ORCHID: Orchestrated Retrieval-Augmented Classification of High-Risk Property with Intelligent Decision-Making

Sanjay Das*
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
dass3@ornl.gov

Maria Mahbub*
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
mahbubm@ornl.gov

Vanessa Lama*
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
lamav@ornl.gov

Brian Starks
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
starksbm@ornl.gov

Christopher Polchek
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
polchekcl@ornl.gov

Saffell Silvers
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
silverssc@ornl.gov

Lauren Deck
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
decklm@ornl.gov

Prasanna Balaprakash
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
pbalapra@ornl.gov

Robert Patton
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
pattonrm@ornl.gov

Tirthankar Ghosal
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
ghosalt@ornl.gov

Abstract

High-Risk Property (HRP) classification is critical at U.S. Department of Energy (DOE) sites, where inventories include sensitive and often dual-use equipment. Compliance must track evolving rules designated by various export control policies to make transparent and auditable decisions. Traditional expert-only workflows are time-consuming, backlog-prone, and struggle to keep pace with shifting regulatory boundaries. We propose ORCHID, a modular agentic framework for HRP classification that pairs retrieval-augmented generation (RAG) with human oversight to produce policy based outputs that can be audited. Small cooperating agents—retrieval, description refiner, classifier, validator, and feedback logger—coordinate via agent-to-agent messaging and invoke tools through the Model Context Protocol (MCP) for model-agnostic on-premise operation. The interface follows an "Item to Evidence to Decision" loop with step-by-step reasoning, on-policy citations, and append-only audit bundles (run-cards, prompts, evidence). In preliminary tests on real HRP cases, ORCHID improves accuracy and traceability over a non-agentic baseline while deferring uncertain items to Subject Matter Experts (SMEs). The demonstration shows single item submission, grounded citations, SME feedback capture, and exportable audit artifacts—illustrating a practical path to trustworthy LLM assistance in sensitive DOE compliance workflows.

1 Introduction

The classification of high-risk properties (HRPs) is a mission-critical task for U.S. DOE sites and national laboratories, with implications for national security, export control, and regulatory compliance (Fergusson and Kerr, 2013). Determining whether a property falls under sensitive categories such as the U.S. Munitions List (USML), Commerce Control List (CCL), or Nuclear Regulatory Commission (NRC) regulations requires careful, context-sensitive reasoning. Traditionally, this process has been manual, labor-intensive, and reliant on subject matter expertise—making it slow, error-prone, and difficult to scale.

As property portfolios grow in volume and complexity, there is a growing need for intelligent systems that can augment human decision-making in this domain. While conventional machine learning (ML) and single-prompt large language models (LLMs) offer some automation potential to assist in HRP classification, they are fundamentally limited by issues such as lack of interpretability, inability to incorporate dynamic rule changes, and poor responsiveness to expert correction. Moreover, tightly coupled architectures make it hard to adapt or debug such systems when requirements shift.

Early efforts to automate export-control and security classification relied on rules and ontologies curated by subject-matter experts (sme), typically wrapping the eCFR United States Munitions List (USML)

*Authors contributed equally to this research.

Proceedings of the First Workshop on Multilingual Report Generation via Retrieval Augmented Generation (RAG4Reports), pages 57–64

Figure 1: ORCHID Workflow: An SME inputs model, vendor, item and description of the asset; the system refines the property description, queries multiple policy databases, retrieves relevant information, classifies the asset, validates the classification and reasoning steps and produces the final response, which gets approved by the SME.

[1], eCFR Nuclear Regulatory Commission (NRC) [2], and the eCFR Commerce Control List (CCL) [3] into machine-readable taxonomies(Li and Chen, 2020; Clark, 2008). These systems improved consistency but struggled with ambiguous cross-category items and frequent rule changes. Recent research has moved toward knowledge-centered and ontology-driven modeling of security/export-control concepts, enabling richer reasoning over product descriptions and technical attributes (Rao et al., 2009). For example, ontology-based security/export-control classification approaches demonstrate that standardized concept graphs can reduce ambiguity and support explainable labeling (Rzepka and Obayashi, 2025), though coverage gaps remain for rapidly evolving technologies (e.g., advanced semiconductors, dual-use AI) (Parizadehgan, 2025).

Retrieval-augmented generation (RAG) has become a dominant strategy for keeping models aligned to authoritative texts and reducing hallucinations in law and governance applications (Reuter et al., 2025; Huang et al., 2025). Industry and academic reports alike emphasize dynamic retrieval from up-to-date regulatory repositories and explicit citation in output. Legal-tech guidance and empirical frameworks such as "dynamic legal RAG,"(Ajay Mukund and Easwarakumar, 2025) Gov-RAG(Yu and Chen, 2025), and SemRAG (Zhong et al., 2025) all report improvements in factuality and traceability when generation is grounded in statutes, regulatory notices, and agency FAQs. These capabilities are essential for export-control determinations.

Recent research in AI has explored various methods for combining human expertise with machine learning models. Active learning and human-in-the-loop systems have been widely used in tasks where expert feedback can help refine models, especially in high-stakes domains like medical diagnostics and security (Ruengsurat et al., 2025; Wang et al., 2024). Additionally, retrieval-augmented models like RAG have shown promise in tasks that require both context-specific information retrieval and generation, such as legal document review and scientific research assistance. However, few systems focus on unifying all these elements while integrating SME feedback in real-time to refine predictions. Furthermore, the domain of classifying high-risk properties for national labs requires both high accuracy and clear explanations of predictions, which presents unique challenges for AI models. Our work aims to address this gap by using a human-in-the-loop approach combined with RAG to improve the classification process.

We present *ORCHID: Orchestrated Retrieval-augmented Classification with Human-in-the-loop Intelligent Decision-making*. ORCHID is a modular, multi-agent system designed to deliver explainable and adaptable HRP classifications, guided by expert feedback and grounded in regulatory source material. At the core of ORCHID is an agentic architecture, where specialized agents perform decomposed tasks such as retrieval, reasoning, validation, and feedback integration. Agents interact via an agent-to-agent (A2A) messaging layer, enabling coordination without monolithic prompts or persistent memory (Ray, 2025). To support modular planning and tool invocation, ORCHID implements a lightweight abstraction called the Model Context Protocol (MCP) (Hou et al., 2025). MCP introduces stateless adapters around external models or tools (e.g., dense retrievers, BM25 indexes, LLMs), all accessible through a common invoke(query, params) interface. This decoupling allows the system to flexibly swap components, combine signals through strategies like reciprocal rank fusion (RRF), and maintain clean separation between retrieval, generation, and orchestration logic.

Unlike traditional LLM pipelines, ORCHID enables adaptive learning, by incorporating SME feedback into ongoing refinement; modular, testable components, supporting rapid iteration and system evolution; operational gains, including reduced human burden, improved classification accuracy, and regulatory auditability.

2 Policy Repository

ORCHID restricts search to a versioned policy corpus (eCFR USML, NRC, CCL, plus EAR99 guidance). Policy text is chunked with stable section IDs and embedded with *mxbai-embed-large-v1*(Lee et al., 2024)(Li and Li, 2023). We maintain a hybrid index: BM25 over nor-

[1]https://www.ecfr.gov/current/title-22/chapter-I/subchapter-M/part-121

[2]https://www.ecfr.gov/current/title-10/chapter-I

[3]https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C/part-774

malized text for lexical matches and a vector index for semantic matches and results are combined with reciprocal rank fusion. The Vector Store tool (MCP) encapsulates this. Agents pass a query object (with/without description, top-k), and receive a ranked list of snippets with section IDs, confidence. A small citation packer filters to minimally sufficient spans that the model must cite verbatim.

3 System Architecture

The system architecture of ORCHID framework is visualized in Figure 2.

The framework is designed as a modular, multi-agent, retrieval-augmented reasoning system that performs auditable policy-grounded classification with minimally structured user inputs. The architecture decomposes the end-to-end decision process into formally defined transformation stages, each implemented by an independent agent and coordinated by a deterministic orchestrator. The primary design objectives are: (i) strict grounding in authoritative policy sources, (ii) deployment-aware security compliance, (iii) explainable reasoning with citation traceability, and (iv) auditable execution with reproducible metadata tracking.

3.1 Problem Formulation

Here, we formalize the high-risk property classification problem. Let the initial structured user input be defined as:

$$\mathcal{I}_0 = \{V, E, M, d_0\} \tag{1}$$

where V denotes the Manufacturer or Vendor name, E denotes the Equipment or Service identifier, M denotes the Model number (optional), and d_0 represents a short free-text description of the property. The objective of the system is to compute a validated classification decision $\mathcal{D}_f$ grounded in policy evidence.

The complete transformation pipeline is expressed as:

$$\mathcal{D}_f = \mathcal{A}_{VR} \circ \mathcal{A}_{HRP} \circ \mathcal{A}_{IR} \circ \mathcal{A}_{DR}(\mathcal{I}_0) \tag{2}$$

where $\mathcal{A}_{DR}$, $\mathcal{A}_{IR}$, $\mathcal{A}_{HRP}$, and $\mathcal{A}_{VR}$ denote the Description Refiner, Information Retrieval, Classification, and Validation agents, respectively.

3.2 Description Refinement Layer

The Description Refiner (DR) agent performs semantic normalization and ambiguity resolution using a large language model (LLM). This is needed to ensure accurate understanding of the user query for the downstream task. It transforms the raw input $\mathcal{I}_0$ into a canonical structured representation:

$$\mathcal{I}_1 = f_{DR}(\mathcal{I}_0; \theta_{LLM}) \tag{3}$$

where θ_{LLM} denotes the model parameters.

The refinement process includes entity normalization, terminology expansion, contextual disambiguation, and optional clarification queries to the user. The refined representation is:

$$\mathcal{I}_1 = \{V', E', M', d_1\} \tag{4}$$

A deployment-dependent security constraint governs external augmentation to ensure sensitive data protection:

$$\text{OnPrem} \Rightarrow \neg\text{ExternalWebAccess} \tag{5}$$

$$\text{OffPrem} \Rightarrow \text{ControlledWebAccess}(V, E, M) \tag{6}$$

This ensures that sensitive environments prohibit outbound search operations.

3.3 Hybrid Information Retrieval Layer

The Information Retrieval (IR) agent constructs hybrid lexical-semantic queries over a policy-scoped corpus. This step is crucial as it provides relevant policy context for the downstream classification task and ensures evidence grounding. Query generation is defined as:

$$Q = g(V', E', M', d_1) \tag{7}$$

Retrieval proceeds via two complementary channels:

Lexical Retrieval (Sparse Retrieval) : This method relies on token-based matching, looking for exact words, phrases, or their variations within a document. Although, it is excellent at finding precise, rare terminology, proper nouns, and exact phrases, it cannot handle rephrases/synonyms.

$$S_{lex} = \text{BM25}(Q) \tag{8}$$

Dense Semantic Retrieval : This method maps the user query into a dense, low-dimensional vector (an embedding) using a pre-trained neural network. Although, it captures the underlying intent and semantic meaning, enabling it to match synonyms, paraphrases, and context-dependent queries, it struggles with highly specific jargon, rare technical acronyms, or exact entity matching.

$$S_{vec} = \text{Embed}(Q) \tag{9}$$

This identifies nearest neighbor over dense vector embeddings.

Therefore, by combining lexical and semantic retrieval, the system mitigates the weaknesses of one with the strengths of the other, which results in a candidate set is formed by union:

$$S_{cand} = S_{lex} \cup S_{vec} \tag{10}$$

A cross-encoder reranker (Reciprocal Rank Fusion(RPF)) optimizes contextual alignment:

$$S_{ranked} = \text{Rerank}(S_{cand}, Q) \tag{11}$$

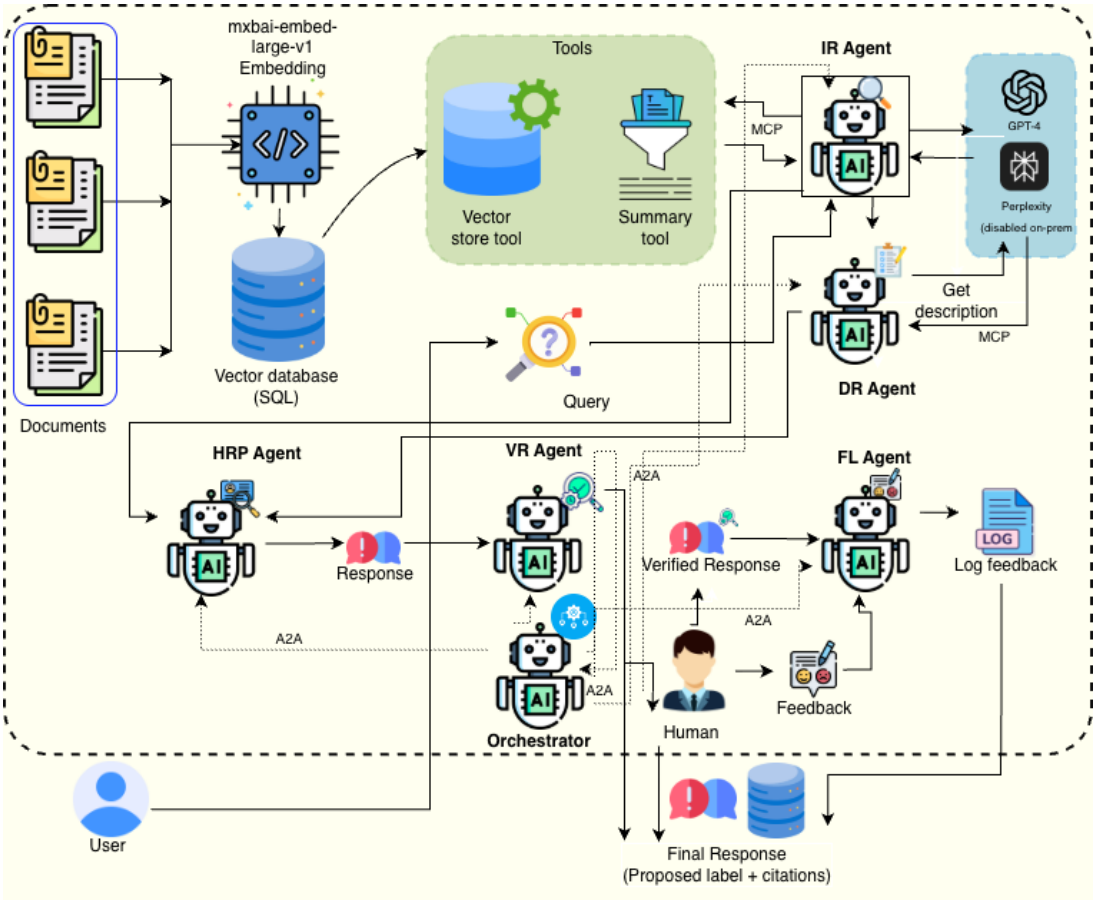

Figure 2: ORCHID agentic architecture. The Orchestrator coordinates agents for Retrieval (IR), Description Refinement (DR), HRP classification, Validation (VR), and Feedback Logging (FL) via agent-to-agent (A2A) messages. IR/DR/HRP access local tools through Model Context Protocol (MCP) adapters (Vector Store, Summary) over a versioned policy corpus. VR either issues a verified decision or routes the case to a human reviewer, whose feedback is recorded in an append-only audit log.

The top-k policy-grounded snippets are selected ($\mathcal{R} = \{s_1, s_2, \ldots, s_k\}$). Each snippet s_i is associated with metadata:

$$s_i = (text_i, doc\_id_i, section_i, version_i) \tag{12}$$

ensuring version traceability and audit compliance.

3.4 Grounded Classification Layer

The HRP classification agent performs citation-grounded classification. It constructs a structured prompt:

$$P = \Phi(\mathcal{I}_1, \mathcal{R}) \tag{13}$$

where Φ enforces explicit inclusion of retrieved evidence. The provisional decision is defined as:

$$\mathcal{D}_p = \{y, c, \mathcal{C}, \mathcal{E}\} \tag{14}$$

where:

- y is the predicted label,
- $c \in [0, 1]$ is the confidence score,
- $\mathcal{C} \subseteq \mathcal{R}$ denotes cited supporting snippets,
- $\mathcal{E}$ represents structured reasoning steps.

A strict grounding constraint is enforced:

$$\forall e_j \in \mathcal{E}, \exists s_i \in \mathcal{C} \tag{15}$$

ensuring that each reasoning step is explicitly supported by retrieved evidence.

3.5 Independent Validation Layer

In this step, we employ a Validation and Review (VR) agent to perform second-order verification of the provisional decision by the previous agent. It evaluates two formal properties.

Coverage Constraint: It checks if each of the reasoning steps ($\forall e_j \in \mathcal{E}$) are supported by any existing retrieved policy snippets ($\exists s_i \in \mathcal{R}$).

$$\text{Coverage} = 1\,(\forall e_j \in \mathcal{E}, \exists s_i \in \mathcal{R}) \tag{16}$$

Conflict Detection: It checks if there exists any policy snippet (S_a) in the citations that contradicts with another

policy snippet (S_b).

$$\text{Conflict} = 1\left(\exists s_a, s_b \in \mathcal{R} \mid s_a \perp s_b\right) \tag{17}$$

Based on these checks, the VR agent emits a verdict:

$$v \in \{\text{AGREE}, \text{REVIEW}, \text{CONFLICT}\} \tag{18}$$

with validation confidence c_v.
A gating mechanism determines routing:

$$\mathcal{D}_f = \begin{cases} \mathcal{D}_p & \text{if } v = \text{AGREE} \wedge c_v > \tau \\ \text{RouteToSME} & \text{otherwise} \end{cases} \tag{19}$$

where τ is a configurable threshold controlling human escalation.

3.6 Feedback Logging and Audit Layer

To ensure traceability and auditability of the system generated outputs, all finalized decisions are appended to an immutable audit store:

$$\mathcal{L}_{t+1} = \mathcal{L}_t \cup \{(\mathcal{I}_1, y, c, v, r_{SME}, \mathcal{RC})\} \tag{20}$$

where r_{SME} denotes reviewer rationale and $\mathcal{RC}$ denotes run-card metadata. An append-only property is ensured to satisfy non-destructive forensic traceability.

3.7 Orchestration and Deterministic Control

To orchestrate and maintain the whole workflow, a non-generative orchestrator coordinates agent execution and maintains state transitions:

$$\mathcal{S}_{t+1} = \mathcal{O}(\mathcal{S}_t, \mathcal{A}_k) \tag{21}$$

Thus, the orchestrator performs, context propagation between agents, execution scheduling, version control annotation, and run-card capture. the run-card metadata captures the LLM_version, Embedding_version, k, τ, and BM25_params.

Importantly, the orchestrator never generates classification content and operates purely at the control plane. Through modular agent decomposition, hybrid retrieval grounding, independent validation, deterministic orchestration, and append-only auditing, the system achieves high-assurance, explainable, and compliance-aligned property identification suitable for regulated environments.

4 Evaluation

In this section, we evaluate the retrieval and classification performance of the framework by comparing tool outputs with SME feedback.

4.1 Experimental Setup

4.1.1 Platform and Tools

The proposed framework was implemented and evaluated on a system equipped with an NVIDIA A100 80GB GPU. The agent-based architecture was developed in Python using the PydanticAI framework, with retrieval and indexing components implemented via chromadb and LlamaIndex and model integration handled through HuggingFace Transformers and PyTorch. The backend service was deployed using FastAPI with Uvicorn for asynchronous serving, while the interactive user interface was built using Streamlit. For semantic retrieval, the system employs the embedding model "intfloat/multilingual-e5-large" and performs cross-encoder reranking using "cross-encoder/ms-marco-MiniLM-L6-v2". Local language model inference is performed using "meta-llama/Llama-3.1-8B-Instruct", while selected reasoning and validation tasks leverage the OpenAI API with the GPT-4o model to achieve state-of-the-art performance.

4.1.2 Data for Evaluation

We evaluate 160 property descriptions with ground truths from SMEs. The ground truths inform if an item is controlled under International Traffic in Arms Regulations (ITAR|USML), Nuclear Regulatory Commission (NRC), or the Commerce Control List (CCL), making them high-risk. If the item poses low-risk, it is labelled EAR99 (i.e., not controlled under any of the lists mentioned above). For each experiment, the model makes a determination of what "category" the item falls under; ITAR|USML, NRC, CCL, or EAR99.

4.1.3 Performance Metrics

The system is currently in a prototype phase, and we have defined a set of core evaluation metrics that are used in assessments. These metrics aim to quantitatively and qualitatively measure the system's adaptability, reliability, and operational efficiency, particularly in the context of human-in-the-loop feedback from Subject Matter Experts (SMEs). To evaluate the retrieval effectiveness and classification precision of our pipeline, we employ the following metrics:

1. **Recall@5**: Measures the ability of the system to retrieve all relevant documents within the top 5 results.

$$\text{Recall@5} = \frac{|\text{Relevant Docs} \cap \text{Retrieved Top 5}|}{|\text{Total Relevant Docs}|} \tag{22}$$

2. **NDCG@5**: Normalized Discounted Cumulative Gain evaluates ranking quality by discounting relevant documents found at lower ranks.

$$\text{DCG}_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}; \quad \text{NDCG}_k = \frac{\text{DCG}_k}{\text{IDCG}_k} \tag{23}$$

3. **Accuracy (4-Class)**: Assesses the model's ability to correctly categorize samples into one of four distinct categories (e.g., USML|ITER, CCL, NRC, EAR99).

$$\text{Acc}_{\text{4-class}} = \frac{\sum_{i=1}^{4} TP_i}{N} \tag{24}$$

4. **Accuracy (Binary/2-Class)**: Evaluates the system's ability to distinguish between "Non-Risk" and any "Sensitive/Risk" class (grouping all 3 classes USML|ITER, CCL, NRC into a single class).

$$\text{Acc}_{\text{2-class}} = \frac{\sum_{i=1}^{2} TP_i}{N} \tag{25}$$

Metric Description: The retrieval metrics (*Recall@5*, *NDCG@5*) focus on the system's ability to surface relevant context, while the **4-class accuracy** provides a granular view of classification performance. The **binary accuracy** specifically isolates the model's capability to act as a safety gate, distinguishing neutral content from any level of sensitive or high-risk information.

4.2 Retrieval Performance

The retrieval performance, summarized in Table 1, illustrates the architectural transition from standalone keyword matching to a multi-stage hybrid pipeline. The baseline Lexical approach utilizing BM25 provides a foundation for exact term matching but exhibits the lowest Recall@5, as it fails to account for semantic variations in policy language. The integration of Semantic Retrieval via a union of candidate sets significantly expands the search breadth, capturing underlying intent and rephrased queries. This performance is further optimized through Reciprocal Rank Fusion (RRF) reranking, which aligns the top candidates more effectively with the user query, thus ensuring that the most contextually relevant and grounded policy snippets are prioritized for downstream tasks.

Table 1: Retrieval Performance Evolution from Lexical Baselines to Hybrid Pipelines.

Configuration	Recall@5	NDCG@5
Lexical (S_{lex} via BM25)	0.17	0.20
Lexical + Semantic ($S_{lex} \cup S_{vec}$)	0.20	0.24
Lexical + Semantic + Rerank (S_{ranked})	0.22	0.27

4.3 Classification Performance

Preliminary results appear in Table 2, and the corresponding confusion matrix is provided in Fig. 4. The system achieves a Weighted Average Accuracy of 63.12% across the four-class taxonomy, with peak performance observed in the NRC (90%) and USML (88%) domains. The diminished accuracy in EAR99 (40%) suggests challenges in classifying baseline or 'catch-all' regulatory content compared to highly specialized policies. Notably, the Binary Accuracy reaches 70.37%, indicating that the IR-driven pipeline is significantly more effective at performing initial risk triage than at granular classification. This performance gap suggests that while the retrieval context successfully surfaces relevant risk indicators, the final classification layer encounters semantic confusion between overlapping regulatory frameworks like CCL and EAR99.

The ORCHID framework improves classification reliability, transparency, and reproducibility through evidence-based policy-aware decision-making. Using RAG, each classification is grounded in traceable citations, ensuring verifiable reasoning. Its hybrid retrieval mechanism integrates domain-specific regulatory corpora, ITAR/USML, NRC, CCL, EAR99, for policy compliance, while a human-in-the-loop design incorporates expert feedback to refine performance and prevent recurring errors. The feedbacks are captured in a structured table with its input context, which is queried in later classifications for similar items. This simple structure currently shows promise and we plan to rigorously evaluate it further to analyze and improve. ORCHID's modular, agentic architecture supports scalability and reproducibility, and its single-click interface streamlines the decision process for efficient, auditable outcomes.

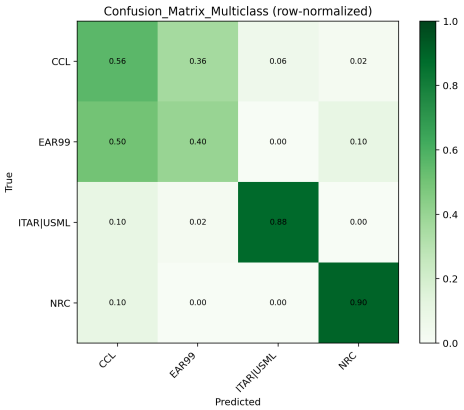

Figure 4: Heatmap with true classes on the y-axis and predicted classes on the x-axis (USML, NRC, CCL, EAR99); values are row-normalized.

The current implementation of ORCHID faces several practical limitations. Its performance depends on curated policy corpora, making it sensitive to coverage gaps and drift when source texts become outdated. Boundary ambiguity persists in fine-grained classifications, particularly in distinguishing CCL and EAR99 items, where validator calibration remains an ongoing effort. The framework currently supports only English text and does not process multimodal inputs such as images or technical specification sheets. In addition, the quality of retrieval and classification is reduced with sparse or poorly written descriptions, and the "no-description" mode exhibits reduced classification relia-

Table 2: Comprehensive preliminary accuracy results.

USML	NRC	CCL	EAR99	Weighted Avg.	Binary Acc
88%	90%	56%	40%	63.12%	70.37%

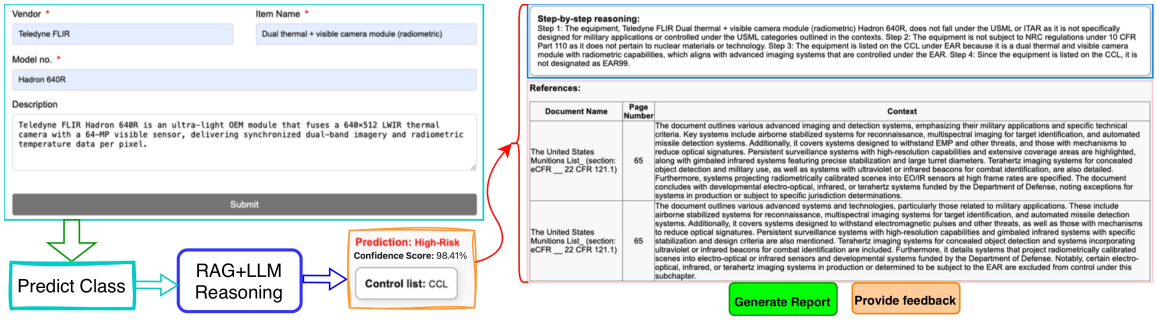

Figure 3: ORCHID UI overview. Submit (vendor, item, model, optional description), inspect policy evidence with citations, review the proposed label and confidence, then record SME feedback.

bility. ORCHID provides decision support but does not constitute legal or regulatory advice, and final determinations must be made by qualified reviewers.

4.4 User Interface

ORCHID's user interface is designed to streamline human–AI interaction in all stages of classification, focusing on transparency, traceability, and minimal analyst effort. After setup, users are directed to the main interface where submissions, results, and feedback are managed through a unified workflow, as visible in Fig. 3.

Submission Workflow: Users initiate the classification workflow by entering a vendor name, item name, and model number, with an optional description, and then trigger processing with a one-click action via the 'Submit' button. This workflow supports both individual entries and batch uploads for multiple items.

Outputs: Upon submission, the interface displays the system's prediction (HRP or Not HRP), the predicted control category, and a single confidence score summarizing model certainty.

Reasoning and Evidence: Each result includes a concise, step-by-step reasoning trace supported by clickable citations. An evidence table presents the underlying documents, sections or pages, and extracted text, with actions for quick copy or open, as well as trace IDs for provenance tracking.

Feedback and Review: Users can provide structured feedback – agreement status, notes, rating, or policy reference – through a simple form submitted via the 'Submit Feedback' button. This input is stored for audit and model refinement.

Batch Mode and Export: For large-scale reviews, the same interface supports batch processing with per-item status indicators and downloadable results. Completed analyses can be exported in JSON, CSV, or PDF formats, each embedding a version strip that records the model identifier, index snapshot, and timestamp to ensure auditability.

The application's frontend provides a cohesive, context-aware experience, guiding users seamlessly from submission to reasoning review, feedback, and export, ensuring both operational efficiency and traceable decision support.

5 Conclusion

In conclusion, we present ORCHID, an agentic framework designed for the scalable, evidence-grounded classification of high-risk properties (HRP). Our initial implementation demonstrates robust performance, particularly within controlled regulatory classes, achieving accuracies up to 90%. Current efforts are focused on refining the architecture to enhance cross-domain classification performance and integrating the framework into mission-critical HRP classification workflows to aid and ease SME burden while maintaining end-to-end traceability and auditability.

References

S Ajay Mukund and KS Easwarakumar. 2025. Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation. *Symmetry*, 17(5):633.

K Clark. 2008. Automated security classification. *Master's thesis, Vrije Universiteit*.

Ian F Fergusson and Paul K Kerr. 2013. The us export control system and the president's reform initiative. Library of Congress, Congressional Research Service.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Tong Li and Zhishuai Chen. 2020. An ontology-based learning approach for automatically classifying security requirements. *Journal of Systems and Software*, 165:110566.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Elham Parizadehgan. 2025. Can u.s. export law handle ai?

Jinghai Rao, Alberto Sardinha, and Norman Sadeh. 2009. A meta-control architecture for orchestrating policy enforcement across heterogeneous information sources. *Journal of Web Semantics*, 7(1):40–56.

Partha Pratim Ray. 2025. A review on agent-to-agent protocol: Concept, state-of-the-art, challenges and future directions. *Authorea Preprints*.

Markus Reuter, Tobias Lingenberg, Rūta Liepiņa, Francesca Lagioia, Marco Lippi, Giovanni Sartor, Andrea Passerini, and Burcu Sayin. 2025. Towards reliable retrieval in rag systems for large legal datasets. *Preprint*, arXiv:2510.06999.

Satida Ruengsurat, Jaimai Eawsivigoon, Vidchaphol Sookplang, Karin Sumongkayothin, Prarinya Siritanawan, Razvan Beuran, and Kazunori Kotani. 2025. Human-in-the-loop for machine learning in offensive cybersecurity. pages 0331–0336.

Rafal Rzepka and Akihiko Obayashi. 2025. Effectiveness of security export control ontology for predicting answer type and regulation categories. In *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, ICAAI '24, page 156–161, New York, NY, USA. Association for Computing Machinery.

Haoran Wang, Qiuye Jin, Shiman Li, Siyu Liu, Manning Wang, and Zhijian Song. 2024. A comprehensive survey on deep active learning in medical image analysis. *Medical Image Analysis*, 95:103201.

Miao Yu and Hailiang Chen. 2025. Gov-rag: A retrieval-augmented generation framework for enhancing e-government services. *Available at SSRN 5111865*.

Kezhen Zhong, Basem Suleiman, Abdelkarim Erradi, and Shijing Chen. 2025. Semrag: Semantic knowledge-augmented rag for improved question-answering. *arXiv preprint arXiv:2507.21110*.