beyond the controlled settings adopted here. Future work will also include comprehensive comparisons with existing DSS approaches, as well as direct baselines with standard RAG implementations such as Haystack and the LangChain default pipeline, together with RAG+LLM systems without evidential reasoning, to contextualize the impact of the proposed framework and preprocessing strategies.

5. Conclusions

This study demonstrated that integrating Large Language Models, Retrieval-Augmented Generation, and Paraconsistent Annotated Evidential Logic Eτ enables the development of a conversational, knowledge-based, and contradiction-tolerant decision-support system for poultry production that remains interpretable and robust under uncertainty.

By embedding evidential reasoning at each stage of the conversational workflow, the Decision Support AI-Copilot for Poultry Farming exhibited promising behavior in terms of semantic alignment, inference under uncertainty, and domain attribution. The architecture avoided heuristic shortcuts, relying instead on structured logical evaluation to manage ambiguous or borderline cases. In this context, the present work contributes to the field of AI-based decision support in agriculture by introducing an integrative, multi-domain, knowledge-grounded, and contradiction-tolerant approach tailored to the specific demands of poultry production.

At the same time, the system should be regarded as a proof-of-concept prototype that demonstrates the technical feasibility of integrating evidential logic with LLM-based reasoning. Although the experiments were conducted under controlled conditions and with a restricted corpus, the outcomes provide a solid foundation for subsequent validations in real production environments. These results highlight the system's potential for scalability and operational deployment while consolidating its methodological contribution to explainable decision support.

Building on these results, this study also outlines a roadmap for expanding and validating the framework across broader agricultural contexts. Future research may extend its scope to other livestock systems, incorporate additional data modalities, and refine evidential thresholds based on field feedback. In doing so, the research not only establishes a conceptual foundation but also defines a practical agenda for advancing evidentially guided decision support in smart farming.

Funding: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, Grant Number 88887.199644/2025-00.

Acknowledgments: During the preparation of this manuscript, the authors used GPT-4o for generating synthetic data for testing and experimentation purposes. The authors have carefully reviewed and edited all outputs and take full responsibility for the content of this publication.

References

2. Food balance sheets 2010–2022. Global, regional and country trends. *Statistics*. Available online: https://www.fao.org/statistics/highlights-archive/highlights-detail/food-balance-sheets-2010-2022-global-regional-and-country-trends/en (accessed on 9 July 2025).
3. Mottet, A.; Tempio, G. Global poultry production: Current state and future outlook and challenges. *World's Poult. Sci. J.* 2017, *73*, 245–256. https://doi.org/10.1017/S0043933917000071.
4. Berckmans, D. General introduction to precision livestock farming. *Anim. Front.* 2017, *7*, 6–11. https://doi.org/10.2527/af.2017.0102.

 doi:10.20944/preprints202512.1284.v1

5. Gržinić, G.; Piotrowicz-Cieślak, A.; Klimkowicz-Pawlas, A.; Górny, R. L.; Ławniczek-Wałczyk, A.; Piechowicz, L.; et al. Intensive poultry farming: A review of the impact on the environment and human health. *Sci. Total Environ.* 2023, *858*, 160014. https://doi.org/10.1016/j.scitotenv.2022.160014.
6. Hafez, H. M.; Attia, Y. A. Challenges to the poultry industry: Current perspectives and strategic future after the COVID-19 outbreak. *Front. Vet. Sci.* 2020, *7*, 516. https://doi.org/10.3389/fvets.2020.00516.
7. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv* 2021, arXiv:2005.11401. https://doi.org/10.48550/arXiv.2005.11401.
8. Li, H.; Su, Y.; Cai, D.; Wang, Y.; Liu, L. A survey on retrieval-augmented text generation. *arXiv* 2022, arXiv:2202.01110. https://doi.org/10.48550/arXiv.2202.01110.
9. Leite, M. V.; Abe, J. M.; Souza, M. L. H.; de Alencar Nääs, I. Enhancing environmental control in broiler production: Retrieval-augmented generation for improved decision-making with large language models. *AgriEng* 2025, *7*, 12. https://doi.org/10.3390/agriengineering7010012.
10. Izacard, G.; Grave, E. Leveraging passage retrieval with generative models for open domain question answering. *arXiv* 2021, arXiv:2007.01282. https://doi.org/10.48550/arXiv.2007.01282.
11. Cai, D.; Wang, Y.; Liu, L.; Shi, S. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022)*, New York, NY, USA, July 11–15, 2022. https://doi.org/10.1145/3477495.3532682.
12. Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; et al. Language models are few-shot learners. *arXiv* 2020, arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165.
13. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* 2023, *55*, 248:1–248:38. https://doi.org/10.1145/3571730.
14. Metze, K.; Morandin-Reis, R. C.; Lorand-Metze, I.; Florindo, J. B. Bibliographic research with ChatGPT may be misleading: The problem of hallucination. *J. Pediatr. Surg.* 2024, *59*, 158. https://doi.org/10.1016/j.jpedsurg.2023.08.018.
15. Abe, J. M.; Akama, S.; Nakamatsu, K. *Introduction to Annotated Logics: Foundations for Paracomplete and Paraconsistent Reasoning*; Springer International Publishing: Cham, 2015; Vol. 88. https://doi.org/10.1007/978-3-319-17912-4.
16. Carvalho, F. R. D.; Abe, J. M. *A Paraconsistent Decision-Making Method*; Springer International Publishing: Cham, 2018; Vol. 87. https://doi.org/10.1007/978-3-319-74110-9.
17. de Carvalho Junior, A.; Justo, J. F.; de Oliveira, A. M.; da Silva Filho, J. I. A comprehensive review on paraconsistent annotated evidential logic: Algorithms, applications, and perspectives. *Eng. Appl. Artif. Intell.* 2024, *127*, 107342. https://doi.org/10.1016/j.engappai.2023.107342.
18. Tiwari, A.; Beed, R. S. Applications of Internet of Things in Smart Agriculture. In *AI to Improve e-Governance and Eminence of Life*; Springer: Singapore, 2023; pp 103–115. https://doi.org/10.1007/978-981-99-4677-8_6.
19. Abe, J. M. Remarks on Paraconsistent Annotated Evidential Logic Et. *Unisanta Sci. Technol.* 2014, *3*, 25–29.
20. Petroni, F.; Piktus, A.; Fan, A.; Lewis, P.; Yazdani, M.; Cao, N. D.; et al. KILT: A benchmark for knowledge intensive language tasks. *arXiv* 2021, arXiv:2009.02252. https://doi.org/10.48550/arXiv.2009.02252.
21. Wang, H.; Zhang, D.; Li, J.; Feng, Z.; Zhang, F. Entropy-optimized dynamic text segmentation and RAG-enhanced LLMs for construction engineering knowledge base. *Appl. Sci.* 2025, *15*, 3134. https://doi.org/10.3390/app15063134.
22. Zhu, B.; Vuppalapati, C. Integrating retrieval-augmented generation with large language models for supply chain strategy optimization. In *Applied Cognitive Computing and Artificial Intelligence*; Springer: Cham, 2025; pp 475–486. https://doi.org/10.1007/978-3-031-85628-0_34.
23. Boban, I.; Doko, A.; Gotovac, S. Sentence retrieval using stemming and lemmatization with different length of the queries. *Adv. Sci. Technol. Eng. Syst. J.* 2020, *5*, 349–354. https://doi.org/10.25046/aj050345.
24. Pramana, R.; Debora; Subroto, J. J.; Gunawan, A. A. S.; Anderies. Systematic literature review of stemming and lemmatization performance for sentence similarity. In *Proceedings of the 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, Yogyakarta, Indonesia, Nov 23–25, 2022; IEEE: New York, NY, USA, 2022; pp 1–6. https://doi.org/10.1109/icitda55840.2022.9971451.

25. Sirisha, U.; Kumar, C.; Durgam, R.; Eswaraiah, P.; Nagamani, G. An analytical review of large language models leveraging KDGI fine-tuning, quantum embedding's, and multimodal architectures. *Comput. Mater. Contin.* 2025, *83*, 4031–4059. https://doi.org/10.32604/cmc.2025.063721.
26. Gabín, J.; Parapar, J. Leveraging retrieval-augmented generation for keyphrase synonym suggestion. In *Advances in Information Retrieval*; Springer: Cham, 2025; pp 311–327. https://doi.org/10.1007/978-3-031-88711-6_20.
27. Chung, J.; Kamar, E.; Amershi, S. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Toronto, Canada, July 9–14, 2023; pp 575–593. https://doi.org/10.18653/v1/2023.acl-long.34.
28. Amin, M. M.; Schuller, B. W. On prompt sensitivity of ChatGPT in affective computing. In *Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction (ACII 2024)*, Glasgow, United Kingdom, 15–18 September 2024; pp 203–209. https://doi.org/10.1109/ACII63134.2024.00028.
29. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; et al. GPT-4 technical report. *arXiv* 2024, arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774.
30. Ojo, R. O.; Ajayi, A. O.; Owolabi, H. A.; Oyedele, L. O.; Akanbi, L. A. Internet of things and machine learning techniques in poultry health and welfare management: A systematic literature review. *Comput. Electron. Agric.* 2022, *200*, 107266. https://doi.org/10.1016/j.compag.2022.107266.
31. Astill, J.; Dara, R. A.; Fraser, E. D. G.; Roberts, B.; Sharif, S. Smart poultry management: Smart sensors, big data, and the internet of things. *Comput. Electron. Agric.* 2020, *170*, 105291. https://doi.org/10.1016/j.compag.2020.105291.
32. OpenAI Platform. Models overview. Available online: https://platform.openai.com/docs/models (accessed on February 6, 2025).
33. Akama, S., Ed. *Towards Paraconsistent Engineering*; Springer International Publishing: Cham, 2016. https://doi.org/10.1007/978-3-319-40418-9.
34. DSSAICopilotForPoultryFarming/. Available online: https://github.com/marcusviniciusleite/DSSAICopilotForPoultryFarming/tree/main (accessed on July 23, 2025).
35. Brassó, L. D.; Komlósi, I.; Várszegi, Z. Modern technologies for improving broiler production and welfare: A review. *Animals* 2025, *15* (4), 493. https://doi.org/10.3390/ani15040493.
36. Tareesh, M. T.; Anandhi, M.; Sujatha, G.; Thiruvenkadan, A. K. Digital twins in poultry farming: A comprehensive review of the smart farming breakthrough transforming efficiency, health, and profitability. *Int. J. Vet. Sci. Anim. Husbandry* 2025, *10* (8S), 89–95. https://doi.org/10.22271/veterinary.2025.v10.i8Sb.2476.
37. Nääs, I. A.; Garcia, R. G. The dawn of intelligent poultry science: A Brazilian vision for a global future. *Braz. J. Poult. Sci.* 2025, *27*, eRBCA. https://doi.org/10.1590/1806-9061-2025-2131.
38. Baumhover, A.; Hansen, S. L. Preparing the AI-assisted animal scientist: Faculty and student perspectives on enhancing animal science education with artificial intelligence. *Anim. Front.* 2024, *14* (6), 54–56. https://doi.org/10.1093/af/vfae038.
39. Ghavi Hossein-Zadeh, N. Artificial intelligence in veterinary and animal science: Applications, challenges, and future prospects. *Comput. Electron. Agric.* 2025, *235*, 110395. https://doi.org/10.1016/j.compag.2025.110395.
40. eFarm – Horizon OpenAgri. Available online: https://horizon-openagri.eu/open-source-catalogue/demo-efarm/ (accessed on August 31, 2025).
41. farmOS. Available online: https://farmos.org/ (accessed on August 31, 2025).
42. Poultry Farming Management System – Horizon OpenAgri. Available online: https://horizon-openagri.eu/open-source-catalogue/poultry-farming-management-system/ (accessed on August 31, 2025).
43. Poultry Development Review. Available online: https://www.fao.org/4/i3531e/i3531e00.htm (accessed on August 31, 2025).
44. Environmental guidelines for poultry rearing operations. FAOLEX. Available online: https://www.fao.org/faolex/results/details/en/c/LEX-FAOC204579/ (accessed on August 31, 2025).

 doi:10.20944/preprints202512.1284.v1

45. Mellor, D. J. Operational details of the five domains model and its key applications to the assessment and management of animal welfare. *Animals (Basel)* 2017, *7* (8), 60. https://doi.org/10.3390/ani7080060.
46. Azad, H. K.; Deepak, A. Query expansion techniques for information retrieval: A survey. *Inf. Process. Manag.* 2019, *56* (5), 1698–1735. https://doi.org/10.1016/j.ipm.2019.05.009.
47. Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y. The curious case of neural text degeneration. *arXiv* 2020, arXiv:1904.09751. https://doi.org/10.48550/arXiv.1904.09751.
48. Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; Jordanous, A. Is temperature the creativity parameter of large language models? *arXiv* 2024, arXiv:2405.00492. https://doi.org/10.48550/arXiv.2405.00492.
49. Chen, B.; Zhang, Z.; Langrené, N.; Zhu, S. Unleashing the potential of prompt engineering for large language models. *Patterns* 2025, *6* (6), 101260. https://doi.org/10.1016/j.patter.2025.101260.