Foundations of Formal Reasoning over Knowledge Bases Combining Symbolic and Sub-Symbolic Knowledge

Gianluca Cima, Marco Console, Laura Papi

Sapienza University of Rome
{cima, console, papi}@diag.uniroma1.it

Abstract

More and more organizations are relying on Machine Learning (ML) models to support internal decision-making processes. To better support such processes, it would be highly beneficial to contextualize the inductively acquired knowledge encoded in these models and enable formal reasoning over it. Despite significant progress in Neuro-Symbolic AI, this specific challenge remains largely under-explored. We propose a framework that allows to integrate the knowledge induced by ML classifiers with the knowledge specified by logic-based formalisms. The framework is based on the novel notion of *Hybrid Knowledge Base (HKB)*, consisting of two components: an ontology and a set of ML binary classifiers. As usual, the ontology provides an intensional representation of the modeled domain through logic-based axioms, while the binary classifiers implicitly encode the extensional knowledge. Specifically, a HKB associates to each concept and role mentioned in the ontology a classifier based on a set of features deemed to be relevant for the application domain, thereby virtually populating the concepts and roles with the instances and pairs of instances from the feature space. Besides the definition of the new framework, as a more technical contribution we show how to reason in this framework by studying query answering over HKBs. In particular, we investigate the computational complexity of query answering in a rich language over HKBs in which the ontology is specified in (the Description Logic counterpart of) RDFS, while the binary classifiers are represented by Multi-Layer Perceptrons.

1 Introduction

In recent years, *information* represented by Machine Learning (ML) models (Bishop 2007; Hastie, Tibshirani, and Friedman 2009) is increasingly prominent in Information Systems (Chen, Chiang, and Storey 2012). Techniques leveraging such information are regularly employed in the most diverse application domains, such as finance, healthcare, and law, and have demonstrated strong empirical success in a variety of concrete tasks (Jiang et al. 2017; Bahoo et al. 2023).

Despite their remarkable successes, ML models still exhibit significant limitations that hinder their adoption in mission-critical settings (Thames and Sun 2024). A key limitation lies in the restricted form of interaction they offer. At a high level, an ML model is (the representation of) a function $M\colon A \to B$, and typically, the only *query* that a user can pose to M is *value computation*, i.e. given $a \in A$, return the output $M(a) \in B$. While such point-wise (*local*) access underpins many successful stories of ML-based techniques, numerous crucial tasks implicitly demand a *global* view of the information encoded by M, which remains difficult to access (König et al. 2024). In general, it would be very useful to put the inductively acquired knowledge contained in these models into context and formally reason over it, ideally by combining such knowledge with a logical specification of the domain of interest over which these ML models operate. Despite the great effort in recent years on Neural Symbolic AI systems, this specific problem remains under-explored.

In this paper, we propose a novel logical framework that enables the integration of knowledge induced by ML classifiers with knowledge specified by logic-based formalisms. Our framework falls under the general category of Ontology-Driven Knowledge Bases. The core principle of this paradigm is to represent knowledge via two complementary components: an *intensional part* and an *extensional part*. The intensional part, specified through an *ontology* (also referred to as TBox), comprises the vocabulary, i.e. the involved concepts and roles, and a set of axioms expressed in some formal language that constrain their interrelations, providing a semantically rich intensional representation of the modeled domain. The extensional part refers to the actual data, specifying which objects (resp. pairs of objects) are instances of concepts (resp. roles). Depending on the use case, the extensional component of an Ontology-Driven Knowledge Base methodology can take different forms. For example, in Ontology-Mediated Query Answering setting (Bienvenu and Ortiz 2015), it is realized as an ABox, whereas in the Virtual Knowledge Graph (VKG) paradigm (Xiao et al. 2019) (originally termed Ontology-Based Data Access (OBDA) (Poggi et al. 2008)) it consists of a relational database together with mappings that link the underlying database schema to the concepts and roles in the ontology.

In our novel framework, we introduce a further method to define the extensional component, which leverages ML models. We consider the typical scenario in which an application domain has a set of relevant features over which ML models operate. The framework is based on the notion of a *Hybrid Knowledge Base (HKB)*, consisting of two compo-

nents: an ontology and a set of ML binary classifiers. The former conceptualizes the application domain, while the latter assigns to each concept (resp. role) in the ontology vocabulary a classifier (resp. classifiers over pairs) trained on the relevant features of the domain, thereby virtually specifying the extension of concepts (resp. roles) with feature space elements (resp. pairs of feature space elements).

Example 1. *Consider a medical scenario exploiting classifiers to determine whether a patient is diabetic (classifier κ_{D}), male (κ_{M}), or pregnant (κ_{P}), as well as whether a pair of patients are compatible blood donors (classifier over pairs λ_{cD}) or share a compatible blood type (λ_{cB}).*

In our framework, we can place this inductively obtained knowledge within a semantic context. Specifically, we model the above scenario through an ontology $\mathcal{O}$, and by assigning to each concept and role in the ontology's vocabulary its corresponding classifier. The vocabulary of $\mathcal{O}$ consists of the atomic concepts D, M, and P for diabetics, males, and pregnant patients, respectively, and the atomic roles cD and cB for pairs of patients that are compatible donors and pairs of patients with compatible blood types, respectively. The background knowledge of $\mathcal{O}$ sanctions that pregnant patients cannot be male ($\mathsf{P} \sqsubseteq \neg\mathsf{M}$). Finally, the HKB is composed by $\mathcal{O}$ and the set of classifiers described above, where κ_C is assigned to the atomic concept C (for $C \in \{\mathsf{D}, \mathsf{M}, \mathsf{P}\}$) and λ_R is assigned to the atomic role R (for $R \in \{\mathsf{cD}, \mathsf{cB}\}$).

As in a VKG system, the extensional knowledge of a HKB is defined implicitly. In particular, it is encoded in the classifiers assigned to the concepts and roles of the ontology, whose extensions are fully determined by these classifiers.

We provide a model-theoretic semantics for HKBs. An interpretation for a HKB $\mathcal{K}$ is a pair $(\mathcal{I}, f)$, where $\mathcal{I}$ is a standard first-order interpretation over the ontology vocabulary, and f is a function that maps each feature space element (i.e. each input to the classifiers) to a (possibly empty) set of objects from the domain $\Delta^{\mathcal{I}}$ of $\mathcal{I}$. This reflects the assumption that each feature space element $\bar{a}$ either corresponds to a combination of attribute values that no real-world entity can exhibit ($f(\bar{a}) = \emptyset$) or corresponds to one or more entities sharing exactly those characteristics. The models of a HKB $\mathcal{K}$ are those interpretations $(\mathcal{I}, f)$ such that $\mathcal{I}$ satisfies the ontology axioms, and the extension of each concept and role in the ontology vocabulary exactly follows the classification defined by the corresponding classifier via the mapping f.

Due to the presence of ontology axioms, there will be feature space elements *discarded* by HKB models. For instance, consider the axiom $\mathsf{P} \sqsubseteq \neg\mathsf{M}$ in Example 1, and let $\bar{a}$ be a feature space element simultaneously classified positively by κ_{P} and κ_{M} (i.e. $\kappa_{\mathsf{P}}(\bar{a}) = \kappa_{\mathsf{M}}(\bar{a}) = 1$). In this case, every model of the HKB must discard $\bar{a}$, i.e. set $f(\bar{a}) = \emptyset$. Clearly, it is desirable to discard feature space elements only when necessary, so as to maximize coverage of the feature space and, consequently, preserve as much information as possible from the application of the employed ML classifiers to the sub-symbolic data. For this reason, we introduce the notion of *minimally-discarding models*, which are models that discard a $\subseteq$-minimal set of feature space elements.

Importantly, our model-theoretic semantics enables the possibility to formally reason over the combination of symbolic and sub-symbolic knowledge within a HKB. It provides a foundation for studying classical reasoning tasks from the Knowledge Representation and Reasoning tradition over knowledge acquired through symbolic and inductive techniques, with users interacting via the ontology layer, as is typical of any ontology-driven methodology. In this paper, we focus on two reasoning tasks: *non-trivial consistency checking* and *query answering*. The former checks whether there exists a model that does not discard all elements of the feature space, which is a key reasoning task for assessing whether the ML classifiers contribute any meaningful information, assuming, as usual, that the ontology is flawless and error-free. The latter is arguably one of the most studied tasks in symbolic reasoning, serving as a mechanism to extract information from the various (preferred) models of the knowledge encoded within an application domain.

Besides the novel framework, the other main contribution of this paper is a thorough computational complexity analysis of the aforementioned reasoning tasks over HKBs in a meaningful setting. We consider the Description Logic (DL) *DL-Lite*$^{\neg}_{\mathtt{RDFS}}$ as the ontology language and unions of conjunctive queries with inequalities (UCQ$^{\neq}$s) as the query language. This setting has been the subject of recent thorough investigations in the ontology-mediated query answering literature (see (Cima, Lenzerini, and Poggi 2020; Cima et al. 2025)). The DL *DL-Lite*$^{\neg}_{\mathtt{RDFS}}$ corresponds to the DL counterpart of RDFS (often called *DL-Lite*$_{\mathtt{RDFS}}$ (Cuenca Grau 2004; Rosati 2007)) extended with disjointness axioms, while the query language is the well-known union of conjunctive queries extended with inequality atoms. For the classifiers, we consider Multi-Layer Perceptrons (MLPs), a widely used class for classification tasks over structured data.

Similarly to the so-called *data complexity* which assumes that both the ontology and the query (if applicable) are fixed, we focus on the *classifiers complexity*, i.e. the complexity in which only the set of classifiers is regarded as the input. We show that non-trivial consistency checking is NP-complete, whereas query answering is significantly more intractable, and more specifically coNExpTime-complete. Due to this latter negative result, we explore both a restricted setting and an alternative semantics for query answering. The restricted setting forbids the use of roles in ontology axioms, resulting in the fragment $\mathcal{H}^{\neg}$ of *DL-Lite*$^{\neg}_{\mathtt{RDFS}}$. In this restricted setting, we prove that both tasks are NP-complete. As for the alternative semantics, we define a semantics for query answering based on the "When In Doubt Throw It Out" principle from belief revision, which restricts reasoning to those feature space elements never discarded by any minimally-discarding model. We prove that query answering under this alternative semantics becomes Σ^p_2-complete, while it remains NP-complete in the restricted setting. Interestingly, for the results related to query answering, all the lower bounds already hold for conjunctive queries.

The remainder of this paper is organized as follows. In Section 2, we discuss some related works. In Section 3, we introduce the necessary background on ontologies and classifiers. In Section 4, we present our novel notion of HKB, including the semantics of queries. In Section 5, we provide

the computational complexity analysis discussed above. Finally, in Section 6, we conclude and outline future work.

2 Related Work

The idea of integrating ML techniques and symbolic reasoning is at the core of the field of Neuro-Symbolic Artificial Intelligence (NSAI) (Hitzler and Sarker 2022). Many approaches proposed in NSAI prescribe the use of some formal language to interact with the information encoded within ML models. However, to the best of our knowledge, no previous work allows to fully integrate the information encoded into a set of ML models with the axioms of an ontology and perform logical reasoning. In the remainder of this section, we survey some of the most relevant related works and identify the main differences in our approach.

A prominent approach to integrate logical formalisms and ML models consists in using the former to express constraints over the latter (see (Giunchiglia, Stoian, and Lukasiewicz 2022) for an interesting survey). This gives rise to a family of ML models that we can call *Logically Constrained* (LCML). The behavior of an LCML model m is restricted using a set of logical constraints T expressed by encoding T into m *at training time*. This is achieved either by encoding the constraints into the loss-functions used by the learning algorithms (Xu et al. 2018) or by modifying the structure of m altogether (Giunchiglia et al. 2024). While the framework of LCML models may seem akin to the notion of HKBs, there are several crucial differences. Firstly, the aim of the LCML is to define models that comply with a specification: no form of logical reasoning over the knowledge stored in the model is allowed as is the case of HKBs. Secondly, constraints are expressed over one single model using the attributes of the data. Thus, there is no form of integration and contextualization across distinct models. Finally, LCML models do not provide any mechanism for information extraction other than value computation while one of the core features of HKBs is query answering.

Information extraction via logical formalisms is another important trend of NSAI, especially in the context of knowledge graph embeddings (KGE) (Wang et al. 2017). This line of work advocates the use of logics to express queries directly over a KGE in such a way to consider edges and entities that were not explicitly present in the original graph but labeled as possible in the embedding (see, e.g. (Hamilton et al. 2018; Fischer et al. 2019; Arakelyan et al. 2021)). A similar approach has been investigated for general ML models, not necessarily in the KGE framework (Fischer et al. 2019). These approaches lack some of the distinctive features of HKBs. Firstly, they are based on a very simple form of semantics (grounded on the likelihood of edges and entities in the embedding) and cannot capture the subtle nuances of a full-fledged logical formalism. Secondly, since there is no external ontology defining a shared vocabulary, they cannot integrate the information coming from multiple ML models in any meaningful way. Finally, there is no tool to express logical constraints (as the case of LCML models above), thus answers are subject to the characteristic uncertainty of ML models despite their rigorous logical definition.

One last line of work that we deem close to ours is the formal verification of ML models. In this line of work, the compliance of an ML model on a set of constraints is tested using automated reasoning techniques (see (König et al. 2024) for a thorough survey). We believe that HKBs could help experts in the verification of ML models by providing a conceptual language for specifying properties of interest, i.e. ontologies and queries for HKBs.

3 Preliminaries

We assume the reader is familiar with *function-free first-order logic with equality* (*FOL*). We define Σ_C, Σ_R, and Var to be the pairwise disjoint, countably infinite sets of *atomic concepts*, *atomic roles*, and *variables*, respectively.

Ontologies. Similarly to (Xiao et al. 2019), in this paper an ontology $\mathcal{O}$ consists of a finite set of declarations of concepts and roles from Σ_C and Σ_R, respectively, and a finite set of FOL axioms. We let $\mathsf{sig}(\mathcal{O})$, called the *signature* of $\mathcal{O}$, be the set of declared atomic concepts and atomic roles. We impose that axioms in $\mathcal{O}$ use symbols from $\mathsf{sig}(\mathcal{O}) \cup \{=\}$ for predicate names and symbols from Var for terms.

We are interested in *DL-Lite*$^{\neg}_{\mathtt{RDFS}}$ ontologies (Cima et al. 2025) in the technical sections. The axioms in a *DL-Lite*$^{\neg}_{\mathtt{RDFS}}$ ontology take the following forms (we use the DL notation):

$$\begin{array}{lll} B \sqsubseteq C & P_1 \sqsubseteq P_2 & \text{(concept/role inclusion)} \\ B_1 \sqsubseteq \neg B_2 & P_1 \sqsubseteq \neg P_2 & \text{(concept/role disjointness),} \end{array}$$

where $C \in \Sigma_C$ and for $i = 1$ and $i = 2$ (i) P_i is a *basic role*, i.e. either an atomic role $R \in \Sigma_R$ or the inverse of an atomic role $R \in \Sigma_R$, which we denoted by R^-, and (ii) B_i is a *basic concept*, i.e. either an atomic concept $C \in \Sigma_C$ or an expression of the form $\exists P$ with P a basic role.

Interpretations. Given an ontology $\mathcal{O}$, an *interpretation for* $\mathcal{O}$ is a pair $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where $\Delta^{\mathcal{I}}$ is a non-empty set of objects, called *interpretation domain*, and the *interpretation function* $\cdot^{\mathcal{I}}$ assigns (i) a set $C^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ to each atomic concept $C \in \mathsf{sig}(\mathcal{O})$ and (ii) a set $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$ to each atomic role $R \in \mathsf{sig}(\mathcal{O})$. We denote by $\mathcal{I} \models \mathcal{O}$ the fact that $\mathcal{I}$ satisfies all the axioms in $\mathcal{O}$, i.e. $\mathcal{I} \models \alpha$ for each $\alpha \in \mathcal{O}$.

Given a *DL-Lite*$^{\neg}_{\mathtt{RDFS}}$ ontology $\mathcal{O}$ and an interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ for $\mathcal{O}$, the interpretation function $\cdot^{\mathcal{I}}$ extends to basic concepts and roles as follows: $(\exists R)^{\mathcal{I}} = \{o \mid \exists o'.(o, o') \in R^{\mathcal{I}}\}$, $(R^-)^{\mathcal{I}} = \{(o, o') \mid (o', o) \in R^{\mathcal{I}}\}$, and $(\exists R^-)^{\mathcal{I}} = \{o \mid \exists o'.(o', o) \in R^{\mathcal{I}}\}$. Then, we say that $\mathcal{I}$ satisfies a concept inclusion $B \sqsubseteq C$ (resp. role inclusion $P_1 \sqsubseteq P_2$) if $B^{\mathcal{I}} \subseteq C^{\mathcal{I}}$ (resp. $P_1^{\mathcal{I}} \subseteq P_2^{\mathcal{I}}$) and satisfies a concept disjointness $B_1 \sqsubseteq \neg B_2$ (resp. role disjointness $P_1 \sqsubseteq \neg P_2$) if $B_1^{\mathcal{I}} \cap B_2^{\mathcal{I}} = \emptyset$ (resp. $P_1^{\mathcal{I}} \cap P_2^{\mathcal{I}} = \emptyset$).

Queries. Given an ontology $\mathcal{O}$, a *FOL query* q *over* $\mathcal{O}$ is an expression of the form $q = \{\bar{x} \mid \varphi(\bar{x})\}$, where $\bar{x} = (x_1, \ldots, x_m)$ is a tuple of variables from Var (m is the *arity* of q), and $\varphi(\bar{x})$ is a FOL formula with the variables in $\bar{x}$ as the free variables, predicate names from $\mathsf{sig}(\mathcal{O}) \cup \{=\}$, and terms from Var. For a FOL query $q = \{\bar{x} \mid \varphi(\bar{x})\}$ over an ontology $\mathcal{O}$ with $\bar{x} = (x_1, \ldots, x_m)$, an interpretation $\mathcal{I}$ for $\mathcal{O}$, and a tuple of objects $\bar{o} = (o_1, \ldots, o_m)$ from $\Delta^{\mathcal{I}}$, we denote by $\mathcal{I} \models \varphi(\bar{x}/\bar{o})$ the fact that $\mathcal{I}$ satisfies the FOL

sentence $\varphi(\bar{x}/\bar{o})$ obtained by replacing each occurrence of the free variable x_i with the domain object o_i, for $i \in [m]$.

We consider two query languages: *conjunctive queries* (CQ) and *unions of conjunctive queries with inequalities* (UCQ$^{\neq}$). A union of conjunctive queries with inequalities (UCQ$^{\neq}$) q over an ontology $\mathcal{O}$ takes the form $q = \{\bar{x} \mid \exists \bar{y}_1.\, \phi_1(\bar{x}, \bar{y}_1) \wedge \xi_1(\bar{x}, \bar{y}_1) \vee \ldots \vee \exists \bar{y}_p.\, \phi_p(\bar{x}, \bar{y}_p) \wedge \xi_p(\bar{x}, \bar{y}_p)\}$, where for each $i \in [p]$: (i) $\bar{y}_i$ is a tuple of variables from Var with $\bar{x} \cap \bar{y}_i = \emptyset$, (ii) $\phi_i(\bar{x}, \bar{y}_i)$ is a conjunction of atoms with predicate names from $\mathsf{sig}(\mathcal{O})$ and terms from $\bar{x} \cup \bar{y}_i$, and (iii) $\xi_i(\bar{x}, \bar{y}_i)$ is a conjunction of *inequality atoms* (i.e. an atom of the form $t_1 \neq t_2$ with $t_1, t_2 \in \bar{x} \cup \bar{y}_i$). We say that q is a conjunctive query (CQ) if $p = 1$ and $\xi_1(\bar{x}, \bar{y}_1)$ is empty.

Classifiers. We fix a countably infinite set $\mathbb{A}$ of symbols for *attributes*. To each $A \in \mathbb{A}$, we associate an *attribute domain* D_A, consisting of the non-empty set of possible values for the attribute A (e.g. D_A can be a finite set of *categories*, the natural numbers $\mathbb{N}$, or even the reals $\mathbb{R}$). Given a tuple $\mathcal{A} = (A_1, \ldots, A_n)$ of attributes from $\mathbb{A}$, we denote by $\mathbb{F}(\mathcal{A})$ the *feature space of* $\mathcal{A}$, defined as $\mathbb{F}(\mathcal{A}) = D_{A_1} \times \ldots \times D_{A_n}$.

Given a tuple $\mathcal{A} = (A_1, \ldots, A_n)$ of attributes from $\mathbb{A}$, a *binary classifier based on* $\mathcal{A}$ (when $\mathcal{A}$ is clear from the context, a *binary classifier*) is a function $\kappa \colon \mathbb{F}(\mathcal{A}) \to \{0, 1\}$ assigning to each vector of values for the attributes in $\mathcal{A}$ a label in $\{0, 1\}$, with the usual meaning that the label 1 (resp. 0) corresponds to the positive (resp. negative) class. We also make use of the notion of a *binary classifier over pairs based on* $\mathcal{A}$ (when $\mathcal{A}$ is clear from the context, a *binary classifier over pairs*), which is a function $\lambda \colon \mathbb{F}(\mathcal{A}) \times \mathbb{F}(\mathcal{A}) \to \{0, 1\}$.

4 A Formal Framework for HKBs

We assume that the extensional part of a Knowledge Base is defined via a set of binary classifiers, one for each atomic concept mentioned in the ontology, and a set of binary classifiers over pairs, one for each atomic role mentioned in the ontology. This leads us to the following formal definition.

Definition 1. *A* Hybrid Knowledge Base *(*HKB*)* $\mathcal{K}$ *is a pair* $\mathcal{K} = (\mathcal{O}, \Psi)$*, where* $\mathcal{O}$ *is an ontology and* Ψ *is a set of classifiers based on a tuple* $\mathcal{A}$ *of attributes from* $\mathbb{A}$ *(and we let* $\mathsf{sig}(\Psi) = \mathcal{A}$*). More precisely,* Ψ *contains:*

- *a binary classifier* κ_C*, for each* $C \in \mathsf{sig}(\mathcal{O})$*;*
- *a binary classifier over pairs* λ_R*, for each* $R \in \mathsf{sig}(\mathcal{O})$*.*

Note that the classifiers in Ψ operate over a tuple $\mathcal{A}$ of attributes. As such, $\mathcal{A}$ conveys the information from an Information System deemed relevant for supporting the decision-making processes of an organization. It can be thought of as the result of a feature selection process.

Example 2. *Recall Example 1, and suppose the classifiers are based on the tuple* $\mathcal{A} = (\mathsf{age}, \mathsf{bmi}, \mathsf{ant}, \mathsf{a1c}, \mathsf{hb})$ *of patient attributes, where* age *is the patient age, with* $D_{\mathsf{age}} = \{0, 1, 2\}$ *(*0 *corresponds to young,* 1 *to middle-aged, and* 2 *to senior);* bmi *is the body mass index, with* $D_{\mathsf{bmi}} = \{0, 1, 2\}$ *(*0 *corresponds to low BMI,* 1 *medium, and* 2 *high);* ant *is the number of ABO blood group antigens, with* $D_{\mathsf{ant}} = \{0, 1, 2\}$ *(*0 *corresponds to blood type O,* 1 *to either A or B, and* 2 *to AB);* $\mathsf{a1c}$ *indicates the value of blood glucose, with* $D_{\mathsf{a1c}} = \{0, 1, 2\}$ *(*0 *corresponds to low glucose level,* 1 *medium, and* 2 *high); and* hb *contains the hemoglobin levels according to blood tests, with* $D_{\mathsf{hb}} = \{0, 1, 2\}$ *(*0 *corresponds to low hemoglobin level,* 1 *medium, and* 2 *high). Suppose the classifiers in* Ψ *are defined as follows:*

- *For each* $\bar{a} = (a_1, a_2, a_3, a_4, a_5) \in \mathbb{F}(\mathcal{A})$*, we have:*
  - $\kappa_{\mathsf{D}}(\bar{a}) = 1$ *if and only if* $a_2 + a_3 - 3 \geq 0$*;*
  - $\kappa_{\mathsf{M}}(\bar{a}) = 1$ *if and only if* $3a_2 + a_5 - 5 \geq 0$*;*
  - $\kappa_{\mathsf{P}}(\bar{a}) = 1$ *if and only if* $-a_2 + \frac{1}{2}a_3 + a_4 - \frac{5}{2} \geq 0$
- *For each pair* $(\bar{a}, \bar{b}) \in \mathbb{F}(\mathcal{A}) \times \mathbb{F}(\mathcal{A})$*, where* $\bar{a} = (a_1, a_2, a_3, a_4, a_5)$ *and* $\bar{b} = (b_1, b_2, b_3, b_4, b_5)$*, we have:*
  - $\lambda_{\mathsf{cD}}(\bar{a}, \bar{b}) = 1$ *if and only if* $-a_3 - a_4 + b_3 + b_4 - 1 \geq 0$*;*
  - $\lambda_{\mathsf{cB}}(\bar{a}, \bar{b}) = 1$ *if and only if* $a_3 - b_3 \geq 0$*.*

The HKB $\mathcal{K}$ *for the medical scenario is then defined as the pair* $\mathcal{K} = (\mathcal{O}, \Psi)$*, where* $\mathcal{O}$ *is the ontology with* $\mathsf{sig}(\mathcal{O}) = \{\mathsf{D}, \mathsf{M}, \mathsf{P}, \mathsf{cB}, \mathsf{cD}\}$ *and which states the set* $\{\mathsf{P} \sqsubseteq \neg \mathsf{M}\}$ *of axioms, while* $\Psi = \{\kappa_{\mathsf{D}}, \kappa_{\mathsf{M}}, \kappa_{\mathsf{P}}, \lambda_{\mathsf{cD}}, \lambda_{\mathsf{cB}}\}$*.*

We now turn to describe the semantics of HKBs, which we formalize in logical terms through first-order interpretations. Before proceeding, a few clarifications are in order. As previously noted, the data deemed relevant is conveyed by the set $\mathcal{A}$ of attributes over which the classifiers of a HKB operate, and, more specifically, is captured by the feature space $\mathbb{F}(\mathcal{A})$. Note that each element $(a_1, \ldots, a_n) \in \mathbb{F}(\mathcal{A})$ either does not correspond to any real-world entity (i.e. no actual entity can exhibit its combination of attribute values) or represents one or more real entities sharing exactly those characteristics. With this observation in mind, we are now ready to define the notion of interpretation for a HKB.

Definition 2. *Given a HKB* $\mathcal{K} = (\mathcal{O}, \Psi)$ *with* $\mathsf{sig}(\Psi) = \mathcal{A}$*, an* interpretation for $\mathcal{K}$ *is a pair* $\mathfrak{I} = (\mathcal{I}, f)$*, where*

- $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ *is an interpretation for* $\mathcal{O}$*;*
- $f \colon \mathbb{F}(\mathcal{A}) \to \mathcal{P}(\Delta^{\mathcal{I}})$ *associates to each element* $\bar{a} \in \mathbb{F}(\mathcal{A})$ *a (possibly empty) set* $f(\bar{a})$ *of objects from* $\Delta^{\mathcal{I}}$*;*
- $f(\bar{a}) \cap f(\bar{b}) = \emptyset$ *holds for every pair* $(\bar{a}, \bar{b}) \in \mathbb{F}(\mathcal{A}) \times \mathbb{F}(\mathcal{A})$ *such that* $\bar{a} \neq \bar{b}$*.*

Note that f acts as a bridge from the raw input vectors, which constitute the sub-symbolic representation of data, to the symbolic representation embodied by the objects in the interpretation domain. The third bullet point ensures that distinct elements of the feature space are mapped to distinct objects of the interpretation domain (similarly to the Unique Name Assumption). Also, the definition allows modeling situations in which a combination of attribute values $\bar{a} \in \mathbb{F}(\mathcal{A})$ is infeasible, expressible by setting $f(\bar{a}) = \emptyset$, capturing the fact that no real-world entity exhibits such attributes according to the considered interpretation.

In what follows, for $\mathfrak{I} = (\mathcal{I}, f)$, we denote by $\mathsf{disc}(\mathfrak{I})$ the set of $\mathfrak{I}$*-discarded elements*, i.e. $\mathsf{disc}(\mathfrak{I}) = \{\bar{a} \in \mathbb{F}(\mathcal{A}) \mid f(\bar{a}) = \emptyset\}$. We also say that $\mathfrak{I}$ is *trivial* if $\mathsf{disc}(\mathfrak{I})$ coincides with the feature space $\mathbb{F}(\mathcal{A})$, i.e. if $\mathsf{disc}(\mathfrak{I}) = \mathbb{F}(\mathcal{A})$. We are now ready to define the notion of model for a HKB.

Definition 3. *Let* $\mathcal{K} = (\mathcal{O}, \Psi)$ *be a HKB with* $\mathsf{sig}(\Psi) = \mathcal{A}$*, and let* $\mathfrak{I} = (\mathcal{I}, f)$ *be an interpretation for* $\mathcal{K}$*. We say that* $\mathfrak{I}$ is a model of $\mathcal{K}$ *if* $\mathcal{I} \models \mathcal{O}$ *and the next conditions hold:*

1. *for each atomic concept* $C \in \mathsf{sig}(\mathcal{O})$*:* $C^{\mathcal{I}} = \{o \mid \exists \bar{a}.\, \kappa_C(\bar{a}) = 1$ *and* $o \in f(\bar{a})\}$*;*
2. *for each atomic role* $R \in \mathsf{sig}(\mathcal{O})$*:* $R^{\mathcal{I}} = \{(o, o') \mid \exists \bar{a}, \bar{b}.\, \lambda_R(\bar{a}, \bar{b}) = 1$ *and* $o \in f(\bar{a})$ *and* $o' \in f(\bar{b})\}$*.*

When both conditions 1 and 2 hold, we write $\mathfrak{I} \models \Psi$*.*

In other words, an interpretation $\mathfrak{I} = (\mathcal{I}, f)$ for a HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ is a model of $\mathcal{K}$ if both $\mathcal{I} \models \mathcal{O}$ and $\mathfrak{I} \models \Psi$. As for the condition $\mathcal{I} \models \mathcal{O}$, we require that $\mathcal{I}$ satisfies all the FOL axioms in $\mathcal{O}$. As for $\mathfrak{I} \models \Psi$, once the universe of discourse $\Delta^{\mathcal{I}}$ and the function f are fixed, we treat the classifiers in Ψ as *exact mappings*, requiring that the extension of atomic concepts and roles faithfully follows the classifiers in Ψ. So we do not actually have any incomplete information about predicate extensions, since Ψ fully determines them. Thus, like the disjointness axioms in the ontology, the inclusions axioms also act as constraints in a given HKB.

Example 3. *Recall Example 2. Let* $\mathcal{I}$ *be an interpretation for* $\mathcal{O}$ *such that* $\Delta^{\mathcal{I}} = \{o_{\bar{a}} \mid \kappa_{\mathsf{P}}(\bar{a}) = 0 \vee \kappa_{\mathsf{M}}(\bar{a}) = 0\}$*,* $\mathsf{D}^{\mathcal{I}} = \{o_{\bar{a}} \mid \kappa_{\mathsf{D}}(\bar{a}) = 1\}$*,* $\mathsf{P}^{\mathcal{I}} = \{o_{\bar{a}} \mid \kappa_{\mathsf{P}}(\bar{a}) = 1\}$*,* $\mathsf{M}^{\mathcal{I}} = \{o_{\bar{a}} \mid \kappa_{\mathsf{M}}(\bar{a}) = 1\}$*,* $\mathsf{cD}^{\mathcal{I}} = \{(o_{\bar{a}}, o_{\bar{b}}) \mid \lambda_{\mathsf{cD}}(\bar{a}, \bar{b}) = 1\}$*, and* $\mathsf{cB}^{\mathcal{I}} = \{(o_{\bar{a}}, o_{\bar{b}}) \mid \lambda_{\mathsf{cB}}(\bar{a}, \bar{b}) = 1\}$*. Let* $f \colon \mathbb{F}(\mathcal{A}) \to \mathcal{P}(\Delta^{\mathcal{I}})$ *be the function such that* $f(\bar{a}) = \emptyset$ *if both* $\kappa_{\mathsf{P}}(\bar{a}) = 1$ *and* $\kappa_{\mathsf{M}}(\bar{a}) = 1$*; otherwise,* $f(\bar{a}) = \{o_{\bar{a}}\}$*. Consider now the interpretation* $\mathfrak{I} = (\mathcal{I}, f)$ *for* $\mathcal{K}$*. By construction, we have* $\mathfrak{I} \models \Psi$*. Furthermore, since* $\mathfrak{I}$ *discards all feature space elements classified positively by both* κ_{P} *and* κ_{M}*, we can conclude that* $\mathcal{I} \models \mathcal{O}$*. It follows that* $\mathfrak{I}$ *is a model of* $\mathcal{K}$*.*

We say that a HKB is *consistent* if it has at least one model and *inconsistent* otherwise. Inconsistency can arise due to an unsatisfiable set of axioms (e.g. $\mathcal{O} = \{\forall x.\, (C(x) \wedge \neg C(x))\}$) or due to the classifiers being incompatible with the ontology axioms (e.g. $\mathcal{O} = \{\exists x.\, C(x)\}$ but $\kappa_C(\bar{a}) = 0$, for each $\bar{a} \in \mathbb{F}(\mathcal{A})$). Although consistent, the next example shows that a HKB may admit only trivial models.

Example 4. *Consider a HKB* $\mathcal{K}' = (\mathcal{O}', \Psi')$ *of a university scenario, where* $\mathsf{sig}(\mathcal{O}')$ *contains the atomic concepts* AP*,* FP*,* FL*,* F*, and* L *for associate professors, full professors, foreign lab coordinators, foreigners, and lab coordinators, respectively. The set of axioms in* $\mathcal{O}'$ *is* $\{\mathsf{AP} \sqsubseteq \neg\mathsf{FP}, \mathsf{FL} \sqsubseteq \mathsf{L}, \mathsf{FL} \sqsubseteq \mathsf{F}\}$*. The set* $\Psi' = \{\kappa_{\mathsf{AP}}, \kappa_{\mathsf{FP}}, \kappa_{\mathsf{FL}}, \kappa_{\mathsf{F}}, \kappa_{\mathsf{L}}\}$ *of classifiers is such that* $\mathsf{sig}(\Psi') = \mathcal{A}' = \{A_1, A_2\}$*, with* $D_{A_1} = D_{A_2} = \{0, 1\}$*. Furthermore, for each* $\bar{a} = (a_1, a_2) \in \mathbb{F}(\mathcal{A}')$*, we have* $\kappa_{\mathsf{AP}}(\bar{a}) = 1$ *if and only if* $a_1 - 1 \geq 0$*;* $\kappa_{\mathsf{FP}}(\bar{a}) = 1$ *if and only if* $a_1 + a_2 - 1 \geq 0$*;* $\kappa_{\mathsf{FL}}(\bar{a}) = 1$ *if and only if* $-a_1 - a_2 + 1 \geq 0$*;* $\kappa_{\mathsf{F}}(\bar{a}) = 1$ *if and only if* $a_2 - 1 \geq 0$*; and* $\kappa_{\mathsf{L}}(\bar{a}) = 1$ *if and only if* $-a_2 \geq 0$*.*

One can easily verify that each $\bar{a} \in \mathbb{F}(\mathcal{A}')$ *is such that either* (i) $\kappa_{\mathsf{AP}}(\bar{a}) = 1 \wedge \kappa_{\mathsf{FP}}(\bar{a}) = 1$ *or* (ii) $\kappa_{\mathsf{FL}}(\bar{a}) = 1$*. Consider now an element* $\bar{a} \in \mathbb{F}(\mathcal{A}')$*. If* $\bar{a}$ *satisfies* (i)*, then it should be discarded due to* $\mathsf{AP} \sqsubseteq \neg\mathsf{FP}$*. If* $\bar{a}$ *satisfies* (ii)*, then one can see that either* $\kappa_{\mathsf{L}}(\bar{a}) = 0$ *or* $\kappa_{\mathsf{F}}(\bar{a}) = 0$*, and therefore it should be discarded due to* $\mathsf{FL} \sqsubseteq \mathsf{L}$ *or* $\mathsf{FL} \sqsubseteq \mathsf{F}$*, respectively. Thus, we get that* $\mathcal{K}'$ *admits only trivial models.*

Trivial models inherently suppress the information inferred from the classifiers in Ψ, as they discard all the feature space elements. It is thus natural to focus only on non-trivial models, provided they exist. More generally, rather than considering all models of a HKB, it is reasonable to restrict the attention to those that retain maximal information from the application of the classifiers over the sub-symbolic data. In this paper, we focus on models that discard feature space elements in a minimal fashion, according to set inclusion.

Definition 4. *Given a HKB* $\mathcal{K} = (\mathcal{O}, \Psi)$ *and an interpretation* $\mathfrak{I} = (\mathcal{I}, f)$ *for* $\mathcal{K}$*, we say that* $\mathfrak{I}$ *is a* minimally-discarding model of $\mathcal{K}$ *if* (i) $\mathfrak{I}$ *is a model of* $\mathcal{K}$ *and* (ii) *there is no model* $\mathfrak{I}' = (\mathcal{I}', f')$ *of* $\mathcal{K}$ *such that* $\mathsf{disc}(\mathfrak{I}') \subsetneq \mathsf{disc}(\mathfrak{I})$*.*

Given a HKB $\mathcal{K}$*, we denote by* $\mathsf{MinDisc}(\mathcal{K})$ *the set of minimally-discarding models of* $\mathcal{K}$*.*

Interestingly, a HKB may admit minimally-discarding models that discard different subsets of the feature space.

Example 5. *Recall Example 2. Let* $\mathcal{K}'' = (\mathcal{O}'', \Psi)$ *be the HKB obtained from* $\mathcal{K}$ *by adding to* $\mathcal{O}$ *the axiom:* $\mathsf{cD} \sqsubseteq \mathsf{cB}$*. Consider* $\bar{a} = (1, 1, 1, 2, 0)$ *and* $\bar{b} = (1, 0, 2, 2, 1)$*. Note that* (i) $\lambda_{\mathsf{cD}}(\bar{a}, \bar{b}) = 1$ *and* $\lambda_{\mathsf{cB}}(\bar{a}, \bar{b}) = 0$*, and* (ii) $\kappa_{\mathsf{M}}(\bar{a}) = \kappa_{\mathsf{M}}(\bar{b}) = 0$ *and* $\lambda_{\mathsf{cB}}(\bar{a}, \bar{a}) = \lambda_{\mathsf{cB}}(\bar{b}, \bar{b}) = 1$*. Due to* (i)*, we derive that there can be no model* $\mathfrak{I}$ *of* $\mathcal{K}''$ *such that both* $\bar{a} \notin \mathsf{disc}(\mathfrak{I})$ *and* $\bar{b} \notin \mathsf{disc}(\mathfrak{I})$*. Due to* (ii)*, we derive that there can be models* $\mathfrak{I}$ *of* $\mathcal{K}''$ *such that* $\bar{a} \notin \mathsf{disc}(\mathfrak{I})$ *and models* $\mathfrak{I}'$ *of* $\mathcal{K}''$ *such that* $\bar{b} \notin \mathsf{disc}(\mathfrak{I}')$*. As a result, there exist at least two distinct minimally-discarding models of* $\mathcal{K}''$ *that discard different subsets of the feature space.*

The notion of minimally-discarding model shares similar characteristics with the notion of *repair*, widely studied in the literature on consistent query answering. To draw such a correspondence, we associate to each tuple $\mathcal{A}$ of attributes with a finite domain a *set of facts* $D_{\mathcal{A}} = \{s(c_{\bar{a}}) \mid \bar{a} \in \mathbb{F}(\mathcal{A})\}$, where $s \in \Sigma_C$ is an atomic concept assumed to be never mentioned in any ontology $\mathcal{O}$ and $c_{\bar{a}}$ is a *constant* associated with element $\bar{a}$. To each HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ with $\mathsf{sig}(\Psi)$ being a tuple $\mathcal{A}$ of attributes with finite domains, we associate a *knowledge base (KB)* $\mathcal{B}_{\mathcal{K}} = (\Sigma_{\mathcal{K}}, D_{\mathcal{A}})$, where $\Sigma_{\mathcal{K}}$ is a set of axioms that includes the axioms in $\mathcal{O}$ together with the following set of axioms: $\{s(c_{\bar{a}}) \rightarrow C(c_{\bar{a}}) \mid C \in \mathsf{sig}(\mathcal{O})$ is an atomic concept and $\kappa_C(\bar{a}) = 1\} \cup \{s(c_{\bar{a}}) \wedge s(c_{\bar{b}}) \rightarrow R(c_{\bar{a}}, c_{\bar{b}}) \mid R \in \mathsf{sig}(\mathcal{O})$ is an atomic role and $\lambda_R(\bar{a}, \bar{b}) = 1\}$. A *repair* of a KB $\mathcal{B} = (\Sigma, D)$ formed by a set of facts D and a set of axioms Σ is a $\subseteq$-maximal set of facts $\mathcal{R} \subseteq D$ such that $\mathcal{B}' = (\Sigma, \mathcal{R})$ is consistent. Given a HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ with $\mathsf{sig}(\Psi)$ being a tuple $\mathcal{A}$ of attributes with finite domains, a set of facts $\mathcal{R} \subseteq D_{\mathcal{A}}$, and a model $\mathfrak{I}$ of $\mathcal{K}$ such that $\mathsf{disc}(\mathfrak{I}) = \{\bar{a} \mid s(c_{\bar{a}}) \in D_{\mathcal{A}} \setminus \mathcal{R}\}$, one can verify that $\mathcal{R}$ is a repair of $\mathcal{B}_{\mathcal{K}}$ if and only if $\mathfrak{I} \in \mathsf{MinDisc}(\mathcal{K})$.

4.1 Reasoning Tasks over HKBs

In any Ontology-Driven Knowledge Base framework, the typical form of interaction to extract information consists in posing queries over the vocabulary of the ontology. We adopt the same interaction model in our framework, and formally define the notion of answers to queries over HKBs.

Definition 5. *Let* $\mathcal{K} = (\mathcal{O}, \Psi)$ *be a HKB, let* $q = \{\bar{x} \mid \varphi(\bar{x})\}$ *be a FOL query over* $\mathcal{O}$ *of arity* m*, and let* $\bar{t} = (\bar{a}_1, \ldots, \bar{a}_m)$ *be an* m*-tuple of elements from* $\mathbb{F}(\mathcal{A})$*, i.e.* $\bar{t} \in \mathbb{F}(\mathcal{A})^m$*. We say that* $\bar{t}$ *is a* skeptical-answer to q over $\mathcal{K}$ *if, for every* $\mathfrak{I} =$

$(\mathcal{I}, f) \in \mathsf{MinDisc}(\mathcal{K})$*, the following condition holds: there exists an* m*-tuple* $\bar{o} = (o_1, \ldots, o_m)$ *of objects from* $\Delta^{\mathcal{I}}$ *such that* $\mathcal{I} \models \varphi(\bar{x}/\bar{o})$ *and* $o_i \in f(\bar{a}_i)$*, for each* $i \in [m]$*.*

For a HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ *and a FOL query* q *over* $\mathcal{O}$*, we let* $\mathsf{Sans}(q, \mathcal{K})$ *be the set of skeptical-answers to* q *over* $\mathcal{K}$*.*

In other words, in the absence of a principled criterion to single out a unique "correct" minimally-discarding model, we adopt a cautious approach by considering only those answers obtainable by all minimally-discarding models. This form of reasoning is reminiscent of classical skeptical reasoning over all repairs of a KB, a common strategy for handling query answering with respect to inconsistent KBs (Lembo et al. 2010; Bienvenu and Bourgaux 2016).

Example 6. *Recall Example 3, and consider the UCQ*$^{\neq}$ $q = \{(x, y) \mid \mathsf{cD}(x, y) \wedge x \neq y \wedge (\mathsf{P}(x) \wedge \mathsf{P}(y)) \vee (\mathsf{D}(x) \wedge \mathsf{D}(y))\}$ *over* $\mathcal{O}$*, asking for pairs of distinct patients that are compatible blood donors and are either both pregnant or both diabetic. Consider now* $\bar{a} = (1, 2, 1, 0, 2)$ *and* $\bar{b} = (0, 2, 1, 1, 2)$*. It is not hard to verify that* $(\bar{a}, \bar{b}) \in \mathsf{Sans}(q, \mathcal{K})$*.*

In what follows, we say that a HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ is an $(\mathcal{L}_{\mathcal{O}}, \mathcal{L}_{\mathcal{C}})$-*HKB* if (i) $\mathcal{O}$ is specified in the fragment $\mathcal{L}_{\mathcal{O}}$ of FOL and (ii) each classifier in Ψ belongs to the class $\mathcal{L}_{\mathcal{C}}$ of classifiers. Given the general framework presented so far, it is natural to consider the following reasoning tasks, for specific FOL fragments $\mathcal{L}_{\mathcal{O}}$ of ontology languages, classes $\mathcal{L}_{\mathcal{C}}$ of classifiers, and FOL fragments $\mathcal{L}_{\mathcal{Q}}$ of query languages:

- *(non-trivial) consistency*: given a $(\mathcal{L}_{\mathcal{O}}, \mathcal{L}_{\mathcal{C}})$-HKB $\mathcal{K}$, check whether $\mathcal{K}$ admits a (non-trivial) model;
- *skeptical entailment*: given a $(\mathcal{L}_{\mathcal{O}}, \mathcal{L}_{\mathcal{C}})$-HKB $\mathcal{K}$, a query $q \in \mathcal{L}_{\mathcal{Q}}$, and a tuple $\bar{t}$, check whether $\bar{t} \in \mathsf{Sans}(q, \mathcal{K})$.

We are interested in the *classifiers complexity* variant of the above decision problems, which is the complexity where only the set Ψ of classifiers is regarded as the input, while the ontology and the query (if applicable) are assumed to be fixed. This complexity measure is analogous to the widely studied *data complexity* (Vardi 1982) measure of reasoning tasks over KBs, where the input is only the set of facts.

5 Computational Complexity Analysis

We now provide a detailed computational complexity analysis of the decision problems defined in the previous section. In this first investigation, we focus on the setting in which $\mathcal{L}_{\mathcal{O}} = \textit{DL-Lite}^{\neg}_{\mathtt{RDFS}}$, $\mathcal{L}_{\mathcal{Q}}$ is either CQ or $\mathsf{UCQ}^{\neq}$, and the classifiers operate on tuples of attributes having a *finite* domain. More specifically, similarly to other theoretical works that study reasoning tasks over ML models (Barceló et al. 2020a; Arenas et al. 2021; Alfano et al. 2025), we let the domain of each attribute be $\{0, 1\}$. This is only for the sake of presentation, and we observe that all our results can be easily extended to the case in which the domains are finite sets of arbitrary size. We further assume that each binary classifier over pairs $\lambda_R \colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ is treated as a standard classifier $\kappa_R \colon \{0,1\}^{2n} \to \{0,1\}$ operating on concatenated input pairs, i.e. $\lambda_R(x, x') = 1$ if and only if $\kappa_R(x \| x') = 1$. Within this context, we consider the class of `Multi-Layer Perceptron` (`MLP`), i.e. $\mathcal{L}_{\mathcal{C}} = \mathtt{MLP}$.

As defined in (Barceló et al. 2020a), an MLP $\mathcal{M}$ is a function $\mathcal{M} \colon \{0,1\}^l \to \{0,1\}$ defined by a sequence of *weight* matrices $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(k)}$, *bias* vectors $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(k)}$, and *activation* functions $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(k)}$, where k is the number of *layers* of $\mathcal{M}$. Given $\bar{a} \in \{0,1\}^l$, for $i \in [k]$, assuming that $\mathbf{h}^{(0)} = \bar{a}$, we inductively define $\mathbf{h}^{(i)} = \mathbf{a}^{(i)}(\mathbf{h}^{(i-1)}\mathbf{W}^{(i)} + \mathbf{b}^{(i)})$. The output of $\mathcal{M}$ on $\bar{a}$ is defined as $\mathcal{M}(\bar{a}) = \mathbf{h}^{(k)}$. We assume all weights and biases to be rational numbers from $\mathbb{Q}$. That is, we assume that there exists a sequence of positive integers $d_0, \ldots, d_k$ such that $\mathbf{W}^{(i)} \in \mathbb{Q}^{d_{i-1} \times d_i}$ and $\mathbf{b}^{(i)} \in \mathbb{Q}^{d_i}$, for $i \in [k]$. Given that we are interested in binary classifiers, we assume that $d_k = 1$. The size of an MLP is the total size of its weights and biases, in which the size of a rational number $\frac{p}{q}$ is $\mathsf{log}_2(p) + \mathsf{log}_2(q)$ (with the convention that $\mathsf{log}_2(0) = 1$). We focus on MLPs in which all internal functions $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(k-1)}$ are the ReLU function $\mathsf{relu}(x) = \max(0, x)$. Usually, MLP binary classifiers are trained using the sigmoid as the output function $\mathbf{a}^{(k)}$. Nevertheless, when an MLP classifies an input (after training), it takes decisions by simply using the preactivations, also called logits. Based on this and on the fact that we only consider already trained MLPs, we can assume without loss of generality that the output function $\mathbf{a}^{(k)}$ is the binary *step* function $\mathsf{step}(x) = 1$ if $x \geq 0$ and $\mathsf{step}(x) = 0$ otherwise.

For instance, the HKB $\mathcal{K}'$ defined in Example 4 is a $(\textit{DL-Lite}^{\neg}_{\mathtt{RDFS}}, \mathtt{MLP})$-HKB. It is straightforward to verify that a $(\textit{DL-Lite}^{\neg}_{\mathtt{RDFS}}, \mathtt{MLP})$-HKB always admit trivial models. So, in our setting, while the consistency problem becomes trivial, we now characterize the complexity of verifying whether a given $(\textit{DL-Lite}^{\neg}_{\mathtt{RDFS}}, \mathtt{MLP})$-HKB admit non-trivial models.

Theorem 1. *For* $\textit{DL-Lite}^{\neg}_{\mathtt{RDFS}}$ *ontologies and* `MLP` *classifiers, the non-trivial consistency problem is* NP*-complete in classifiers complexity.*

The result comes from the observation that a non-trivial model exists if and only if there exists a non *self-conflicting* feature space element, which can be guessed in nondeterministic polynomial time. Intuitively, a self-conflicting feature space element is an element discarded by all models.

We now turn to the query answering problem over HKBs. Unfortunately, the following result proves that this problem is highly intractable, even in classifiers complexity and for the query language $\mathcal{L}_{\mathcal{Q}} = \mathsf{CQ}$. The upper bound is from a simple guess-and-check procedure, i.e. guess the (in general exponentially large) model that serves as a counterexample to query entailment, while the lower bound is from the complement of the 3-colorability problem for succinct graphs, a NEXPTIME-complete problem (Papadimitriou and Yannakakis 1986). The lower bound exploits the fact that from a Boolean circuit it is possible to obtain in polynomial time an equivalent MLP (Barceló et al. 2020b, Lemma 13).

Theorem 2. *For* $\textit{DL-Lite}^{\neg}_{\mathtt{RDFS}}$ *ontologies,* `MLP` *classifiers, and queries in* $\mathsf{UCQ}^{\neq}$*, the skeptical entailment problem is* CONEXPTIME*-complete in classifiers complexity. The hardness holds already for queries in* CQ*.*

In light of the above negative result, we further investigate the query answering problem over HKBs by considering a

fragment of the ontology language *DL-Lite*$^{\neg}_{\text{RDFS}}$ as well as an alternative semantics for query answers that avoid skeptically reasoning over all the minimally-discarding models.

We point out that in Theorem 2 the main source of complexity comes from the usage of axioms involving atomic roles. This naturally leads us to consider the fragment $\mathcal{H}^{\neg}$ of *DL-Lite*$^{\neg}_{\text{RDFS}}$ which forbids the presence of atomic roles in axioms. Note that $\mathcal{H}^{\neg}$ remains a useful ontology language, as it allows to express taxonomies (i.e. hierarchy of atomic concepts) as well as disjointness between atomic concepts. For instance, the ontologies $\mathcal{O}$ and $\mathcal{O}'$ illustrated, respectively, in Example 2 and Example 4 are $\mathcal{H}^{\neg}$ ontologies.

Interestingly, while for non-trivial consistency the lower bound provided in Theorem 1 already applies to $\mathcal{L}_{\mathcal{O}} = \mathcal{H}^{\neg}$, we now show that the complexity of the query answering problem significantly decreases for $(\mathcal{H}^{\neg}, \texttt{MLP})$-HKBs.

Theorem 3. *For $\mathcal{H}^{\neg}$ ontologies, MLP classifiers, and queries in $\mathsf{UCQ}^{\neq}$, the skeptical entailment problem becomes* NP*-complete in classifiers complexity. The hardness holds already for queries in CQ.*

The complexity decreases are due to the fact that each minimally-discarding model of a $(\mathcal{H}^{\neg}, \texttt{MLP})$-HKB $\mathcal{K}$ discards only the self-conflicting feature space elements.

5.1 Alternative Semantics for Query Answering

We now consider a sound approximation of the query answering semantics provided in Definition 5. This semantics is inspired by the well-behaved *IAR semantics* adopted for query answering over inconsistent KBs (Lembo et al. 2015), which follows the *WIDTIO* (*When In Doubt Throw It Out*) approach of the belief revision and update research area.

Definition 6. *Let $\mathcal{K} = (\mathcal{O}, \Psi)$ with $\mathsf{sig}(\Psi) = \mathcal{A}$ be a (DL-Lite$^{\neg}_{\text{RDFS}}$, MLP)-HKB, and let $\mathfrak{I}$ be a model of $\mathcal{K}$. We say that $\mathfrak{I}$ is a* minimally-discarding WIDTIO-model *of $\mathcal{K}$ if, for every $\bar{a} \in \mathbb{F}(\mathcal{A})$, the following holds: $\bar{a} \in \mathsf{disc}(\mathfrak{I})$ if and only if there exists a $\mathfrak{I}' \in \mathsf{MinDisc}(\mathcal{K})$ such that $\bar{a} \in \mathsf{disc}(\mathfrak{I}')$.*

Let now $q = \{\bar{x} \mid \varphi(\bar{x})\}$ be a UCQ$^{\neq}$ query over $\mathcal{O}$ of arity m and $\bar{t} = (\bar{a}_1, \ldots, \bar{a}_m)$ be an m-tuple of elements from $\mathbb{F}(\mathcal{A})$. We say that $\bar{t}$ is a WIDTIO-answer to q over $\mathcal{K}$ *if, for every minimally-discarding WIDTIO-model $\mathfrak{I}$ of $\mathcal{K}$: there exists an m-tuple $\bar{o} = (o_1, \ldots, o_m)$ of objects from $\Delta^{\mathcal{I}}$ such that $\mathcal{I} \models \varphi(\bar{x}/\bar{o})$ and $o_i \in f(\bar{a}_i)$, for each $i \in [m]$.*

For a (DL-Lite$^{\neg}_{\text{RDFS}}$, MLP)-HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ and a UCQ$^{\neq}$ query q over $\mathcal{O}$, we denote by $\mathsf{Wans}(q, \mathcal{K})$ the set of WIDTIO-answers to q over $\mathcal{K}$.

In other words, a WIDTIO-model implements the WIDTIO approach by discarding all and only the feature space elements discarded by at least one minimally-discarding model. Thus, the evaluation of a query under the WIDTIO semantics reasons only on those feature space elements that belong to every possible minimally-discarding model.

It is not hard to see that a minimally-discarding WIDTIO-model of a (*DL-Lite*$^{\neg}_{\text{RDFS}}$, MLP)-HKB $\mathcal{K}$ always exists (at worst, it will be a trivial model). Furthermore, this semantics is a sound approximation of the previously studied semantics, in the sense that $\mathsf{Wans}(q, \mathcal{K}) \subseteq \mathsf{Sans}(q, \mathcal{K})$ holds for every (*DL-Lite*$^{\neg}_{\text{RDFS}}$, MLP)-HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ and UCQ$^{\neq}$ query q over $\mathcal{O}$. Actually, the next example shows that there are cases in which the subset relation can even be strict.

Example 7. *Consider the (DL-Lite$^{\neg}_{\text{RDFS}}$, MLP)-HKB $\mathcal{K} = (\mathcal{O}, \Psi)$, where $\mathsf{sig}(\mathcal{O})$ contains the atomic roles B, P, and F, and $\mathcal{O}$ states the axiom $\mathsf{P} \sqsubseteq \neg\mathsf{F}$. The set $\Psi = \{\lambda_{\mathsf{B}}, \lambda_{\mathsf{P}}, \lambda_{\mathsf{F}}\}$ of classifiers is such that $\mathsf{sig}(\Psi) = \mathcal{A} = \{A_1, A_2\}$, and for each pair $(\bar{a}, \bar{b}) \in \mathbb{F}(\mathcal{A}) \times \mathbb{F}(\mathcal{A})$ such that $\bar{a} = (a_1, a_2)$ and $\bar{b} = (b_1, b_2)$, we have: $\lambda_{\mathsf{B}}(\bar{a}, \bar{b}) = \kappa_{\mathsf{B}}(\bar{a}||\bar{b}) = 1$ if and only if $2(a_1 + a_2) - b_1 - b_2 - 3 \geq 0$; $\lambda_{\mathsf{F}}(\bar{a}, \bar{b}) = \kappa_{\mathsf{F}}(\bar{a}||\bar{b}) = 1$ if and only if $a_1 + b_2 - 3(a_2 + b_1) \geq 0$; $\lambda_{\mathsf{P}}(\bar{a}, \bar{b}) = \kappa_{\mathsf{P}}(\bar{a}||\bar{b}) = 1$ if and only if $-a_1 - b_2 - 2(a_2 + b_1) + 2 \geq 0$. Let $\bar{a} = (1, 1)$, $\bar{b} = (1, 0)$, $\bar{c} = (0, 1)$, and $\bar{d} = (0, 0)$. Since $(\bar{b}, \bar{c})$, $(\bar{d}, \bar{c})$, $(\bar{b}, \bar{d})$, and $(\bar{d}, \bar{d})$ are the pairs (x, x') such that $\kappa_{\mathsf{P}}(x||x') = \kappa_{\mathsf{F}}(x||x') = 1$, we derive that every $\mathfrak{I} \in \mathsf{MinDisc}(\mathcal{K})$ is such that either $\mathsf{disc}(\mathfrak{I}) = \{\bar{d}, \bar{b}\}$ or $\mathsf{disc}(\mathfrak{I}) = \{\bar{d}, \bar{c}\}$. Consider now the CQ $q = \{(x) \mid \exists y.\mathsf{B}(x, y)\}$ over $\mathcal{O}$. Since $(\bar{a}, \bar{d})$, $(\bar{a}, \bar{b})$, and $(\bar{a}, \bar{c})$ are the pairs (x, x') such that $\kappa_{\mathsf{B}}(x||x') = 1$, we get that $\mathsf{Sans}(q, \mathcal{K}) = \{\bar{a}\}$ while $\mathsf{Wans}(q, \mathcal{K}) = \emptyset$.*

We now study the *WIDTIO-entailment* problem, for $\mathcal{L}_{\mathcal{O}} \in \{\textit{DL-Lite}^{\neg}_{\text{RDFS}}, \mathcal{H}^{\neg}\}$ and $\mathcal{L}_{\mathcal{Q}} = \{\mathsf{CQ}, \mathsf{UCQ}^{\neq}\}$, defined as the problem of deciding, given a $(\mathcal{L}_{\mathcal{O}}, \texttt{MLP})$-HKB $\mathcal{K}$, a query $q \in \mathcal{L}_{\mathcal{Q}}$, and a tuple $\bar{t}$, whether $\bar{t} \in \mathsf{Wans}(q, \mathcal{K})$. We start with $\mathcal{L}_{\mathcal{O}} = \mathcal{H}^{\neg}$, and prove that the two query answering semantics actually coincide in this setting (in fact, this is an immediate consequence of the next result).

Proposition 1. *Let $\mathcal{K}$ be a $(\mathcal{H}^{\neg}, \texttt{MLP})$-HKB and let $\mathfrak{I}$ be an interpretation for $\mathcal{K}$. We have that $\mathfrak{I} \in \mathsf{MinDisc}(\mathcal{K})$ if and only if $\mathfrak{I}$ is a WIDTIO-model of $\mathcal{K}$.*

It immediately follows that the complexity results derived in Theorem 3 also apply to the WIDTIO-entailment problem. We conclude this section by analyzing the remaining case of the complexity of query answering over (*DL-Lite*$^{\neg}_{\text{RDFS}}$, MLP)-HKBs under the WIDTIO semantics.

Theorem 4. *For DL-Lite$^{\neg}_{\text{RDFS}}$ ontologies, MLP classifiers, and queries in $\mathsf{UCQ}^{\neq}$, the WIDTIO entailment problem is Σ^p_2-complete in classifiers complexity. The hardness holds already for queries in CQ.*

For the upper bound, it is enough to guess an assignment to the variables that makes the query true, and check whether the guessed assignment involves only feature space elements that are never discarded in minimally-discarding models.

6 Conclusion and Future Work

We presented a novel framework that provides semantic context for the knowledge induced by ML models. This framework allows us to define formal reasoning tasks over the knowledge derived from both symbolic (i.e. ontologies) and sub-symbolic (i.e. classifiers) representations. Finally, we studied the complexity of two fundamental reasoning tasks in this context: consistency checking and query answering.

Directions for future work are many. Firstly, we would like to extend HKBs to account for the probabilistic nature of ML outputs. Secondly, we would like to study the complexity of other reasoning tasks connected to HKBs. Finally, it would be valuable to provide an ASP encoding for the query answering problem (under the WIDTIO approach) and validate it empirically with a thorough experimental evaluation.

Acknowledgments

This work has been supported by MUR under the PNRR project FAIR (PE0000013).

References

Alfano, G.; Greco, S.; Mandaglio, D.; Parisi, F.; Shahbazian, R.; and Trubitsyna, I. 2025. Even-if Explanations: Formal Foundations, Priorities and Complexity. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI 25)*, 15347–15355.

Arakelyan, E.; Daza, D.; Minervini, P.; and Cochez, M. 2021. Complex Query Answering with Neural Link Predictors. In *Proceedings of the Ninth International Conference on Learning Representations (ICLR 2021)*.

Arenas, M.; Báez, D.; Barceló, P.; Pérez, J.; and Subercaseaux, B. 2021. Foundations of Symbolic Languages for Model Interpretability. In *Proceedings of the Thirty-Fourth Annual Conference on Neural Information Processing Systems (NeurIPS 2021)*, 11690–11701.

Bahoo, S.; Cucculelli, M.; Goga, X.; and Mondolo, J. 2023. Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis. *SN Business & Economics*, 4(23).

Barceló, P.; Monet, M.; Pérez, J.; and Subercaseaux, B. 2020a. Model Interpretability through the lens of Computational Complexity. In *Proceedings of the Thirty-Third Annual Conference on Neural Information Processing Systems (NeurIPS 2020)*.

Barceló, P.; Monet, M.; Pérez, J.; and Subercaseaux, B. 2020b. Model Interpretability through the Lens of Computational Complexity. *CoRR*, abs/2010.12265.

Bienvenu, M.; and Bourgaux, C. 2016. Inconsistency-Tolerant Querying of Description Logic Knowledge Bases. In Pan, J. Z.; Calvanese, D.; Eiter, T.; Horrocks, I.; Kifer, M.; Lin, F.; and Zhao, Y., eds., *Proceedings of the Twelfth International Summer School on Reasoning Web: Logical Foundation of Knowledge Graph Construction and Query Answering (RW 2016)*, volume 9885 of *Lecture Notes in Computer Science*, 156–202. Springer.

Bienvenu, M.; and Ortiz, M. 2015. Ontology-Mediated Query Answering with Data-Tractable Description Logics. In *Proceedings of the Eleventh International Summer School Tutorial Lectures (RW 2015)*, 218–307.

Bishop, C. M. 2007. *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer.

Chen, H.; Chiang, R. H. L.; and Storey, V. C. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4): 1165–1188.

Cima, G.; Console, M.; Delfino, R. M.; Lenzerini, M.; and Poggi, A. 2025. Answering Conjunctive Queries with Safe Negation and Inequalities over RDFS Knowledge Bases. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI 25)*, 14824–14831.

Cima, G.; Lenzerini, M.; and Poggi, A. 2020. Answering Conjunctive Queries with Inequalities in *DL-Lite*$_{\mathscr{R}}$. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*, 2782–2789.

Cuenca Grau, B. 2004. A possible simplification of the semantic web architecture. In *Proceedings of the Thirteenth International World Wide Web Conference (WWW 2004)*, 704–713.

Fischer, M.; Balunovic, M.; Drachsler-Cohen, D.; Gehr, T.; Zhang, C.; and Vechev, M. T. 2019. DL2: Training and Querying Neural Networks with Logic. In *Proceedings of the Thirty-Sixth International Conference on Machine Learning (ICML 2019)*, 1931–1941.

Giunchiglia, E.; Stoian, M. C.; and Lukasiewicz, T. 2022. Deep Learning with Logical Constraints. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI 2022)*, 5478–5485.

Giunchiglia, E.; Tatomir, A.; Stoian, M. C.; and Lukasiewicz, T. 2024. CCN+: A neuro-symbolic framework for deep learning with requirements. *International Journal of Approximate Reasoning*, 171: 109124.

Hamilton, W. L.; Bajaj, P.; Zitnik, M.; Jurafsky, D.; and Leskovec, J. 2018. Embedding Logical Queries on Knowledge Graphs. In *Proceedings of the Thirty-First Annual Conference on Advances in Neural Information Processing Systems (NeurIPS 2018)*, 2030–2041.

Hastie, T.; Tibshirani, R.; and Friedman, J. H. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer.

Hitzler, P.; and Sarker, M. K. 2022. *Neuro-symbolic artificial intelligence: The state of the art*. IOS press.

Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; and Wang, Y. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4).

König, M.; Bosman, A. W.; Hoos, H. H.; and van Rijn, J. N. 2024. Critically Assessing the State of the Art in Neural Network Verification. *Journal of Machine Learning Research*, 25: 12:1–12:53.

Lembo, D.; Lenzerini, M.; Rosati, R.; Ruzzi, M.; and Savo, D. F. 2010. Inconsistency-tolerant Semantics for Description Logics. In *Proceedings of the Fourth International Conference on Web Reasoning and Rule Systems (RR 2010)*, 103–117.

Lembo, D.; Lenzerini, M.; Rosati, R.; Ruzzi, M.; and Savo, D. F. 2015. Inconsistency-tolerant query answering in ontology-based data access. *Journal of Web Semantics*, 33: 3–29.

Papadimitriou, C. H.; and Yannakakis, M. 1986. A Note on Succinct Representations of Graphs. *Information and Control*, 71(3): 181–185.

Poggi, A.; Lembo, D.; Calvanese, D.; De Giacomo, G.; Lenzerini, M.; and Rosati, R. 2008. Linking Data to Ontologies. *Journal on Data Semantics*, X: 133–173.

Rosati, R. 2007. The Limits of Querying Ontologies. In *Proceedings of the Eleventh International Conference on Database Theory (ICDT 2007)*, volume 4353 of *Lecture Notes in Computer Science*, 164–178. Springer.

Thames, C.; and Sun, Y. 2024. A Survey of Artificial Intelligence Approaches to Safety and Mission-Critical Systems. In *2024 Integrated Communications, Navigation and Surveillance Conference (ICNS)*, 1–12.

Vardi, M. Y. 1982. The Complexity of Relational Query Languages (Extended Abstract). In *Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing (STOC 1982)*, 137–146.

Wang, Q.; Mao, Z.; Wang, B.; and Guo, L. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE transactions on knowledge and data engineering*, 29(12): 2724–2743.

Xiao, G.; Ding, L.; Cogrel, B.; and Calvanese, D. 2019. Virtual Knowledge Graphs: An Overview of Systems and Use Cases. *Data Intelligence*, 1(3): 201–223.

Xu, J.; Zhang, Z.; Friedman, T.; Liang, Y.; and den Broeck, G. V. 2018. A Semantic Loss Function for Deep Learning with Symbolic Knowledge. In *Proceedings of the Thirty-Fifth International Conference on Machine Learning (ICML 2018)*, 5498–5507.