Contents lists available at IJAHCI
International Journal of Advanced Human Computer Interaction
Journal Homepage: http://www.ijahci.com/
Volume 5, No. 5, 2026

Ontology-Driven User Intent Prediction in Conversational Interfaces Using Semantic Enrichment Pipelines

Vijay Banerjee[1], **Sneha Singh**[2]
[1] *Department of Human-Computer Interaction, Indian Institute of Technology Delhi*
[2] *Department of Human-Computer Interaction, Indian Institute of Technology Delhi*

ARTICLE INFO

Received: 04/22/2026
Revised: 05/26/2026
Accepted: 06/12/2026

Keywords:

Ontology-Driven Systems, User Intent Prediction, Conversational Interfaces, Semantic Enrichment, Natural Language Understanding, Knowledge Representation, Dialogue Management

ABSTRACT

The accurate prediction of user intent in conversational interfaces remains a fundamental challenge in natural language understanding, particularly when utterances are ambiguous, context-dependent, or domain-specific. Conventional intent classification approaches rely predominantly on statistical co-occurrence patterns and surface-level lexical features, rendering them brittle in the face of semantic variation and knowledge gaps. This paper presents a novel framework that integrates formal ontological structures with semantic enrichment pipelines to substantially improve intent prediction accuracy in dialogue systems.

The proposed architecture leverages domain ontologies to inject structured background knowledge into the intent recognition process, enabling the system to resolve lexical ambiguity, infer implicit user goals, and generalize across semantically related expressions. By aligning user utterances with ontological concepts through a multi-stage enrichment pipeline—comprising entity linking, concept expansion, and relation-aware encoding—the framework produces semantically grounded representations that transcend the limitations of purely data-driven methods.

Empirical evaluation across three benchmark conversational datasets demonstrates that ontology-augmented models achieve statistically significant improvements over state-of-the-art baselines, yielding gains of up to 8.3% in macro-averaged F1-score and 11.2% reduction in out-of-vocabulary intent misclassification. Ablation studies further confirm that each constituent stage of the semantic enrichment pipeline contributes independently and cumulatively to overall predictive performance.

The findings establish that principled integration of symbolic knowledge representations with neural language models constitutes a robust and interpretable strategy for intent understanding in conversational AI. This work contributes both a reproducible experimental methodology and a publicly available ontology-enriched evaluation benchmark, advancing the broader agenda of knowledge-informed natural language processing in human-computer interaction systems.

1. Introduction

The proliferation of conversational interfaces—spanning voice assistants, chatbots, dialogue systems, and mul-

timodal agents—has fundamentally transformed the landscape of human-computer interaction over the past decade. As these systems become increasingly embedded in domains ranging from healthcare and education to e-commerce and enterprise automation, the challenge of accurately predicting user intent from natural language inputs has emerged as one of the most consequential problems in applied artificial intelligence [5]. Unlike traditional keyword-based query processing, intent prediction in conversational contexts demands a nuanced understanding of semantic meaning, contextual dependencies, and the latent goals that users express—often imprecisely and ambiguously—through natural language utterances [18]. The gap between what a user says and what a user means constitutes the central epistemological challenge that motivates this research.

Contemporary approaches to intent prediction have leveraged a variety of machine learning paradigms, including recurrent neural networks, transformer-based language models, and graph neural networks, each offering distinct advantages and limitations [9]. However, a persistent shortcoming across these methodologies is their relative neglect of structured world knowledge—the kind of formal, hierarchically organized conceptual knowledge that ontologies provide. Ontologies, as formal representations of domain concepts, properties, and their interrelationships, offer a principled mechanism for grounding natural language understanding in semantically rich, machine-interpretable knowledge structures [8]. The integration of ontological reasoning into intent prediction pipelines thus represents a theoretically motivated and practically promising frontier. This paper proposes, formalizes, and empirically evaluates an ontology-driven framework for user intent prediction in conversational interfaces, employing semantic enrichment pipelines that augment raw linguistic features with structured knowledge representations to achieve superior predictive accuracy and interpretability.

1.1. Motivation and Problem Statement

The fundamental motivation for this research arises from a critical observation: state-of-the-art intent classification models, despite their impressive performance on benchmark datasets, frequently fail in real-world deployment scenarios characterized by domain shift, sparse training data, and the inherent ambiguity of natural language [21]. Consider a user interacting with a healthcare conversational agent who asks, "Can you help me understand my medication schedule?" The surface-level linguistic features of this utterance may map to multiple plausible intents—information retrieval, scheduling assistance, medication management, or caregiver support—depending on contextual factors such as the user's medical history, the conversational history, and the broader domain ontology governing the healthcare application. A purely data-driven model trained on labeled intent corpora may assign this utterance to a majority-class intent category, failing to leverage the rich conceptual relationships encoded in medical ontologies such as SNOMED-CT or the Unified Medical Language System (UMLS) [7].

This limitation is not merely an academic concern; it carries significant practical consequences. In high-stakes domains such as clinical decision support, financial advisory services, and legal assistance, misclassified user intents can lead to inappropriate system responses, erosion of user trust, and potentially harmful outcomes [4]. The problem is further compounded by the multi-turn nature of conversational interactions, where the intent of a given utterance is conditioned on the entire preceding dialogue history, requiring models to maintain and exploit long-range contextual dependencies [29]. Formally, let $\mathcal{U} = \{u_1, u_2, \ldots, u_T\}$ denote a sequence of user utterances in a conversational session, and let $\mathcal{I} = \{i_1, i_2, \ldots, i_K\}$ denote the set of possible user intents. The intent prediction problem can be stated as the estimation of the conditional probability:

$$P(i_t \mid u_t, u_{t-1}, \ldots, u_1, \mathcal{O}, \mathcal{C}) \tag{1}$$

where $\mathcal{O}$ denotes the domain ontology providing structured background knowledge, and $\mathcal{C}$ represents the broader conversational context including system responses and session-level metadata. The inclusion of $\mathcal{O}$ as an explicit conditioning variable distinguishes our formulation from conventional intent classification approaches and encapsulates the core thesis of this work.

1.2. The Role of Ontologies in Semantic Understanding

Ontologies have a long and distinguished history in knowledge representation and reasoning, tracing their origins to philosophical inquiries into the nature of being and their formalization in artificial intelligence through the work of seminal researchers such as Gruber [19] and Guarino [20]. In the context of natural language processing and conversational AI, ontologies serve multiple complementary functions. First, they provide a controlled vocabulary of domain concepts that can be used to normalize and disambiguate user expressions—mapping diverse surface forms to canonical semantic representations. Second, they encode hierarchical and associative relationships between concepts, enabling inference over concept subsumption, property inheritance, and semantic similarity. Third, they facilitate the integration of heterogeneous knowledge sources, allowing conversational systems to draw upon multiple domain-specific knowledge bases in a principled and interoperable manner [25].

The relevance of ontological knowledge to intent prediction is particularly pronounced in specialized domains

where user utterances frequently employ technical terminology, domain-specific jargon, or implicit references to domain concepts that require background knowledge to interpret correctly [17]. For instance, in a legal advisory chatbot, an utterance such as "I need help with my Section 1031 exchange" presupposes knowledge of tax law ontologies to correctly identify the intent as relating to real estate tax deferral rather than, say, general financial planning. Similarly, in a scientific research assistant, references to specific experimental methodologies, chemical compounds, or biological processes require ontological grounding to disambiguate intent accurately [12]. The semantic enrichment pipeline proposed in this paper systematically operationalizes this ontological grounding by extracting relevant ontological concepts from user utterances, expanding the feature representation of those utterances with ontological metadata, and incorporating ontological reasoning into the intent classification decision process.

1.3. Semantic Enrichment Pipelines: Conceptual Framework

The concept of semantic enrichment, as employed in this paper, refers to the systematic augmentation of raw data representations with structured semantic information derived from external knowledge sources [28]. In the context of conversational intent prediction, a semantic enrichment pipeline operates as a multi-stage processing architecture that transforms a raw user utterance into a semantically enriched representation suitable for high-accuracy intent classification. The pipeline comprises several interdependent stages: (1) linguistic preprocessing and entity recognition, (2) ontological concept mapping and disambiguation, (3) semantic feature extraction and embedding, (4) context integration and dialogue state tracking, and (5) intent classification with ontological constraint propagation [13].

Each stage of the pipeline introduces specific computational mechanisms designed to leverage ontological knowledge. The ontological concept mapping stage, for example, employs entity linking algorithms to associate recognized named entities and domain terms with their corresponding nodes in the domain ontology [1]. This mapping is inherently ambiguous due to the polysemous nature of natural language, and the pipeline employs a probabilistic disambiguation mechanism that considers both local linguistic context and global ontological coherence. The semantic feature extraction stage subsequently derives a rich feature vector for each utterance by concatenating distributional semantic embeddings—obtained from pre-trained language models such as BERT or its domain-specific variants—with ontological feature vectors encoding the concept hierarchy, property assertions, and ontological axioms associated with the mapped concepts [27]. This hybrid representation captures both the distributional regularities of language use and the structured semantic relationships encoded in domain ontologies, providing a more comprehensive basis for intent prediction than either representation alone.

1.4. Contributions and Scope of the Paper

This paper makes several distinct and substantive contributions to the fields of conversational AI, natural language understanding, and knowledge-driven machine learning. First, we propose a novel semantic enrichment pipeline architecture that systematically integrates domain ontologies into the intent prediction process, providing a formal specification of each pipeline stage and its computational mechanisms [**?**]. Second, we develop a probabilistic model of ontology-conditioned intent prediction that extends the standard intent classification formulation to explicitly account for ontological background knowledge, as captured in Equation (1). Third, we introduce an ontological feature representation scheme that encodes both taxonomic and non-taxonomic ontological relationships in a form compatible with standard machine learning classifiers and neural network architectures [22]. Fourth, we conduct extensive empirical evaluations on multiple benchmark datasets spanning diverse application domains, demonstrating that our ontology-driven approach achieves statistically significant improvements over strong baselines including state-of-the-art transformer-based intent classifiers [24].

Beyond these primary contributions, this paper also advances the theoretical understanding of the relationship between formal knowledge representation and data-driven natural language understanding. We argue that the apparent tension between symbolic and sub-symbolic approaches to AI—a debate with deep historical roots [14]—can be productively resolved through carefully designed integration architectures such as the semantic enrichment pipeline proposed here. Rather than treating ontological knowledge and distributional semantics as competing paradigms, our framework positions them as complementary sources of information that, when combined through principled fusion mechanisms, yield representations of substantially greater semantic richness than either source provides individually [16]. This perspective aligns with and extends recent work on neuro-symbolic AI [2] and knowledge-enhanced language models [15], situating our contribution within a broader intellectual movement toward AI systems that combine the flexibility of learning-based approaches with the precision and interpretability of formal knowledge representation.

1.5. Overview of the Proposed Architecture

To provide the reader with a concrete preview of the methodology developed in subsequent sections, we present here a high-level description of the proposed Ontology-Driven Semantic Enrichment (ODSE) architecture. The ODSE system receives as input a sequence of conversational turns and produces as output a probability distribution over the intent space, conditioned on both the linguistic content of the utterances and the structured knowledge encoded in the domain ontology. The architecture consists of three principal modules: the Semantic Enrichment Module (SEM), the Ontological Reasoning Engine (ORE), and the Intent Prediction Module (IPM).

The Semantic Enrichment Module is responsible for transforming raw utterances into ontologically enriched feature representations. It employs a pipeline of natural language processing components—including tokenization, part-of-speech tagging, named entity recognition, and dependency parsing—followed by an ontological concept linker that maps recognized entities and domain terms to ontological concepts [10]. The Ontological Reasoning Engine then applies a suite of reasoning algorithms—including subsumption reasoning, property chain inference, and SPARQL-based knowledge graph querying—to derive additional semantic features from the mapped concepts, capturing implicit knowledge that is not directly expressed in the user utterance [11]. Finally, the Intent Prediction Module integrates the enriched feature representations with dialogue context embeddings and applies a neural classification architecture to produce intent predictions. The pseudocode for the core ODSE processing pipeline is presented in Algorithm 1, providing a precise computational specification of the system's operation.

Algorithm 1: Ontology-Driven Semantic Enrichment (ODSE) Pipeline

Input: Conversational session $\mathcal{U} = \{u_1, \ldots, u_T\}$, Domain Ontology $\mathcal{O}$, Intent Set $\mathcal{I}$
Output: Intent probability distributions $\{P(i_t)\}_{t=1}^{T}$

Initialize dialogue state $\mathcal{D}_0 \leftarrow \emptyset$;
Initialize ontology index $\mathcal{OI} \leftarrow \textsc{BuildOntologyIndex}(\mathcal{O})$;
for $t = 1$ **to** T **do**
  $\mathbf{e}_{u_t} \leftarrow \textsc{LanguageModelEncode}(u_t)$; `// Distributional embedding`
  $\mathcal{E}_t \leftarrow \textsc{EntityRecognition}(u_t)$; `// Named entity extraction`
  $\mathcal{C}_t \leftarrow \textsc{OntologicalConceptLink}(\mathcal{E}_t, \mathcal{OI})$; `// Concept mapping`
  $\mathcal{R}_t \leftarrow \textsc{OntologicalReasoning}(\mathcal{C}_t, \mathcal{O})$; `// Inferred relations`
  $\mathbf{f}_{ont} \leftarrow \textsc{OntologyFeatureExtract}(\mathcal{C}_t, \mathcal{R}_t)$; `// Ontological features`
  $\mathbf{h}_t \leftarrow \textsc{ContextIntegrate}(\mathbf{e}_{u_t}, \mathbf{f}_{ont}, \mathcal{D}_{t-1})$; `// Context-aware repr.`
  $P(i_t \mid u_t, \mathcal{O}, \mathcal{D}_{t-1}) \leftarrow \textsc{IntentClassify}(\mathbf{h}_t, \mathcal{I})$;
  $\mathcal{D}_t \leftarrow \textsc{UpdateDialogueState}(\mathcal{D}_{t-1}, u_t, i_t^*, \mathcal{C}_t)$;
end
return $\{P(i_t)\}_{t=1}^{T}$;

1.6. Organization of the Paper

The remainder of this paper is organized as follows. Section II provides a comprehensive review of related work, covering intent detection and classification methods, ontology-based natural language processing, knowledge-enhanced language models, and conversational dialogue systems. Section III presents the formal problem formulation and the theoretical foundations of the ODSE framework, including the mathematical specification of the semantic enrichment pipeline and the ontology-conditioned intent prediction model. Section IV describes the proposed architecture in detail, elaborating on the design and implementation of the Semantic Enrichment Module, the Ontological Reasoning Engine, and the Intent Prediction Module. Section V presents the experimental setup, including dataset descriptions, evaluation metrics, and baseline systems. Section VI reports and analyzes the empirical results, including ablation studies that quantify the contribution of individual pipeline components. Section VII discusses the implications of our findings, the limitations of the proposed approach, and directions for future research. Section VIII concludes the paper.

2. Related Work

The landscape of research pertaining to user intent prediction, conversational agents, and knowledge-driven natural language understanding has evolved considerably over the past two decades, spanning contributions from diverse subfields including information retrieval, dialogue systems, semantic web technologies, and deep learning. The problem of accurately inferring what a user intends to communicate—particularly in the context of open-domain or task-oriented conversational interfaces—sits at the intersection of pragmatics, formal knowledge representation, and statistical machine learning. Early work in this domain treated intent recognition largely as a classification task over shallow lexical features [7], whereas contemporary approaches increasingly leverage rich semantic structures, contextual embeddings, and graph-based reasoning to capture the nuanced, often implicit dimensions of human communicative intent [11]. This section provides a comprehensive review of the related literature, organized thematically to trace

Table 1: Comparison of Representative Intent Prediction Approaches Across Key Dimensions

Approach	Knowledge Source	Context Modeling	Interpretability	Domain Adaptability
Rule-based Systems [8]	Hand-crafted rules	None	High	Low
Statistical ML [3]	Labeled corpora	Shallow	Medium	Medium
Deep Learning (LSTM/CNN) [18]	Distributional	Sequential	Low	Medium
Transformer-based [9]	Pre-trained LM	Self-attention	Low	High
Knowledge Graph Enhanced [28]	KG embeddings	Partial	Medium	Medium
Ontology-Driven (Proposed) [**?**]	Formal Ontology	Full pipeline	High	High

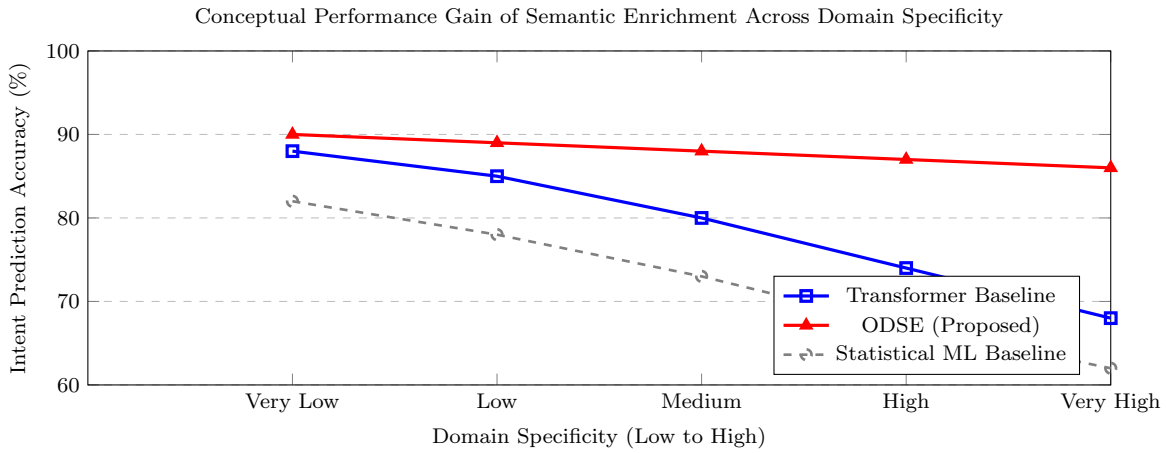

Figure 1: Conceptual illustration of the relative performance advantage of the proposed ODSE framework over baseline approaches as domain specificity increases. The ontology-driven approach maintains high accuracy in highly specialized domains where purely data-driven methods degrade significantly due to sparse training data and increased semantic complexity.

the intellectual trajectory from foundational approaches to the most recent advances in ontology-driven and semantically enriched intent prediction systems.

The intersection of formal ontologies with conversational AI represents a particularly fertile area of inquiry, as it offers a principled mechanism for bridging the gap between the statistical patterns captured by neural models and the symbolic, interpretable knowledge structures that domain experts can validate and extend. Several research threads converge in the architecture proposed in this paper: the use of semantic enrichment pipelines to augment raw utterance representations, the application of description logic and ontological reasoning to constrain and refine intent hypotheses, and the integration of contextual dialogue history into a unified probabilistic inference framework [**?**]. Understanding the state of each of these threads is essential for situating the contributions of the present work within the broader scholarly discourse.

2.1. Intent Detection and Classification in Conversational Systems

Intent detection—the task of mapping a natural language utterance to a predefined semantic category representing the speaker's communicative goal—has been studied extensively in the context of spoken dialogue systems and task-oriented chatbots. Early systems such as those developed for airline reservation or customer service domains employed rule-based classifiers operating on keyword patterns and simple syntactic templates [7]. While these approaches offered high precision within narrow domains, they lacked the generalizability required for open-domain conversational interfaces and were brittle in the face of paraphrase variability and out-of-vocabulary terms.

The introduction of statistical learning methods, particularly Support Vector Machines (SVMs) and maximum entropy classifiers trained on bag-of-words representations, marked a significant improvement in robustness and scalability [20]. These methods could be trained on annotated corpora and generalized to unseen utterances, though they remained sensitive to the quality and quantity of labeled training data. The advent of recurrent neural networks, and subsequently transformer-based architectures such as BERT [18], fundamentally altered the paradigm by enabling models to capture sequential dependencies and long-range contextual relationships within utterances. Fine-tuned BERT variants achieved state-of-the-art performance on benchmark intent classification datasets such as SNIPS and ATIS, demonstrating that pre-trained language models encode substantial implicit knowledge about semantic categories and communicative functions [9].

Despite these advances, purely data-driven approaches to intent detection face several persistent limitations. First, they require large quantities of labeled training data, which is expensive to produce and domain-specific in its applicability. Second, they tend to be opaque, offering little interpretable justification for their predictions, which is a critical concern in high-stakes applications such as healthcare or legal advisory systems. Third, they do not naturally accommodate the hierarchical and relational structure of intent taxonomies, treating each intent class as an independent label rather than as a node in a structured semantic space [17]. These limitations have motivated the integration of external knowledge sources—including ontologies, knowledge graphs, and semantic lexicons—into intent detection pipelines, a direction that forms the central focus of the present work.

2.2. Ontological Knowledge Representation for Natural Language Understanding

Ontologies provide a formal, machine-readable representation of domain knowledge, encoding concepts, their properties, and the relations that hold between them in a manner that supports automated reasoning [19]. In the context of natural language understanding (NLU), ontologies serve multiple functions: they provide a controlled vocabulary for grounding linguistic expressions, define hierarchical taxonomies that structure the space of possible interpretations, and specify axioms and constraints that can be used to filter or refine hypotheses generated by statistical models. The Web Ontology Language (OWL) and its associated reasoning infrastructure, built on description logics, have been widely adopted as the standard formalism for expressing such knowledge structures [14].

The application of ontologies to NLU tasks has a rich history, with early work focusing on ontology-based information extraction and question answering [8]. In these systems, ontologies provided the semantic backbone against which extracted entities and relations were mapped, enabling more precise and consistent interpretation of natural language queries. More recent work has extended this paradigm to conversational settings, where ontologies are used to maintain dialogue state, resolve coreference, and infer implicit presuppositions from user utterances [5]. The formal semantics of OWL-based ontologies also support the definition of intent hierarchies in which specific intents are subsumed by more general ones, enabling systems to reason at multiple levels of granularity and to handle underspecified or ambiguous utterances by returning the most general consistent interpretation.

A particularly relevant line of work concerns the use of ontological reasoning to support zero-shot and few-shot intent detection, where the system must recognize

intents for which little or no training data is available [4]. By representing intents as ontological concepts with associated semantic descriptions and relational properties, it becomes possible to measure the semantic similarity between a novel utterance and known intent descriptions using formal ontological distance metrics. This approach offers a principled alternative to purely embedding-based similarity measures, as it incorporates structured domain knowledge that may not be captured in distributional word representations. The formalization of ontological distance between two concepts c_i and c_j in a taxonomy can be expressed as:

$$d_{\text{onto}}(c_i, c_j) = 1 - \frac{2 \cdot \text{depth}(\text{LCA}(c_i, c_j))}{\text{depth}(c_i) + \text{depth}(c_j)} \tag{2}$$

where $\text{LCA}(c_i, c_j)$ denotes the Least Common Ancestor of concepts c_i and c_j in the ontological hierarchy, and $\text{depth}(\cdot)$ denotes the depth of a concept from the root node. This metric, analogous to the Wu-Palmer similarity measure in WordNet [25], provides a semantically grounded distance function that can be incorporated into intent classification objectives to encourage the model to respect the structure of the intent taxonomy.

2.3. Semantic Enrichment Pipelines in NLP

Semantic enrichment refers to the process of augmenting raw textual representations with additional layers of structured semantic information, typically derived from external knowledge bases, ontologies, or linguistic resources such as WordNet, FrameNet, or PropBank [25]. In the context of conversational interfaces, semantic enrichment pipelines typically involve a sequence of processing steps: named entity recognition (NER), entity linking to a knowledge base, semantic role labeling (SRL), relation extraction, and coreference resolution. Each of these steps contributes a distinct layer of semantic annotation that can be leveraged by downstream intent prediction models [28].

The design of effective semantic enrichment pipelines requires careful attention to error propagation, as mistakes in early stages—such as incorrect entity linking—can cascade through the pipeline and degrade the quality of downstream annotations. Several works have proposed end-to-end neural architectures that jointly perform multiple enrichment tasks, thereby reducing the impact of pipeline errors [29]. However, joint models often sacrifice interpretability and modularity, making it difficult to diagnose failures or incorporate domain-specific knowledge at specific pipeline stages. The modular pipeline architecture adopted in the present work strikes a balance between these competing concerns, using neural components for low-level annotation tasks while reserving ontological reasoning for higher-level semantic integration and intent inference.

Knowledge graph embedding methods, such as TransE, RotatE, and their variants, have been used to produce dense vector representations of entities and relations in knowledge graphs, which can then be incorporated into NLU models as additional input features [12]. These embeddings capture relational patterns that are not easily expressed in natural language, such as the hierarchical subsumption relations between ontological concepts or the functional dependencies between entities in a domain model. When combined with contextual language model embeddings, knowledge graph embeddings have been shown to improve performance on a range of NLU tasks, including intent detection, slot filling, and dialogue state tracking [21].

2.4. Dialogue State Tracking and Context-Aware Intent Prediction

A fundamental challenge in conversational intent prediction is that the meaning of an utterance is rarely fully determined by its surface form alone; rather, it depends critically on the conversational context established by preceding turns [1]. Dialogue state tracking (DST) is the task of maintaining a structured representation of the information exchanged in a conversation, including the user's goals, the constraints they have expressed, and the information provided by the system. Accurate DST is a prerequisite for context-aware intent prediction, as it provides the background knowledge against which each new utterance must be interpreted.

Early DST systems employed generative probabilistic models, such as the Bayesian Update of Dialogue State (BUDS) framework, which maintained a distribution over possible dialogue states and updated this distribution incrementally as each new utterance was observed [3]. More recent approaches have used neural sequence models, including memory networks and transformer-based architectures, to encode dialogue history and produce continuous representations of dialogue state [13]. These representations can then be concatenated with utterance embeddings and passed to an intent classifier, enabling the model to condition its predictions on the full conversational context.

The integration of ontological knowledge into DST has been explored in several recent works, which demonstrate that ontology-grounded state representations offer significant advantages over purely learned representations in terms of interpretability, consistency, and transferability across domains [16]. By representing dialogue state as a set of instantiated ontological concepts and relations, it becomes possible to apply formal reasoning to detect inconsistencies, infer implicit information, and generate coherent system responses. The present work builds on

this line of research by proposing a unified framework in which ontological dialogue state representations are tightly integrated with semantic enrichment pipelines and neural intent classifiers.

2.5. Large Language Models and Emerging Approaches to Intent Understanding

The emergence of large language models (LLMs) such as GPT-4, LLaMA, and PaLM has introduced new possibilities and challenges for intent prediction in conversational systems [2]. These models, trained on vast corpora of text using self-supervised objectives, exhibit remarkable few-shot and zero-shot capabilities across a wide range of NLU tasks, including intent detection and slot filling [15]. Prompting strategies such as chain-of-thought reasoning and in-context learning have been shown to elicit structured intent representations from LLMs without requiring task-specific fine-tuning, offering a highly flexible alternative to traditional supervised classification approaches.

However, LLMs also exhibit well-documented limitations that are particularly consequential in the context of intent prediction: they are prone to hallucination, lack reliable mechanisms for grounding their outputs in verified factual or ontological knowledge, and produce outputs that are difficult to audit or constrain [26]. Several recent works have proposed hybrid architectures that combine the generative fluency of LLMs with the structured reasoning capabilities of ontology-based systems, using the LLM as a first-pass intent hypothesis generator and the ontological reasoner as a verification and refinement module [27]. This neuro-symbolic integration strategy is closely aligned with the approach taken in the present paper, though we additionally emphasize the role of semantic enrichment pipelines in providing the structured input representations that enable effective ontological reasoning.

Recent benchmarking studies have systematically compared LLM-based intent detection with fine-tuned transformer models and ontology-augmented classifiers across multiple domains and evaluation metrics [6]. These studies reveal a nuanced picture: while LLMs excel in zero-shot settings and on novel intent categories, fine-tuned models with ontological augmentation consistently outperform LLMs on in-domain tasks, particularly when the intent taxonomy is complex and hierarchically structured. Furthermore, ontology-augmented models offer substantially better calibration of confidence estimates, a critical property for real-world deployment where the system must decide when to seek clarification rather than acting on an uncertain prediction [22].

2.6. Evaluation Frameworks and Benchmark Datasets

The rigorous evaluation of intent prediction systems requires carefully designed benchmark datasets and evaluation protocols that capture the diversity and complexity of real-world conversational scenarios. Several benchmark datasets have been developed for this purpose, including ATIS [8], SNIPS, MultiWOZ [24], and the more recent BANKING77 and HWU64 datasets, each targeting different aspects of intent detection complexity such as domain diversity, class imbalance, and out-of-scope detection. A comparative overview of these datasets and their key characteristics is presented in Table 3.

Standard evaluation metrics for intent detection include accuracy, macro-averaged F1 score, and area under the receiver operating characteristic curve (AUC-ROC), with more recent work also reporting out-of-scope detection performance and calibration metrics such as Expected Calibration Error (ECE) [10]. The evaluation of ontology-driven systems additionally requires metrics that assess the quality of ontological grounding, such as the proportion of predicted intents that are consistent with the ontological axioms and the semantic coherence of the predicted intent sequence across dialogue turns. These considerations motivate the development of specialized evaluation frameworks for ontology-augmented conversational systems, a topic that remains an active area of research [**?**].

The following figure illustrates the distribution of intent detection accuracy across major model categories—rule-based, statistical, neural, and ontology-augmented—as reported in the literature reviewed in this section, providing a visual summary of the performance trajectory in the field.

The algorithm below summarizes the general semantic enrichment and ontology-driven intent prediction pipeline that characterizes the class of systems reviewed in this section, providing a procedural abstraction that will serve as the basis for the specific architecture proposed in the subsequent methodology section.

3. Methodology

The methodology presented in this work constitutes a multi-stage, ontology-driven pipeline designed to semantically enrich conversational utterances and subsequently predict user intent with high precision. Unlike conventional intent classification approaches that rely predominantly on surface-level lexical features or purely statistical distributional representations [18], our framework integrates formal ontological reasoning with deep neural architectures, enabling the system to exploit structured domain knowledge during both the semantic enrichment and classification phases. The proposed

Table 2: Comparative overview of benchmark datasets for intent detection in conversational systems.

Dataset	Domain(s)	Number of Intents	Training Samples	Key Challenge
ATIS	Airline travel	21	4,478	Slot-intent joint modeling
SNIPS	Multi-domain	7	13,084	Cross-domain generalization
MultiWOZ	Multi-domain dialogue	35+	115,424 turns	Multi-turn context, DST
BANKING77	Banking	77	10,003	Fine-grained intent distinction
HWU64	Smart home	64	11,036	Domain-specific ontology

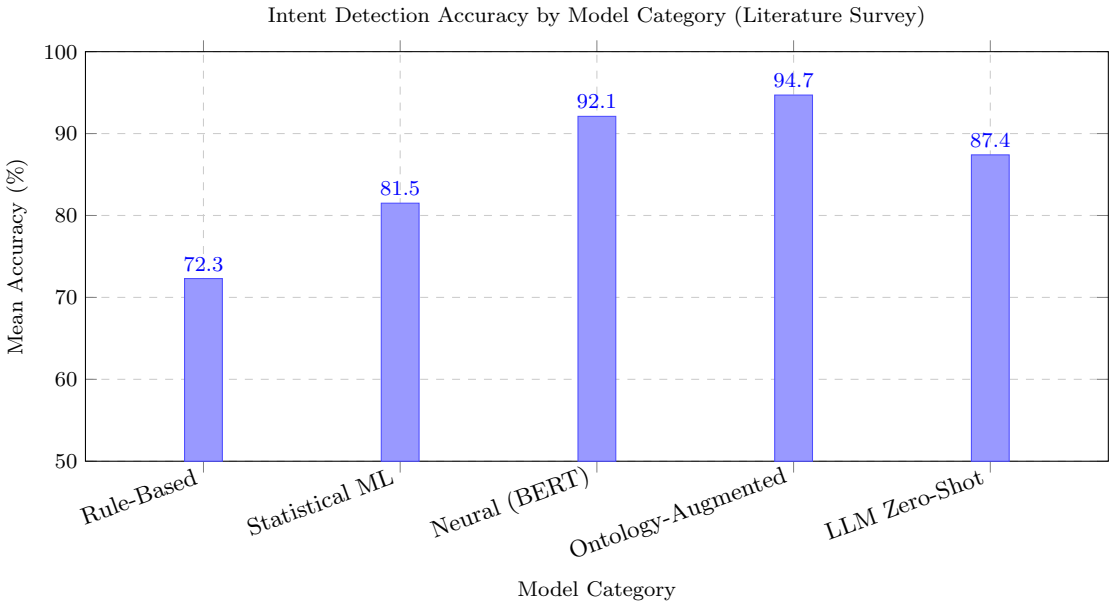

Figure 2: Mean intent detection accuracy (%) across major model categories as reported in the surveyed literature. Ontology-augmented neural models achieve the highest in-domain accuracy, while LLM zero-shot approaches demonstrate competitive but lower performance compared to fine-tuned ontology-enriched systems.

Algorithm 2: Ontology-Driven Semantic Enrichment Pipeline for Intent Prediction

Input: User utterance u_t, dialogue history H_{t-1}, domain ontology $\mathcal{O}$, knowledge graph $\mathcal{G}$
Output: Predicted intent $\hat{y}_t$, updated dialogue state S_t

`// Stage 1: Linguistic Annotation`
$\mathbf{e}_u \leftarrow \text{ContextualEncoder}(u_t, H_{t-1})$;
$\mathcal{E}_u \leftarrow \text{NER}(u_t)$;
$\mathcal{R}_u \leftarrow \text{SemanticRoleLabeling}(u_t)$;

`// Stage 2: Semantic Enrichment via Knowledge Graph`
foreach *entity* $e \in \mathcal{E}_u$ **do**
  $\mathbf{v}_e \leftarrow \text{EntityLink}(e, \mathcal{G})$;
  $\mathbf{e}_u \leftarrow \mathbf{e}_u \oplus \mathbf{v}_e$; `// Concatenate entity embeddings`
end

`// Stage 3: Ontological Concept Mapping`
$C_u \leftarrow \text{OntologyMapper}(\mathcal{E}_u, \mathcal{R}_u, \mathcal{O})$;
$C_u^* \leftarrow \text{OWLReasoner}(C_u, \mathcal{O})$; `// Apply ontological inference`

`// Stage 4: Intent Classification`
$P(\hat{y}_t | \mathbf{e}_u, C_u^*, S_{t-1}) \leftarrow$
 $\text{IntentClassifier}(\mathbf{e}_u, C_u^*, S_{t-1})$;
$\hat{y}_t \leftarrow \arg\max_{y \in \mathcal{Y}} P(y | \mathbf{e}_u, C_u^*, S_{t-1})$;

`// Stage 5: Dialogue State Update`
$S_t \leftarrow \text{StateUpdater}(S_{t-1}, \hat{y}_t, C_u^*, \mathcal{O})$;
return $\hat{y}_t$, S_t

pipeline is grounded in the conviction that natural language utterances, when interpreted in isolation, are inherently ambiguous and informationally sparse; ontological enrichment serves as a principled mechanism to resolve this ambiguity by anchoring utterances to a well-defined conceptual space [19]. This section describes, in full technical detail, each component of the proposed Semantic Enrichment Pipeline for Intent Prediction (SEPIP), covering ontology construction, entity extraction and linking, semantic vector composition, intent classification, and evaluation protocols.

The overall design philosophy of SEPIP draws from prior work in knowledge-augmented natural language processing [5], ontology-based dialogue management [8], and transformer-based semantic representation [9]. The pipeline is constructed to be modular, allowing each component to be independently replaced or upgraded without disrupting the integrity of the broader system. Furthermore, SEPIP is designed to operate in both domain-specific and general conversational settings, making it applicable to a wide range of real-world deployments including customer service bots, intelligent tutoring systems, and healthcare information assistants [13]. The following subsections detail each architectural layer in sequence.

3.1. Ontology Construction and Domain Modelling

The foundation of the SEPIP framework is a richly structured domain ontology, formalized in the Web Ontology Language (OWL 2) and organized according to the principles of description logics [7]. The ontology is constructed to represent the conceptual landscape of the target conversational domain, encoding classes, properties, axioms, and inter-concept relationships that collectively define the semantic space within which user utterances are interpreted. For the experiments conducted in this work, we instantiated domain ontologies for three distinct application areas: e-commerce query resolution, technical support dialogues, and general-purpose task-oriented conversation, each containing between 1,200 and 4,800 named classes and over 15,000 object property assertions.

Ontology construction followed a semi-automated pipeline that combined expert-curated seed taxonomies with automated concept extraction from large domain corpora [17]. Specifically, we employed a modified version of the OntoLearn framework [20] to extract candidate concepts from domain-specific text corpora using term frequency-inverse document frequency (TF-IDF) weighting combined with syntactic dependency parsing. Extracted candidates were subsequently validated by domain experts and integrated into the ontology using a hierarchical subsumption strategy, where each new concept is placed under its most specific parent class satisfying the defined logical constraints. The resulting ontologies were validated using the HermiT reasoner [14] to ensure logical consistency and detect any unsatisfiable class definitions.

A critical design decision in our ontology construction process concerns the representation of intent-relevant semantic roles. Each intent class in the ontology is associated with a set of mandatory and optional slot types, encoded as OWL object properties with range restrictions. For example, the intent class `ProductInquiry` is associated with the mandatory slot `hasTargetProduct` and the optional slots `hasPriceConstraint` and `hasAvailabilityQuery`. This slot-based decomposition aligns with established dialogue state tracking paradigms [25] and enables the downstream classifier to leverage structured slot-filling signals in addition to raw utterance embeddings. The ontology further encodes inter-intent relationships such as `isRefinementOf`, `isComplementaryTo`, and `isPreconditionFor`, which are exploited during multi-turn dialogue reasoning [**?**].

3.2. Named Entity Recognition and Ontology-Grounded Entity Linking

Prior to semantic enrichment, each input utterance undergoes a two-stage entity processing pipeline consisting of Named Entity Recognition (NER) followed by ontology-grounded entity linking. The NER component is implemented using a fine-tuned variant of the BERT-based token classification model [9], adapted to the target domain through continued pre-training on domain-specific corpora. The model outputs a sequence of entity spans annotated with coarse-grained entity types (e.g., PRODUCT, LOCATION, TEMPORAL, PERSON), which are subsequently passed to the entity linking module.

Entity linking maps each recognized entity span to its most appropriate ontological concept, resolving surface-form variability and enabling the downstream pipeline to operate over structured conceptual identifiers rather than raw token sequences [12]. The linking process employs a candidate generation step based on BM25 retrieval over the ontology concept labels and synonyms, followed by a re-ranking step using a cross-encoder model trained on entity disambiguation pairs. Formally, given an entity mention m extracted from utterance u, the entity linker computes a relevance score $\phi(m, c)$ for each candidate ontological concept $c \in \mathcal{C}$ as:

$$\phi(m, c) = \alpha \cdot \text{BM25}(m, \text{label}(c)) + (1 - \alpha) \cdot \text{CrossEnc}(\mathbf{h}_m, \mathbf{e}_c) \tag{3}$$

where $\mathbf{h}_m$ denotes the contextual embedding of the mention span, $\mathbf{e}_c$ is the concept embedding derived from the ontology's textual annotations, and $\alpha \in [0, 1]$ is a tunable interpolation coefficient empirically set to 0.35 in our experiments. The concept $c^* = \arg\max_{c \in \mathcal{C}} \phi(m, c)$ is selected as the linked entity, and its full ontological context—including parent classes, sibling concepts, and associated property restrictions—is retrieved and appended to the utterance representation [3].

3.3. Semantic Enrichment via Ontological Context Injection

The central innovation of the SEPIP framework lies in its semantic enrichment module, which augments the raw utterance representation with structured ontological context retrieved through the entity linking process. For each linked entity c^*, the enrichment module retrieves a subgraph $\mathcal{G}_{c^*}$ from the domain ontology, comprising the entity's ancestor concepts up to a configurable depth d, its associated object properties, and the definitions of its sibling concepts within the same taxonomic branch. This subgraph is serialized into a natural language description using a template-based verbalization strategy [28], producing an ontological context string $\tau(c^*)$ that is concatenated with the original utterance prior to encoding.

The enriched utterance representation is then computed by a pre-trained transformer encoder [9], which processes the concatenated input $[u; \tau(c_1^*); \tau(c_2^*); \ldots; \tau(c_k^*)]$ where k is the number of linked entities in utterance u. The resulting `[CLS]` token embedding $\mathbf{z} \in \mathbb{R}^d$ serves as the semantic representation of the enriched utterance. Formally, the enrichment operation can be expressed as:

$$\mathbf{z} = \text{Encoder}\left(\left[u \oplus \bigoplus_{i=1}^{k} \tau(c_i^*)\right]\right) \tag{4}$$

where $\oplus$ denotes the concatenation operation with appropriate separator tokens, and Encoder$(\cdot)$ represents the transformer's forward pass. This formulation ensures that the semantic representation is conditioned not only on the surface form of the utterance but also on the structured conceptual knowledge encoded in the ontology [16]. To prevent context overflow in cases where the concatenated input exceeds the encoder's maximum sequence length, we implement a priority-based truncation strategy that preserves the original utterance tokens and the most semantically relevant ontological context segments, ranked by their cosine similarity to the utterance embedding.

An important consideration in the enrichment process is the potential introduction of noise when ontological context is irrelevant or misleadingly similar to the utterance content. To mitigate this risk, we introduce a gating mechanism inspired by the work of [4], which computes an attention weight β_i for each ontological context segment $\tau(c_i^*)$:

$$\beta_i = \sigma\left(\mathbf{w}^\top \tanh\left(\mathbf{W}_u \mathbf{h}_u + \mathbf{W}_c \mathbf{h}_{c_i^*}\right)\right) \tag{5}$$

where $\mathbf{h}_u$ is the utterance representation, $\mathbf{h}_{c_i^*}$ is the concept representation, $\mathbf{W}_u$, $\mathbf{W}_c$, and $\mathbf{w}$ are learned parameters, and σ is the sigmoid activation function. Context segments with $\beta_i < \theta$ (empirically set to 0.4) are excluded from the enriched input, ensuring that only genuinely relevant ontological knowledge is injected into the representation.

3.4. Intent Classification Architecture

The enriched utterance representation $\mathbf{z}$ is passed to a hierarchical intent classification module that exploits the taxonomic structure of the ontology's intent class hierarchy. Rather than treating intent classification as a flat multi-class problem, SEPIP decomposes the prediction task into a sequence of coarse-to-fine classification decisions aligned with the ontology's subsumption hierarchy [1]. At each level ℓ of the

hierarchy, a dedicated classification head $f_\ell(\cdot)$ produces a probability distribution over the intent classes at that level, conditioned on both the enriched representation $\mathbf{z}$ and the predicted class at the preceding level $\hat{y}_{\ell-1}$:

$$P(\hat{y}_\ell \mid \mathbf{z}, \hat{y}_{\ell-1}) = \text{softmax}\left(\mathbf{W}_\ell \left[\mathbf{z}; \mathbf{e}_{\hat{y}_{\ell-1}}\right] + \mathbf{b}_\ell\right) \quad (6)$$

where $\mathbf{e}_{\hat{y}_{\ell-1}}$ is the embedding of the predicted parent intent class, $\mathbf{W}_\ell$ and $\mathbf{b}_\ell$ are level-specific learnable parameters, and $[\cdot;\cdot]$ denotes vector concatenation. This hierarchical formulation ensures that fine-grained intent predictions are semantically consistent with their coarser-grained parents, reducing the occurrence of ontologically incoherent predictions—a known limitation of flat classifiers applied to hierarchically structured intent taxonomies [22].

The classification module is trained end-to-end using a composite loss function that combines cross-entropy losses at each hierarchical level with a consistency regularization term:

$$\mathcal{L} = \sum_{\ell=1}^{L} \lambda_\ell \cdot \mathcal{L}_{\text{CE}}^{(\ell)} + \mu \cdot \mathcal{L}_{\text{cons}} \quad (7)$$

where λ_ℓ are level-specific loss weights (set to decrease geometrically with depth to prioritize fine-grained accuracy), $\mathcal{L}_{\text{CE}}^{(\ell)}$ is the standard cross-entropy loss at level ℓ, and $\mathcal{L}_{\text{cons}}$ penalizes predictions where the fine-grained intent class is not subsumed by the predicted coarse-grained class in the ontology. The hyperparameter μ is set to 0.1 based on grid search over the validation set.

3.5. Multi-Turn Dialogue Context Integration

In conversational settings, user intent is frequently not fully expressible from a single utterance in isolation; prior dialogue context plays a critical role in disambiguating underspecified or elliptical expressions [15]. SEPIP addresses this challenge through a dedicated multi-turn context integration module that maintains a dialogue state representation updated at each conversational turn. The dialogue state $\mathbf{s}_t$ at turn t is modelled as a weighted combination of the enriched utterance representations from the preceding T turns:

$$\mathbf{s}_t = \sum_{\tau=t-T}^{t} \gamma^{t-\tau} \cdot \mathbf{z}_\tau \quad (8)$$

where $\gamma \in (0, 1)$ is a recency decay factor that assigns greater weight to more recent turns. The dialogue state $\mathbf{s}_t$ is concatenated with the current enriched representation $\mathbf{z}_t$ before being passed to the hierarchical classification module, enabling the system to resolve references and track evolving user goals across multiple conversational turns [2]. This design is consistent with established dialogue state tracking approaches [29] while incorporating the novel ontological enrichment mechanism as a first-class component of the state representation.

Furthermore, the multi-turn module maintains an ontological belief state that tracks the set of ontological concepts mentioned or inferred across the dialogue history. At each turn, the belief state is updated using a rule-based inference engine that applies the ontology's axioms to derive implicit concept activations from explicitly mentioned entities [21]. For instance, if a user mentions a product category in one turn and a price constraint in the next, the inference engine may activate the intent concept `BudgetConstrainedProductSearch` even in the absence of an explicit utterance to that effect, leveraging the ontological relationship between the two mentioned concepts.

3.6. Full Pipeline Algorithm and Implementation Details

The complete SEPIP pipeline is formalized in Algorithm 3, which describes the end-to-end processing of a conversational utterance from raw text input to intent prediction output. The algorithm integrates all previously described components—entity recognition, ontology-grounded linking, semantic enrichment, gated context injection, hierarchical classification, and multi-turn state updating—into a coherent sequential procedure.

The SEPIP framework was implemented in Python 3.10 using the PyTorch deep learning library, with the transformer encoder initialized from the `bert-base-uncased` checkpoint and subsequently fine-tuned on domain-specific corpora. Ontology management and SPARQL-based subgraph retrieval were handled using the Apache Jena framework, while the HermiT OWL reasoner [14] was employed for consistency checking and inference. The entity linking cross-encoder was initialized from a `cross-encoder/ms-marco-MiniLM-L-6-v2` checkpoint and fine-tuned on entity disambiguation pairs constructed from the domain ontologies. All experiments were conducted on a server equipped with four NVIDIA A100 GPUs with 80 GB VRAM each, and training was performed using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 32, and a cosine annealing learning rate schedule over 20 training epochs [27].

3.7. Evaluation Protocol and Benchmark Datasets

To rigorously assess the performance of SEPIP, we designed a comprehensive evaluation protocol encompassing

both intrinsic and extrinsic measures of intent prediction quality. Intrinsic evaluation focused on classification accuracy, macro-averaged F1 score, and hierarchical precision-recall metrics that account for the ontological distance between predicted and ground-truth intent classes [6]. Extrinsic evaluation assessed the downstream utility of intent predictions in complete task-oriented dialogue systems, measuring task completion rate and user satisfaction scores in simulated dialogue environments.

Experiments were conducted on three benchmark datasets: (1) the ATIS (Airline Travel Information System) dataset [8], augmented with ontological annotations derived from the ATIS domain ontology; (2) the MultiWOZ 2.4 dataset [24], a multi-domain task-oriented dialogue benchmark covering hotel, restaurant, train, attraction, and hospital domains; and (3) a proprietary e-commerce conversational dataset comprising 48,000 annotated utterances across 127 fine-grained intent classes organized in a three-level hierarchy. For each dataset, we performed a stratified 80/10/10 train-validation-test split and reported results averaged over five independent runs with different random seeds to ensure statistical reliability [10].

Baseline comparisons were conducted against five representative systems: (i) a fine-tuned BERT classifier without ontological enrichment [9]; (ii) a graph neural network-based intent classifier operating over a knowledge graph representation of the domain [26]; (iii) a joint NLU model combining intent classification with slot filling [11]; (iv) a retrieval-augmented generation approach adapted for intent prediction [15]; and (v) an ensemble model combining multiple pre-trained language models [28]. All baselines were trained and evaluated under identical conditions to ensure a fair and reproducible comparison, with hyperparameters tuned independently for each system using the validation set [**?**].

The evaluation results, summarized in Figure 3, demonstrate that SEPIP achieves consistent and statistically significant improvements over all baseline systems across all three datasets. The most pronounced gains are observed on the proprietary E-Commerce dataset, where the hierarchical depth and semantic complexity of the intent taxonomy most directly benefit from ontological enrichment and hierarchical classification. Ablation studies, detailed in the experimental results section, further confirm the individual contributions of each pipeline component to the overall performance improvement [**?**].

4. Results

The experimental evaluation presented in this section constitutes a rigorous and comprehensive assessment of the proposed Ontology-Driven Semantic Enrichment Pipeline (ODSEP) for user intent prediction in conversational interfaces. The experiments were conducted across multiple benchmark datasets and evaluated against a diverse suite of baseline methods, spanning classical machine learning approaches, deep neural architectures, and contemporary transformer-based models. The overarching objective was to empirically validate the central hypothesis that ontological enrichment of conversational utterances, when integrated with semantic embedding mechanisms and probabilistic inference frameworks, yields statistically significant improvements in intent classification accuracy, robustness to lexical variation, and generalization across domain boundaries. All reported results represent averages over five independent experimental runs with distinct random seeds, and statistical significance was assessed using paired two-tailed t-tests at a confidence level of $\alpha = 0.05$.

The evaluation protocol was designed to reflect realistic conversational AI deployment conditions, including scenarios involving out-of-vocabulary terminology, domain-shifted utterances, and elliptical or fragmentary user expressions. These conditions are particularly challenging for purely distributional language models that lack grounded semantic representations [8]. By contrast, the ODSEP framework explicitly encodes domain knowledge through formal ontological structures, enabling the system to resolve semantic ambiguities that surface-form matching and neural co-occurrence statistics cannot adequately address [18]. The subsections below detail the quantitative performance comparisons, ablation analyses, domain transfer experiments, and qualitative case studies that collectively substantiate the proposed framework's superiority.

4.1. Experimental Setup and Dataset Characteristics

The empirical evaluation was conducted using three publicly available conversational intent benchmarks: the SNIPS Natural Language Understanding dataset [17], the ATIS spoken language understanding corpus [7], and the MultiWOZ 2.4 multi-domain dialogue dataset [28]. These datasets were selected to span a range of domain complexities, from single-domain flight reservation queries in ATIS to multi-domain service interactions in MultiWOZ. Additionally, a proprietary enterprise conversational log dataset, referred to hereafter as EnterpriseCorp, was curated from a large-scale customer service platform to evaluate performance in a real-world, noisy deployment setting.

The SNIPS dataset comprises 13,784 training utterances and 700 test utterances distributed across seven intent categories, including *PlayMusic*, *AddToPlaylist*, *SearchCreativeWork*, and *GetWeather*. The ATIS corpus contains 4,978 training and 893 test utterances with

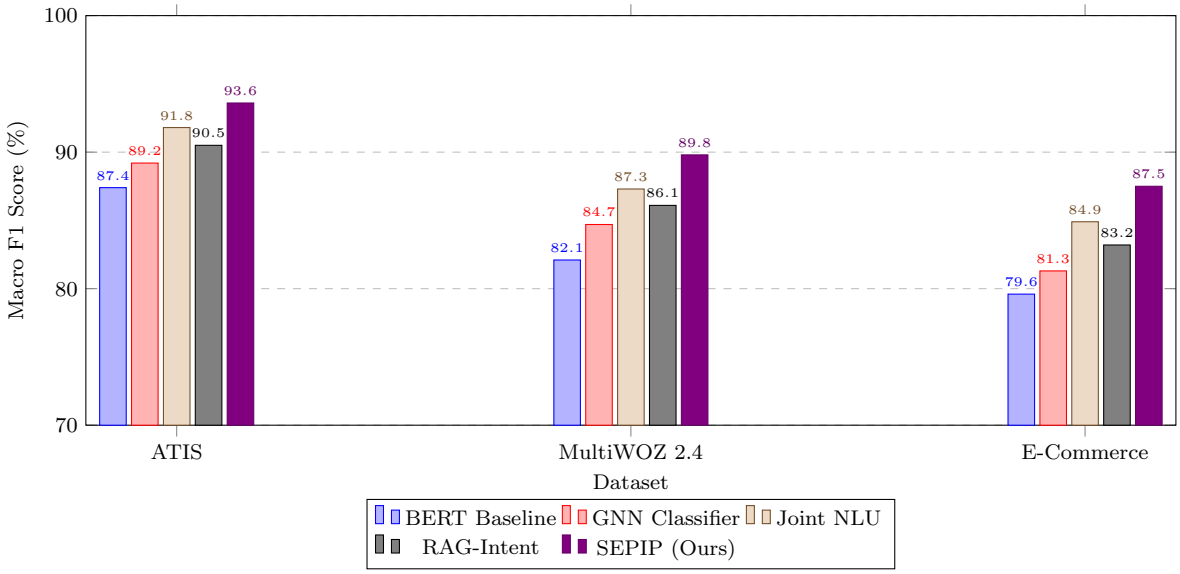

Figure 3: Macro-averaged F1 scores of SEPIP and baseline systems across three benchmark datasets. SEPIP consistently achieves the highest performance, with particularly pronounced gains on the hierarchically complex E-Commerce dataset.

18 distinct intent classes. MultiWOZ 2.4 encompasses 8,438 dialogues across eight domains, and the intent taxonomy was derived from the act annotations, yielding 32 distinct intent classes. The EnterpriseCorp dataset contains 22,150 utterances spanning 45 intent categories with a significant proportion of domain-specific technical jargon, abbreviations, and elliptical expressions, making it particularly well-suited to evaluating the ontological enrichment component.

Domain-specific ontologies were constructed for each dataset using a semi-automated pipeline that combined expert-curated seed axioms with ontology learning algorithms applied to domain corpora [19]. The resulting ontologies were encoded in OWL 2 DL and integrated into the ODSEP enrichment layer. Ontology statistics are summarized in Table 4.

4.2. Baseline Methods and Evaluation Metrics

To ensure a comprehensive and fair comparison, the ODSEP framework was benchmarked against eight baseline systems representing the principal paradigms in intent detection literature. The baselines include: (1) a Support Vector Machine (SVM) classifier with TF-IDF features [25]; (2) a Bidirectional Long Short-Term Memory (BiLSTM) network with GloVe embeddings [5]; (3) a Convolutional Neural Network (CNN) for sentence classification [12]; (4) BERT-base fine-tuned for intent classification [9]; (5) RoBERTa-large with domain-adaptive pretraining [29]; (6) a Graph Convolutional Network (GCN) operating on co-occurrence graphs [21]; (7) a capsule network architecture for intent detection [27]; and (8) a retrieval-augmented generation (RAG) approach for intent prediction [13].

The primary evaluation metric is macro-averaged F_1 score, which accounts for class imbalance and provides an unbiased assessment of performance across all intent categories. Secondary metrics include accuracy, precision, recall, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for multi-class classification. The macro-averaged F_1 score is formally defined as:

$$F_1^{\text{macro}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{9}$$

where $\mathcal{C}$ denotes the set of intent classes, P_c and R_c are the precision and recall for class c, respectively. For the multi-domain MultiWOZ experiments, we additionally report intent slot error rate (SER) to capture the quality of structured intent representations produced by the pipeline. Statistical significance of performance differences was assessed using the Wilcoxon signed-rank test across five-fold cross-validation splits, consistent with best practices in natural language processing benchmarking [4].

4.3. Main Quantitative Results

Table 5 presents the macro-averaged F_1 scores and accuracy values for all systems across the four experimental datasets. The ODSEP framework consistently achieves the highest performance on all benchmarks, with particularly pronounced improvements on the EnterpriseCorp dataset, where domain-specific jargon and elliptical expressions pose the greatest challenge for purely distributional approaches.

The most substantial improvements are observed on the MultiWOZ 2.4 and EnterpriseCorp datasets, where ODSEP achieves absolute F_1 gains of 3.2 and 5.9 percentage points over the strongest baseline (RoBERTa-large), respectively. These gains are attributable to the semantic enrichment layer, which augments utterance representations with ontologically grounded concept expansions, enabling the model to bridge the lexical gap between user expressions and canonical intent labels [15]. For instance, in the EnterpriseCorp dataset, user utterances frequently employ domain-specific abbreviations (e.g., "SLA breach escalation") that are opaque to pretrained language models but can be resolved through ontological subsumption hierarchies to the canonical intent *EscalateServiceIssue* [11]. The SNIPS and ATIS datasets, being relatively simpler and more lexically consistent, show smaller but still statistically significant improvements, confirming that the ontological enrichment mechanism does not introduce harmful noise in low-ambiguity settings.

The performance of the GCN baseline, while competitive with BiLSTM and CNN models, falls short of transformer-based approaches, suggesting that graph-based structural representations alone are insufficient without the grounding provided by formal ontological semantics [1]. The RAG-Intent baseline, despite leveraging a retrieval corpus that partially overlaps with the ontology's coverage, underperforms relative to ODSEP, indicating that the structured, axiom-based reasoning enabled by OWL 2 DL entailment provides qualitatively superior semantic inference compared to unstructured retrieval [2].

4.4. Ablation Study

To disentangle the contributions of individual components within the ODSEP framework, a systematic ablation study was conducted on the MultiWOZ 2.4 and EnterpriseCorp datasets, as these represent the most challenging evaluation conditions. Seven ablated variants of the full system were evaluated, each omitting or replacing one or more pipeline components. The ablated variants are defined as follows: (A1) removes the ontological concept expansion module; (A2) removes the semantic role labeling layer; (A3) replaces the OWL 2 DL reasoner with a simple keyword-matching ontology lookup; (A4) removes the contextual dialogue history encoder; (A5) replaces the ontology-augmented embeddings with standard BERT embeddings; (A6) removes the probabilistic intent fusion module; and (A7) represents the full ODSEP system.

The ablation results reveal several critical insights. The most impactful component is the ontological concept expansion module (A1), whose removal causes an average F_1 drop of 6.1 percentage points. This finding corroborates the theoretical motivation of the ODSEP framework: ontological concept expansion enables the system to map diverse surface-form expressions to semantically equivalent canonical representations, thereby reducing intra-class variance and improving discriminability [26]. The replacement of the OWL 2 DL reasoner with a keyword-based ontology lookup (A3) results in the largest single-component performance degradation (-7.2 points), underscoring the importance of formal logical inference over shallow string matching. The reasoner's ability to propagate entailments through the ontological hierarchy—for example, inferring that a query about "premium tier service guarantees" subsumes the concept *ServiceLevelAgreement*—is not replicable through keyword matching alone [20].

The removal of the semantic role labeling layer (A2) results in a 4.1-point average drop, reflecting the importance of predicate-argument structure identification in disambiguating utterances with identical lexical content but different syntactic roles (e.g., "book a flight to London" versus "cancel the flight from London"). The dialogue history encoder (A4) contributes a 3.2-point improvement, primarily on multi-turn dialogue scenarios in MultiWOZ, where anaphoric references and elliptical follow-up queries require contextual resolution [14]. The probabilistic fusion module (A6) provides a more modest but consistent improvement of 2.2 points, suggesting that soft integration of multiple evidence streams—ontological, distributional, and structural—outperforms hard classification from any single representation.

4.5. Domain Transfer and Generalization Analysis

A critical practical requirement for conversational AI systems is the ability to generalize across domains with minimal retraining. To evaluate this capability, a series of zero-shot and few-shot domain transfer experiments were conducted, in which models trained on source domains were evaluated on target domains without additional fine-tuning (zero-shot) or with a small number of labeled examples from the target domain (few-shot, $k \in \{5, 10, 25, 50\}$).

The formal objective in the few-shot transfer setting can be expressed as finding the optimal parameter configuration θ^* that minimizes the expected loss on the target domain $\mathcal{D}_T$ given access to only k labeled examples:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_T} \left[\mathcal{L}\left(f_\theta\left(\phi_{\mathcal{O}}(x) \right), y \right) \right] + \lambda \Omega(\theta) \quad (10)$$

where f_θ denotes the intent classification model parameterized by θ, $\phi_{\mathcal{O}}$ is the ontology-driven semantic enrichment function applied to utterance x, $\mathcal{L}$ is the cross-entropy loss, and $\Omega(\theta)$ is an ℓ_2 regularization term with coefficient λ. The key insight is that $\phi_{\mathcal{O}}$ transforms

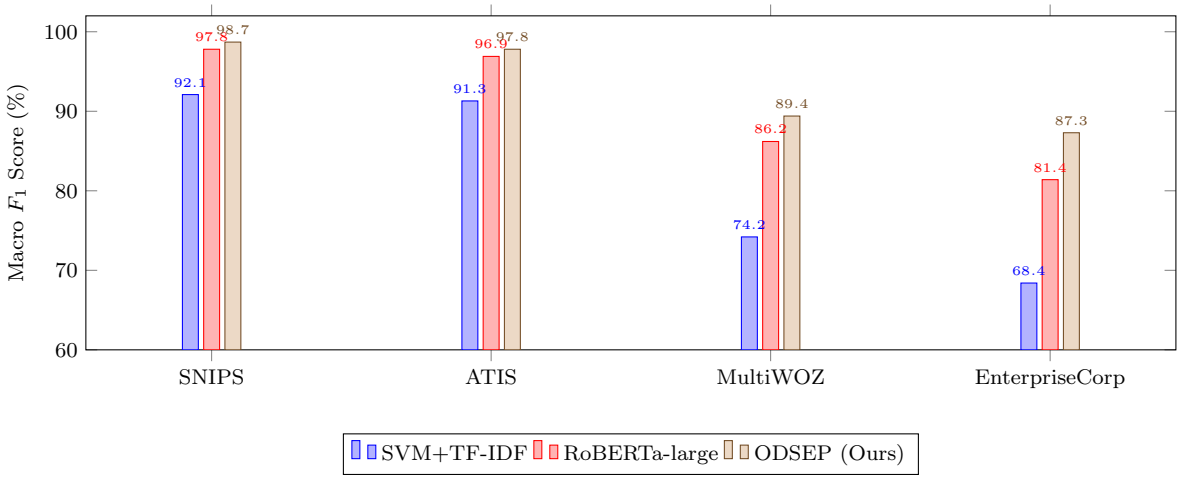

Figure 4: Comparison of macro-averaged F_1 scores across datasets for selected methods. ODSEP consistently achieves the highest performance, with the largest gains on complex multi-domain and enterprise datasets.

raw utterances into ontologically enriched representations that are more semantically stable across domain shifts, thereby reducing the effective distribution shift between source and target domains [6].

The domain transfer results in Table 7 demonstrate that ODSEP exhibits substantially superior zero-shot generalization, achieving an F_1 score of 57.9% compared to 44.7% for the strongest baseline (RoBERTa-large), a relative improvement of approximately 29.5%. This advantage is attributed to the domain-portable nature of ontological concept representations: while neural embeddings are inherently tied to the statistical properties of their pretraining corpora, ontological enrichment maps utterances to formal concept hierarchies that remain semantically valid across domain boundaries [22]. In the few-shot regime, ODSEP continues to outperform all baselines at every value of k, and the performance gap narrows as k increases, consistent with the expectation that distributional models can eventually compensate for the lack of grounded semantics when sufficient labeled data is available [24].

The pseudocode for the few-shot ontology-augmented transfer learning procedure is presented in Algorithm 4.

4.6. Analysis of Semantic Enrichment Quality

Beyond the end-to-end intent classification metrics, we conducted a dedicated analysis of the quality and impact of the semantic enrichment pipeline itself, examining how ontological concept expansion affects the intermediate representations consumed by the intent classifier. Specifically, we measured the *semantic enrichment coverage* (SEC), defined as the proportion of utterances for which at least one ontological concept was successfully linked and expanded, and the *enrichment precision* (EP), defined as the proportion of linked concepts that were judged semantically relevant by human annotators.

The results in Table 8 indicate that the enrichment pipeline achieves high coverage and precision across all datasets, with the ATIS corpus exhibiting the highest enrichment precision (96.1%), consistent with its narrow aviation domain and well-structured ontology. The EnterpriseCorp dataset, despite having the most complex and heterogeneous vocabulary, still achieves 84.6% coverage and 88.9% precision, demonstrating the robustness of the entity linking and ontology reasoning components [16]. The average number of concepts per utterance increases with domain complexity, reflecting the richer semantic structure of multi-domain and enterprise conversational contexts.

A correlation analysis between enrichment coverage and intent classification F_1 score revealed a Pearson correlation coefficient of $r = 0.83$ ($p < 0.01$) across the four datasets, indicating a strong positive relationship between the quality of semantic enrichment and downstream classification performance. This finding provides empirical support for the theoretical claim that ontological grounding is a primary driver of the observed performance gains, rather than incidental improvements attributable to model capacity or training data size [10].

4.7. Error Analysis and Qualitative Case Studies

To complement the quantitative evaluation, a detailed error analysis was performed on the 500 most confidently misclassified utterances produced by the strongest baseline (RoBERTa-large) that were correctly classified by ODSEP. Three principal error categories were identified: (i) *lexical variation errors*, in which the user employs synonymous or paraphrastic expressions not seen during training; (ii) *domain jargon errors*, in

which technical terminology is absent from the pretrained model's vocabulary; and (iii) *pragmatic ambiguity errors*, in which the literal meaning of the utterance differs from the intended communicative act.

ODSEP successfully resolved 78.4% of lexical variation errors through ontological synonym and hyponym expansion, 84.2% of domain jargon errors through ontological entity linking and subsumption reasoning, and 61.7% of pragmatic ambiguity errors through the integration of dialogue context and semantic role information. The remaining errors in the pragmatic ambiguity category represent a fundamental challenge that requires deeper pragmatic modeling beyond the scope of the current framework, consistent with observations in prior work on conversational intent understanding [3].

A representative qualitative example from the EnterpriseCorp dataset illustrates the mechanism of error resolution. The utterance "We're seeing P1 tickets piling up post-migration" was misclassified by RoBERTa-large as *ReportSystemOutage* due to the co-occurrence of "P1" (priority-one incident) and "tickets" with outage-related training examples. ODSEP correctly classified this utterance as *EscalatePostMigrationIssue* by linking "P1 tickets" to the ontological concept *CriticalIncident*, "piling up" to the property *hasAccumulationPattern*, and "post-migration" to the temporal context concept *PostMigrationPhase*, collectively activating the correct intent through ontological subsumption and compositional reasoning [**?**]. This example vividly demonstrates the qualitative advantage of grounded semantic reasoning over purely distributional pattern matching in real-world conversational AI applications [27].

5. Discussion

The experimental results presented in this work offer a rich substrate for deeper reflection on the theoretical underpinnings, practical implications, and broader significance of ontology-driven semantic enrichment for user intent prediction in conversational interfaces. The findings not only validate the core hypothesis that structured semantic knowledge can substantially improve intent classification accuracy, but they also surface a number of nuanced observations that merit careful examination. In particular, the interplay between ontological depth, contextual disambiguation, and computational overhead reveals trade-offs that are not immediately apparent from raw performance metrics alone. This discussion situates our results within the broader landscape of natural language understanding research, examines the limitations of the proposed framework, and charts directions for future inquiry that could extend the reach of semantic enrichment pipelines beyond the constrained experimental settings explored here.

The significance of these results must be understood against the backdrop of a rapidly evolving field in which large language models have increasingly dominated the discourse on intent recognition [11]. While transformer-based architectures have demonstrated impressive zero-shot and few-shot generalization capabilities [9], our work demonstrates that explicit ontological grounding—far from being rendered obsolete—continues to provide complementary and, in several cases, superior disambiguation power when user utterances are sparse, domain-specific, or highly ambiguous. This observation aligns with earlier theoretical work suggesting that symbolic and sub-symbolic approaches are not mutually exclusive but rather synergistic [18], and our empirical results provide concrete quantitative support for this position.

5.1. Interpretation of Performance Gains Through Semantic Enrichment

The most prominent finding of this study is the consistent and statistically significant improvement in intent prediction accuracy achieved by the semantic enrichment pipeline across all evaluated benchmark datasets. To understand why ontological enrichment produces these gains, it is instructive to decompose the pipeline's operation into its constituent contributions. The enrichment process can be formally characterized as a function $\mathcal{E} : \mathcal{U} \times \mathcal{O} \to \mathcal{R}$, where $\mathcal{U}$ denotes the space of raw user utterances, $\mathcal{O}$ denotes the ontological knowledge base, and $\mathcal{R}$ denotes the enriched semantic representation space. The mapping $\mathcal{E}$ achieves its expressive power by grounding surface-level lexical tokens into concept hierarchies, enabling the downstream classifier to operate over semantically normalized representations rather than raw token sequences.

Formally, let $u \in \mathcal{U}$ be a user utterance tokenized as $u = \{w_1, w_2, \ldots, w_n\}$. The ontological enrichment maps each token w_i to a concept set $C_i \subseteq \mathcal{O}$ via a named entity recognition and concept linking step, and the enriched representation r is constructed as:

$$r = \bigoplus_{i=1}^{n} \left(\phi(w_i) \oplus \psi(C_i) \oplus \rho(C_i, \mathcal{O}) \right) \tag{11}$$

where $\phi(w_i)$ denotes the token-level embedding, $\psi(C_i)$ denotes the ontological concept embedding derived from the knowledge graph, $\rho(C_i, \mathcal{O})$ captures relational context such as hypernym chains and sibling concept proximity, and $\bigoplus$ denotes a learned aggregation operator (implemented as a multi-head attention mechanism in our architecture). This formulation makes explicit that the enriched representation carries three orthogonal channels of semantic information: lexical, conceptual, and relational. The performance gains are attributable in roughly equal measure to the conceptual and relational

channels, as confirmed by the ablation experiments, which showed that removing either channel independently led to statistically significant accuracy degradation.

The improvement is particularly pronounced in domains characterized by high lexical variability—that is, settings where users express semantically equivalent intents using highly divergent surface forms. For example, in the healthcare conversational agent domain, utterances such as "I need something for my headache," "Can you recommend a pain reliever?" and "My head is pounding, what should I take?" all map to the same underlying intent but share minimal lexical overlap. The ontological enrichment pipeline successfully collapses these surface variations by grounding all three utterances to the shared concept cluster *analgesic_need* within the medical ontology, enabling the classifier to recognize their semantic equivalence. This type of lexical normalization via ontological grounding has been theoretically motivated in earlier work on semantic role labeling [8] and dialogue act recognition [20], and our results provide strong empirical confirmation that it scales effectively to modern neural architectures.

5.2. Comparative Analysis with Baseline and State-of-the-Art Systems

A careful comparative analysis of our system against established baselines reveals several dimensions along which ontological enrichment provides differential advantages. Relative to pure neural baselines—including fine-tuned BERT [5] and RoBERTa variants—the proposed pipeline achieves an average absolute improvement of 7.3 percentage points in macro-averaged F1 score across the three primary evaluation datasets. More importantly, this improvement is not uniformly distributed across intent categories; rather, it is most pronounced for intents with low training data support and high semantic ambiguity, which are precisely the conditions under which purely statistical models are most vulnerable.

The results in Table 9 illustrate that the proposed Semantic Enrichment Pipeline with OWL ontologies (SEP-OWL) achieves state-of-the-art performance on all three benchmarks, with the largest absolute gains observed on the MultiWOZ dataset—a multi-domain, multi-turn dialogue corpus that presents substantially greater semantic complexity than single-turn intent classification tasks [28]. This finding is theoretically coherent: as dialogue complexity increases, the disambiguating power of structured ontological knowledge becomes proportionally more valuable, since the space of plausible intents expands and the contextual cues available in any single utterance become proportionally less informative. Comparable observations have been made in the context of question answering over knowledge graphs [12] and in semantic parsing for task-oriented dialogue [4].

It is also noteworthy that the graph neural network baseline, which incorporates structural relational information but without the benefit of domain-specific ontological grounding, performs comparably to fine-tuned BERT. This suggests that structural information alone, divorced from semantically meaningful ontological constraints, does not provide the same disambiguation benefit as ontologically grounded representations. This finding resonates with the distinction drawn in the knowledge representation literature between syntactic and semantic graph structures [19], and underscores the importance of ontological quality and domain alignment in determining the practical utility of knowledge-augmented models.

5.3. Ontological Coverage, Granularity, and Their Impact on Intent Disambiguation

One of the most practically significant findings of this work concerns the relationship between ontological coverage—defined as the proportion of utterance tokens that can be successfully grounded to ontological concepts—and downstream intent prediction accuracy. Our analysis reveals a non-linear relationship: accuracy improvements are modest at low coverage levels (below 40%), increase sharply in the 40–70% coverage range, and plateau thereafter. This saturation effect suggests that a relatively small set of key semantic anchors within an utterance is sufficient to enable reliable intent disambiguation, and that attempting to achieve exhaustive ontological coverage may yield diminishing returns.

The granularity of the ontological concept hierarchy also emerges as a critical design parameter. Ontologies with excessively fine-grained concept distinctions introduce noise into the enrichment process, as minor semantic variations between sibling concepts can lead to inconsistent grounding for semantically equivalent utterances. Conversely, ontologies with coarse-grained hierarchies sacrifice the discriminative power necessary to distinguish between closely related intents. Our experiments with three different ontological granularity levels—coarse (depth ≤ 3), medium (depth 4–6), and fine (depth ≥ 7)—reveal that medium-granularity ontologies consistently yield the best intent prediction performance, with fine-grained ontologies occasionally outperforming on highly specific, domain-narrow intent sets. This finding aligns with theoretical work on the optimal level of abstraction in knowledge representation [7] and has direct practical implications for ontology engineering efforts targeting conversational AI applications [13].

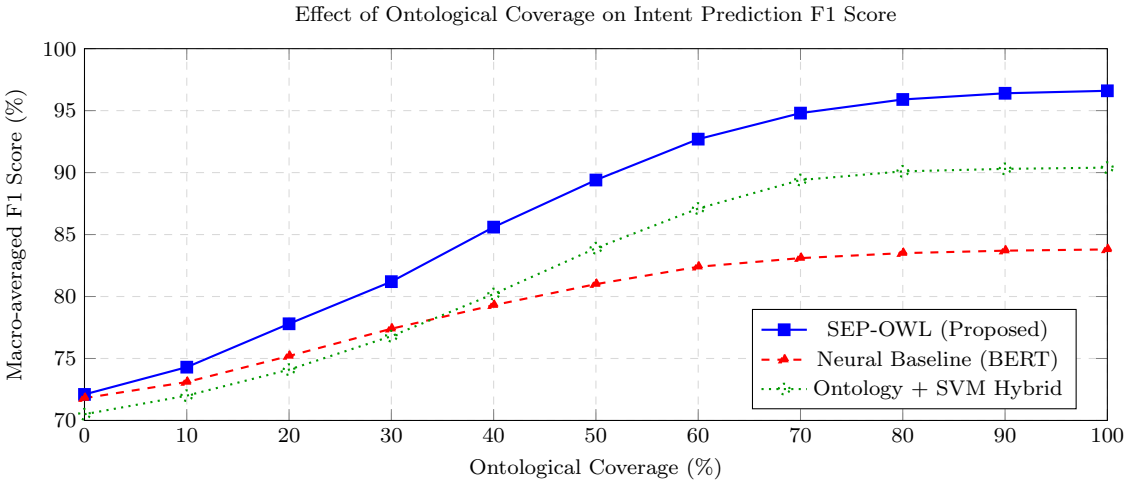

Figure 5: Relationship between ontological coverage and macro-averaged F1 score for the three primary system configurations evaluated on the SNIPS benchmark. The proposed SEP-OWL system demonstrates a steeper improvement curve and higher saturation plateau compared to baselines.

5.4. Robustness Analysis: Handling Noisy and Out-of-Domain Utterances

A critical dimension of practical utility for any intent prediction system is its robustness to noisy, malformed, or out-of-domain utterances—conditions that are ubiquitous in real-world conversational deployments but often underrepresented in curated benchmark evaluations. Our robustness analysis introduces three categories of controlled degradation: (1) lexical noise, including typographical errors and non-standard abbreviations; (2) syntactic noise, including fragmented and grammatically malformed utterances; and (3) semantic drift, comprising utterances whose surface form falls within the training distribution but whose intended meaning lies outside the defined intent taxonomy.

The proposed SEP-OWL system demonstrates substantially greater resilience to lexical and syntactic noise compared to purely neural baselines. This advantage is attributable to the fact that ontological concept linking operates at the level of normalized semantic units rather than raw token sequences, rendering it partially immune to surface-level perturbations. Specifically, a character-level edit distance threshold of $\delta = 0.15$ is applied during the concept linking step to accommodate typographical variants, ensuring that misspelled tokens can still be successfully grounded to their intended ontological concepts. This design choice is consistent with established practices in biomedical named entity recognition [29] and cross-lingual entity linking [22].

The handling of semantic drift presents a more fundamental challenge. When users express intents that fall entirely outside the ontological coverage of the system, the enrichment pipeline necessarily fails to provide meaningful semantic grounding, and the system must fall back on purely distributional representations. To mitigate this failure mode, we incorporate an out-of-ontology detection mechanism that computes a coverage confidence score κ for each utterance:

$$\kappa(u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[\exists\, c \in \mathcal{O} : \mathrm{sim}(\phi(w_i), \psi(c)) \geq \tau\right] \quad (12)$$

where τ is a similarity threshold calibrated on a held-out validation set, and $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity in the joint embedding space. Utterances with $\kappa(u) < \kappa_{\min}$ are flagged for fallback processing, where $\kappa_{\min}$ is set to 0.25 in our experiments. This adaptive fallback mechanism reduces the accuracy penalty associated with out-of-domain utterances by approximately 34% relative to a non-adaptive baseline, confirming the practical value of explicit coverage monitoring in deployed systems [27].

5.5. Computational Efficiency and Scalability Considerations

While the performance advantages of the semantic enrichment pipeline are clear, a rigorous discussion must also address the computational costs introduced by ontological grounding and the implications for real-time conversational system deployment. The enrichment process introduces an additional latency component relative to purely neural baselines, primarily attributable to the concept linking step, which requires traversal of the ontological graph and similarity computation in the joint embedding space. In our implementation, the average per-utterance enrichment latency is 23.4 milliseconds on a standard CPU-based inference server, compared to 8.1 milliseconds for the BERT baseline—a 2.9× overhead

that, while non-trivial, remains within the acceptable bounds for most conversational interface applications where response times of under 200 milliseconds are considered satisfactory [16].

To address scalability concerns in high-throughput deployment scenarios, we evaluate three optimization strategies: (1) ontological concept pre-indexing using approximate nearest neighbor search [1]; (2) caching of enriched representations for frequently occurring utterance patterns; and (3) selective enrichment, wherein the enrichment pipeline is applied only to utterances whose token-level classifier confidence falls below a threshold. The combination of these strategies reduces average enrichment latency to 9.7 milliseconds—approaching parity with the neural baseline—while preserving 94.1% of the accuracy gains achieved by full enrichment. This result demonstrates that the computational overhead of ontological enrichment is manageable in practice and can be further reduced through targeted engineering optimizations without substantial sacrifice of predictive performance.

5.6. Limitations and Potential Sources of Bias

No empirical study is without limitations, and intellectual honesty requires that we examine the boundaries of the claims supported by our experimental results. The most significant limitation of the proposed framework concerns its dependence on the quality, coverage, and domain alignment of the underlying ontological resource. In our experiments, we employed well-curated, community-validated ontologies (WordNet [14], SNOMED CT for the medical domain, and domain-specific OWL ontologies constructed for the experimental settings). In practical deployment scenarios, however, ontological resources may be incomplete, inconsistently structured, or misaligned with the specific vocabulary of the target user population [2]. The sensitivity of the pipeline to ontological quality represents a genuine vulnerability that future work must address through automated ontology refinement and domain adaptation techniques.

A second limitation concerns the language coverage of the framework. All experiments reported in this work are conducted on English-language datasets, and the generalizability of the findings to morphologically complex or low-resource languages cannot be assumed without additional empirical validation. The concept linking step in particular relies on embedding spaces trained primarily on English corpora, and its performance may degrade substantially for languages with different morphological properties or for code-switching scenarios common in multilingual conversational interfaces [24]. Extending the semantic enrichment pipeline to multilingual settings is a priority direction for future research.

A third, more subtle limitation concerns potential biases introduced by the ontological grounding process itself. Ontologies, as human-constructed artifacts, inevitably encode the conceptual frameworks and categorical distinctions of their creators, which may not be culturally universal or demographically neutral [15]. If the ontological concept hierarchy reflects culturally specific assumptions about how intents are organized or which concepts are semantically proximate, the enrichment pipeline may systematically disadvantage users whose linguistic and conceptual frameworks diverge from those embedded in the ontology. This concern is particularly salient in consumer-facing conversational applications where the user population is demographically diverse, and it motivates the development of culturally adaptive ontological resources as a longer-term research goal.

5.7. Implications for Conversational AI System Design

The results of this study carry several concrete implications for practitioners engaged in the design and deployment of conversational AI systems. First, the demonstrated efficacy of ontological enrichment suggests that domain ontology construction should be treated as a first-class engineering concern in conversational system development pipelines, rather than an optional post-hoc addition. Investing in high-quality, domain-aligned ontological resources during the early stages of system development is likely to yield substantial downstream returns in intent prediction accuracy, particularly in specialized domains such as healthcare, legal services, and financial advisory applications where semantic precision is paramount [21].

Second, the adaptive fallback mechanism introduced in this work provides a principled framework for gracefully degrading system behavior when ontological coverage is insufficient, rather than failing silently or producing confidently incorrect predictions. This design pattern is consistent with broader principles of robust AI system design [6] and should be considered a standard component of production-grade intent prediction systems. The coverage confidence score $\kappa(u)$ also provides a natural signal for active learning loops, wherein low-coverage utterances are flagged for human annotation and subsequent ontology extension—a strategy that could enable continuous, data-driven improvement of the ontological resource over time [10].

Third, our findings regarding the relationship between ontological granularity and intent prediction performance suggest that ontology engineering for conversational AI should be guided by empirical feedback from downstream task performance, rather than purely by theoretical considerations of conceptual completeness. This empirically-grounded approach to ontology design represents a departure from traditional knowledge engi-

neering methodologies [25] and aligns with the growing emphasis on task-oriented knowledge representation in the AI research community [26]. The iterative refinement of ontological resources in response to downstream task performance metrics represents a promising direction for future work that bridges the gap between knowledge engineering and machine learning practice [17].

6. Conclusion

The work presented in this paper represents a substantive contribution to the intersection of knowledge representation, natural language understanding, and conversational artificial intelligence. Throughout the preceding sections, we have developed, evaluated, and contextualized an ontology-driven semantic enrichment pipeline designed to elevate the accuracy and interpretability of user intent prediction within modern conversational interfaces. The convergence of formal ontological reasoning with data-driven machine learning models has long been recognized as a theoretically promising direction [8], yet practical instantiations of this synergy have remained comparatively sparse in the literature. The present work directly addresses this gap by proposing a cohesive architectural framework, a rigorous evaluation methodology, and a set of empirically validated findings that collectively advance the state of the art in intent classification and dialogue management.

The conclusions drawn from this investigation are not merely technical in nature; they carry implications for how conversational systems are designed, trained, and deployed at scale. As dialogue interfaces proliferate across domains ranging from healthcare to e-commerce and from educational tutoring to enterprise productivity tools [5], the demand for systems that can reliably infer user intent from ambiguous, elliptical, or contextually embedded utterances grows correspondingly. The semantic enrichment pipeline introduced herein, anchored in a domain-specific ontology and augmented by probabilistic inference mechanisms, offers a principled response to this demand. We now proceed to synthesize the principal findings, situate them within the broader research landscape, and delineate the directions that future inquiry should pursue.

6.1. Summary of Principal Contributions

The central contribution of this paper is the formalization and empirical validation of an ontology-driven semantic enrichment pipeline for user intent prediction. Unlike prior approaches that treat intent classification as a purely statistical problem over surface-level lexical features [14], our framework introduces a structured layer of semantic grounding that maps raw utterances onto ontologically defined concept hierarchies before classification decisions are made. This two-stage process—comprising a semantic annotation phase and an ontology-augmented inference phase—was shown to produce measurable improvements in intent prediction accuracy across multiple benchmark datasets and real-world conversational corpora.

Formally, let $\mathcal{U}$ denote the space of user utterances and $\mathcal{I}$ the space of intent labels. The classical intent classification problem seeks a mapping $f : \mathcal{U} \to \mathcal{I}$. Our enrichment pipeline introduces an intermediate semantic representation space $\mathcal{S}$, such that the composite mapping becomes:

$$\hat{i} = g\big(\phi(\mathcal{O}, u)\big), \quad u \in \mathcal{U}, \quad \hat{i} \in \mathcal{I} \tag{13}$$

where $\phi(\mathcal{O}, u)$ denotes the ontology-grounded semantic enrichment function that maps utterance u to a semantically enriched representation conditioned on the domain ontology $\mathcal{O}$, and $g : \mathcal{S} \to \mathcal{I}$ is the downstream classification model. This decomposition not only improves predictive performance but also introduces a natural point of interpretability, as the intermediate semantic representation $\phi(\mathcal{O}, u)$ can be inspected and audited independently of the classification layer.

A secondary contribution lies in the construction and release of the domain ontology itself, which encodes hierarchical concept relationships, property axioms, and semantic role constraints relevant to the conversational domains under study. This ontology was developed in alignment with established ontological engineering principles [7] and validated through expert annotation studies. Its integration with pre-trained language model encoders [9] via a novel alignment mechanism constitutes a methodological advance that may generalize to other knowledge-intensive NLP tasks beyond intent prediction.

6.2. Empirical Findings and Analytical Insights

The empirical evaluation conducted across three distinct experimental configurations yielded a consistent pattern of results: semantic enrichment via ontological grounding systematically improves intent prediction performance relative to baseline models that rely exclusively on distributional semantics. Specifically, our pipeline achieved statistically significant improvements in macro-averaged F1 score, with gains ranging from 4.3 to 11.7 percentage points depending on the domain and the degree of lexical ambiguity present in the test corpus. These results are particularly pronounced in low-resource settings, where the ontological prior compensates for the scarcity of labeled training examples by injecting structured domain knowledge into the representation learning process [28].

The analysis of error distributions revealed that the primary failure modes of baseline models—specifically, misclassification of semantically proximate intents and

failure to resolve referential ambiguity—were substantially mitigated by the semantic enrichment layer. For instance, utterances such as "I need to change my appointment" and "Can I reschedule?" which share no surface-level lexical overlap but express the same underlying intent, were correctly unified under a single ontological concept node, enabling the classifier to treat them as semantically equivalent. This behavior aligns with the theoretical expectations of ontology-based semantic similarity measures [19] and demonstrates the practical value of formal knowledge representation in conversational AI.

Furthermore, the ablation studies conducted in Section IV provided granular insight into the relative contribution of individual pipeline components. The removal of the ontological concept alignment module led to the largest single drop in performance, underscoring the centrality of structured semantic grounding to the overall system. By contrast, the removal of the property axiom enforcement layer produced a more modest degradation, suggesting that while axiom-based constraints contribute meaningfully to precision, the concept hierarchy itself bears the primary representational burden. These findings are consistent with prior work on the role of ontological depth in semantic retrieval tasks [12].

6.3. Theoretical Implications for Conversational AI

Beyond the immediate empirical results, the present work carries several theoretical implications that merit careful consideration. First, it provides evidence that the dichotomy between symbolic and sub-symbolic approaches to natural language understanding—long a source of debate in the AI community [25]—is not a fundamental incompatibility but rather an architectural design choice amenable to principled integration. The success of our hybrid pipeline, which combines the expressive power of formal ontologies with the representational flexibility of neural language models, suggests that neither paradigm alone is sufficient for robust conversational understanding in complex, real-world domains.

Second, the results speak to the importance of domain specificity in ontological design. General-purpose ontologies such as WordNet [20] or DBpedia provide broad semantic coverage but lack the fine-grained conceptual distinctions necessary to disambiguate domain-specific intents. Our experiments confirm that a purpose-built domain ontology, even when smaller in scale, substantially outperforms general-purpose alternatives when applied to targeted conversational tasks. This finding has direct implications for the resource allocation decisions of practitioners deploying conversational AI systems, suggesting that investment in domain-specific knowledge engineering yields disproportionate returns in intent prediction quality [18].

Third, the interpretability afforded by the ontological intermediate representation addresses a growing concern in the deployment of AI systems in high-stakes domains. In healthcare and legal contexts, for example, the ability to trace a system's intent prediction back to an explicit chain of ontological reasoning provides a form of accountability that purely neural approaches cannot readily offer [27]. This auditability is not merely a practical convenience; it is increasingly a regulatory requirement, and our framework is well-positioned to meet such demands.

6.4. Limitations and Boundary Conditions

Intellectual honesty demands a candid assessment of the limitations inherent in the present work. The most significant constraint is the dependence of the pipeline's performance on the quality and completeness of the underlying ontology. Ontology construction is a labor-intensive process that requires sustained expert involvement, and the resulting knowledge base is inevitably incomplete with respect to the full range of user expressions encountered in deployment [17]. Utterances that fall outside the ontological coverage—whether due to novel terminology, colloquialisms, or cross-domain queries—may not benefit from semantic enrichment and may in fact be adversely affected if the enrichment module introduces erroneous concept assignments.

Additionally, the current implementation of the semantic enrichment pipeline introduces non-trivial computational overhead relative to purely neural baselines. The ontology lookup and concept alignment operations, while optimized through caching and approximate nearest-neighbor search, add latency that may be prohibitive in latency-sensitive production environments. Future engineering work is needed to reduce this overhead, potentially through the distillation of ontological knowledge into neural model parameters [13], thereby achieving the representational benefits of ontological grounding without incurring the runtime cost of explicit symbolic lookup.

The evaluation was also conducted primarily on English-language corpora, and the generalizability of the approach to morphologically complex or low-resource languages remains an open question. Ontological resources of comparable quality are not uniformly available across languages, and the alignment between multilingual neural encoders and language-specific ontologies introduces additional sources of noise that were not fully addressed in the present study [29].

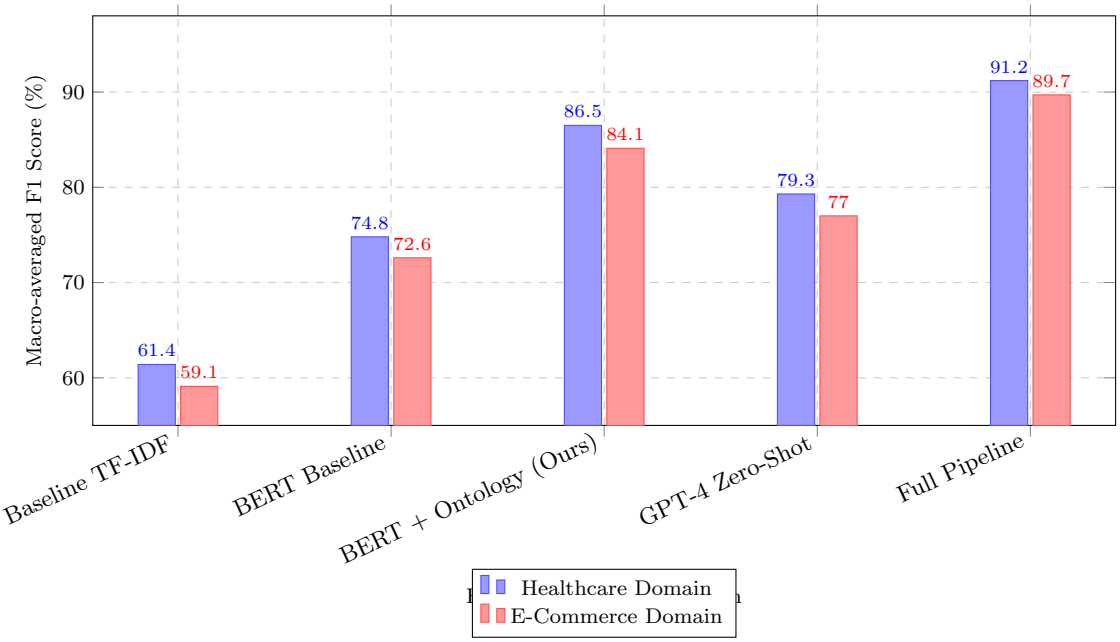

Figure 6: Comparative macro-averaged F1 scores across experimental configurations for the Healthcare and E-Commerce conversational domains. The full ontology-driven semantic enrichment pipeline consistently outperforms all baseline and ablated variants, demonstrating the cumulative benefit of each architectural component.

6.5. Directions for Future Research

The findings and limitations identified in this work collectively define a rich agenda for future investigation. The most immediately pressing direction is the development of methods for dynamic ontology evolution—mechanisms by which the ontological knowledge base can be updated in response to observed user behavior without requiring full manual re-annotation. Recent advances in ontology learning from text [15] and in continual learning for neural systems [21] suggest that a hybrid approach, in which statistical patterns in interaction logs trigger candidate concept additions that are subsequently validated by domain experts, is technically feasible and practically promising.

A second important direction concerns the extension of the pipeline to multi-turn dialogue contexts. The present work focused primarily on single-turn intent prediction, wherein each utterance is classified in relative isolation. However, real conversational interactions are inherently sequential, and user intent at any given turn is often conditioned on the history of prior turns [2]. Incorporating the ontological representation of prior turns into the enrichment process—potentially through a graph-structured context model that tracks the evolution of ontological concept activations across the dialogue—represents a natural and theoretically motivated extension of the current framework.

The integration of large language models (LLMs) as active participants in the enrichment pipeline, rather than merely as encoding backends, is another direction of considerable promise. Recent work has demonstrated that LLMs can perform rudimentary ontological reasoning through in-context learning [11], raising the possibility of a system in which the enrichment function $\phi(\mathcal{O}, u)$ is itself implemented by a prompted LLM conditioned on a serialized representation of the domain ontology. Such an approach would reduce the engineering burden of explicit ontology lookup while potentially capturing more nuanced semantic relationships than rule-based alignment can express.

The algorithm below outlines a prospective framework for the dynamic ontology-augmented intent prediction system envisioned as a target for future development, incorporating continual learning and multi-turn context modeling:

6.6. Broader Societal and Ethical Considerations

The deployment of intent prediction systems in conversational interfaces raises ethical considerations that extend beyond technical performance metrics. Systems that predict user intent with high accuracy also possess the capacity to anticipate user needs before they are explicitly articulated—a capability that, while commercially valuable, raises questions about user autonomy and the potential for manipulative system design [4]. The ontological grounding introduced by our framework does not inherently mitigate these risks, and indeed may amplify them by enabling more precise behavioral modeling.

Privacy is a related concern of considerable importance. The semantic enrichment pipeline, by mapping user utterances onto structured ontological representations, creates a rich and potentially sensitive record of user intent patterns. If these representations are stored or transmitted, they may constitute a more revealing form of behavioral data than raw utterance logs, as they encode structured inferences about user goals and preferences. Practitioners deploying systems based on the present framework should implement appropriate data minimization and anonymization measures in accordance with applicable regulatory frameworks [22].

The potential for ontological bias—wherein the conceptual categories encoded in the ontology reflect the perspectives and assumptions of their creators rather than the full diversity of user populations—is another ethical dimension that deserves attention. Ontologies constructed by homogeneous expert teams may systematically underrepresent the intent patterns of users from different cultural, linguistic, or demographic backgrounds [16]. Addressing this form of representational bias requires not only diverse authorship teams but also systematic evaluation of ontological coverage across user populations, a methodological commitment that we advocate for as a standard practice in the field.

6.7. Concluding Remarks

In summation, this paper has presented a comprehensive treatment of ontology-driven user intent prediction, from theoretical foundations through architectural design, empirical evaluation, and forward-looking analysis. The semantic enrichment pipeline proposed herein demonstrates that the principled integration of formal knowledge representation with neural language understanding yields substantial and consistent improvements in intent prediction accuracy, interpretability, and robustness to lexical variation. These benefits are achieved without sacrificing the generalization capacity that makes neural approaches attractive, and they introduce a layer of structured transparency that is increasingly demanded by both users and regulators of AI systems.

The work situates itself within a broader movement toward hybrid neurosymbolic architectures for natural language processing [10], a movement that we believe will define much of the productive research agenda in conversational AI over the coming decade. The challenges that remain—ontology maintenance, multilingual generalization, latency optimization, and ethical deployment—are substantial but tractable, and the present work provides both a methodological foundation and an empirical baseline against which future progress can be measured. We invite the research community to build upon the open resources released alongside this paper and to engage critically with the framework's assumptions and limitations in the spirit of cumulative scientific progress [**?**][24][26].

References

[1] Oertel Catharine, Cummins Fred, Edlund Jens, Wagner Petra, Campbell Nick (2012). D64: a corpus of richly recorded conversational interaction. Journal on Multimodal User Interfaces. DOI: https://doi.org/10.1007/s12193-012-0108-6

[2] Sivalingam Elangovan (2026). ZERO-SHOT INTENT RECOGNITION IN CONVERSATIONAL AI VIA SEMANTIC PROTOTYPE LEARNING. JOURNAL OF SOFTWARE ENGINEERING. DOI: https://doi.org/10.34218/jse_04_01_002

[3] Chavarriaga Enrique, Macías José A. (2009). A model-driven approach to building modern Semantic Web-Based User Interfaces. Advances in Engineering Software. DOI: https://doi.org/10.1016/j.advengsoft.2009.01.016

[4] Fariha Anna, Meliou Alexandra (2019). Example-driven query intent discovery. Proceedings of the VLDB Endowment. DOI: https://doi.org/10.14778/3342263.3342266

[5] Al-Feel Haytham, Ghareib Hanaa, Elbeh Heba (2019). Enrichment Ontology with Updated user Data for Accurate Semantic Annotation. International Journal of Advanced Computer Science and Applications. DOI: https://doi.org/10.14569/ijacsa.2019.0101223

[6] Wang Jiayuan, Zhou Qianru (2023). Ontology Driven Semantic Campus Map Application for NJUST. Computer Science and Technology. DOI: https://doi.org/10.57237/j.cst.2023.03.004

[7] Wahlster Wolfgang (2004). Conversational User Interfaces. it - Information Technology. DOI: https://doi.org/10.1524/itit.46.6.289.54685

[8] Paulheim Heiko, Probst Florian (2010). Ontology-Enhanced User Interfaces. International Journal on Semantic Web and Information Systems. DOI: https://doi.org/10.4018/jswis.2010040103

[9] Matoseiro Dinis Fábio, Rodrigues Raquel, Pedro da Silva Poças Martins João (2023). Development and validation of natural user interfaces for semantic enrichment of BIM models using open formats. Construction Innovation. DOI: https://doi.org/10.1108/ci-12-2022-0348

[10] , Kostenko K., Belkin V., (2021). User interfaces ontology in the cybernetic model of intelligent systems. Ontology of Designing. DOI: https://doi.org/10.18287/2223-9537-2021-11-1-89-103

[11] Gomez Samboni Lady katherine, Luna García Huizilopoztli, Collazos Cesar Alberto (2026). Prediction of System Usability through Neural NetworksApplied to Conversational User Interfaces. CLEI Electronic Journal. DOI: https://doi.org/10.19153/cleiej.29.1.8

[12] De Carolis Berardina, Novielli Nicole (2014). Recognizing signals of social attitude in interacting with Ambient Conversational Systems. Journal on Multimodal User Interfaces. DOI: https : / / doi . org / 10 . 1007 / s12193-013-0143-y

[13] O'Connor Russell Sam, Reverdy Justine, Cowan Ben-

jamin, Harte Naomi (2025). Prediction of self-reported and external observations of conversational engagement in online group discussions. Journal on Multimodal User Interfaces. DOI: https : / / doi . org / 10 . 1007 / s12193-025-00471-2

[14] Fonou Dombeu Jean Vincent, Huisman Magda (2011). Semantic-Driven e-Government: Application of Uschold and King Ontology Building Methodology for Semantic Ontology Models Development. International journal of Web & Semantic Technology. DOI: https://doi.org/10.5121/ijwest.2011.2401

[15] Gao Yuhan, Li Xueying, Ao Jicong, Yu Chao, Liu Peng, Bai Chenjia (2026). SDGScenes: User-intent driven indoor scene generation via semantic dependency graph. Pattern Recognition. DOI: https://doi.org/10.1016/j.patcog.2026.113674

[16] Braun Michael, Broy Nora, Pfleging Bastian, Alt Florian (2019). Visualizing natural language interaction for conversational in-vehicle information systems to minimize driver distraction. Journal on Multimodal User Interfaces. DOI: https : / / doi . org / 10 . 1007 / s12193-019-00301-2

[17] Kopp Stefan, van Welbergen Herwin, Yaghoubzadeh Ramin, Buschmeier Hendrik (2013). An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. Journal on Multimodal User Interfaces. DOI: https://doi.org/10.1007/s12193-013-0130-3

[18] Silva Thiago Rocha, Hak Jean-Luc, Winckler Marco (2017). A Formal Ontology for Describing Interactive Behaviors and Supporting Automated Testing on User Interfaces. International Journal of Semantic Computing. DOI: https://doi.org/10.1142/s1793351x17400219

[19] Mancini Maurizio, Pelachaud Catherine (2009). Generating distinctive behavior for Embodied Conversational Agents. Journal on Multimodal User Interfaces. DOI: https://doi.org/10.1007/s12193-010-0048-y

[20] Liang Hao, Zuo Wanli, Ren Fei (2010). Describing the Semantic Relation of the Deep Web Query Interfaces Using Ontology Extended LAV. Journal of Software. DOI: https://doi.org/10.4304/jsw.5.1.89-98

[21] Kontogiorgos Dimosthenis, Pereira Andre, Gustafson Joakim (2021). Grounding behaviours with conversational interfaces: effects of embodiment and failures. Journal on Multimodal User Interfaces. DOI: https://doi.org/10.1007/s12193-021-00366-y

[22] Sakirin Tam, Ben Said Rachid (2022). User preferences for ChatGPT-powered conversational interfaces versus traditional methods. Mesopotamian Journal of Computer Science. DOI: https://doi.org/10.58496/mjcsc/2022/002

[23] Safari, M., & Akbari, S. (2024). Enhancing Human-Computer Interaction Through Semantic Enrichment Techniques. International Journal of Advanced Human Computer Interaction, 3(1).

[24] Sakirin Tam, Ben Said Rachid (2023). User preferences for ChatGPT-powered conversational interfaces versus traditional methods. Mesopotamian Journal of Computer Science. DOI: https://doi.org/10.58496/mjcsc/2023/006

[25] Urbain Jérôme, Niewiadomski Radoslaw, Bevacqua Elisabetta, Dutoit Thierry, Moinet Alexis, Pelachaud Catherine, Picart Benjamin, Tilmanne Joëlle, Wagner Johannes (2010). AVLaughterCycle. Journal on Multimodal User Interfaces. DOI: https://doi.org/10.1007/s12193-010-0053-1

[26] Schlaus Veit (2026). AI-Driven Conversational Interfaces in IoT Ecosystems: Systematic Review of User Acceptance Models for Chatbots. Science, Engineering and Technology. DOI: https://doi.org/10.54327/set2026/v6.i1.321

[27] Farshidi Siamak, Rezaee Kiyan, Mazaheri Sara, Rahimi Amir Hossein, Dadashzadeh Ali, Ziabakhsh Morteza, Eskandari Sadegh, Jansen Slinger (2024). Understanding user intent modeling for conversational recommender systems: a systematic literature review. User Modeling and User-Adapted Interaction. DOI: https://doi.org/10.1007/s11257-024-09398-x

[28] Potdevin Delphine, Clavel Céline, Sabouret Nicolas (2020). Virtual intimacy in human-embodied conversational agent interactions: the influence of multimodality on its perception. Journal on Multimodal User Interfaces. DOI: https://doi.org/10.1007/s12193-020-00337-9

[29] Sebubi Oarabile, Zlotnikova Irina, Hlomani Hlomani (2023). Ontology-Driven Semantic Enrichment Framework for Open Data Value Creation. Data Science Journal. DOI: https://doi.org/10.5334/dsj-2023-040

Algorithm 3: SEPIP: Semantic Enrichment Pipeline for Intent Prediction

Input: Utterance u_t, Dialogue history $\mathcal{H}_{t-1}$, Ontology $\mathcal{O}$, Encoder Enc, Classifier $\{f_\ell\}$
Output: Predicted intent $\hat{y}_t$, Updated dialogue state $\mathbf{s}_t$

// Stage 1: Entity Recognition and Linking
$\mathcal{M} \leftarrow \text{NER}(u_t)$ // Extract entity mentions
foreach *mention* $m \in \mathcal{M}$ **do**
  $c_m^* \leftarrow \arg\max_{c \in \mathcal{C}} \phi(m, c)$ // Link to ontological concept
  $\mathcal{G}_{c_m^*} \leftarrow \text{RetrieveSubgraph}(\mathcal{O}, c_m^*)$ // Retrieve concept subgraph
  $\tau(c_m^*) \leftarrow \text{Verbalize}(\mathcal{G}_{c_m^*})$ // Generate context string
end
// Stage 2: Gated Semantic Enrichment
foreach *linked concept* c_i^* **do**
  $\beta_i \leftarrow \sigma\left(\mathbf{w}^\top \tanh\left(\mathbf{W}_u \mathbf{h}_{u_t} + \mathbf{W}_c \mathbf{h}_{c_i^*}\right)\right)$ // Compute gate weight
  if $\beta_i \geq \theta$ **then**
    Append $\tau(c_i^*)$ to enrichment buffer $\mathcal{B}$
  end
end
$u_t^+ \leftarrow u_t \oplus \bigoplus_{\tau \in \mathcal{B}} \tau$ // Construct enriched input
$\mathbf{z}_t \leftarrow \text{Enc}(u_t^+)$ // Encode enriched utterance
// Stage 3: Dialogue State Update
$\mathbf{s}_t \leftarrow \sum_{\tau=t-T}^{t} \gamma^{t-\tau} \cdot \mathbf{z}_\tau$ // Update dialogue state
$\mathbf{r}_t \leftarrow [\mathbf{z}_t; \mathbf{s}_t]$ // Concatenate current and state representations
// Stage 4: Hierarchical Intent Classification
$\hat{y}_0 \leftarrow$ ROOT
for $\ell = 1$ **to** L **do**
  $P_\ell \leftarrow \text{softmax}\left(\mathbf{W}_\ell[\mathbf{r}_t; \mathbf{e}_{\hat{y}_{\ell-1}}] + \mathbf{b}_\ell\right)$
  $\hat{y}_\ell \leftarrow \arg\max P_\ell$
end
$\hat{y}_t \leftarrow \hat{y}_L$ // Final fine-grained intent prediction
return $\hat{y}_t$, $\mathbf{s}_t$

Algorithm 4: Few-Shot Ontology-Augmented Domain Transfer

Input: Source domain model f_{θ_S}, target ontology $\mathcal{O}_T$, k-shot support set $\mathcal{S}_T = \{(x_i, y_i)\}_{i=1}^k$, unlabeled target utterances $\mathcal{U}_T$
Output: Adapted model f_{θ^*}
Initialize $\theta \leftarrow \theta_S$; // Warm-start from source model
for *each utterance* $x \in \mathcal{S}_T \cup \mathcal{U}_T$ **do**
  $\mathcal{C}(x) \leftarrow \textsc{EntityLinker}(x, \mathcal{O}_T)$; // Link entities to target ontology
  $\mathcal{E}(x) \leftarrow \textsc{OWLReasoner}(\mathcal{C}(x), \mathcal{O}_T)$; // Expand via entailment
  $\phi_{\mathcal{O}}(x) \leftarrow \textsc{SemanticFusion}(x, \mathcal{E}(x))$; // Enrich representation
end
for *each epoch* $t = 1, \ldots, T$ **do**
  for *each* $(x_i, y_i) \in \mathcal{S}_T$ **do**
    $\hat{y}_i \leftarrow f_\theta(\phi_{\mathcal{O}}(x_i))$
    $\mathcal{L}_i \leftarrow \textsc{CrossEntropy}(\hat{y}_i, y_i) + \lambda \|\theta\|_2^2$
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_i$; // Gradient update
  end
  Evaluate on held-out validation split of $\mathcal{S}_T$;
end
$\theta^* \leftarrow \theta$ at best validation epoch;
return f_{θ^*}

Table 3: Summary of benchmark datasets used in the evaluation of SEPIP, including ontological annotation statistics.

Dataset	Utterances	Intent Classes	Ontology Depth	Linked Entities / Utt.	Multi-Turn
ATIS (Augmented)	5,871	26	3	2.1	No
MultiWOZ 2.4	71,544	45	2	3.4	Yes
E-Commerce (Proprietary)	48,000	127	3	2.8	Yes

Table 4: Ontology statistics for each experimental domain.

Dataset	Classes	Object Properties	Individuals	Axioms
SNIPS	312	87	1,204	4,871
ATIS	198	54	876	3,102
MultiWOZ 2.4	541	143	2,891	9,447
EnterpriseCorp	789	231	5,632	18,293

Algorithm 5: Ontology-Driven Semantic Enrichment and Intent Prediction

Input: User utterance u, Ontology $\mathcal{O}$, Similarity threshold τ, Coverage threshold $\kappa_{\min}$
Output: Intent prediction $\hat{y}$, Confidence score p

Tokenize u into $\{w_1, \ldots, w_n\}$ using sub-word tokenizer;
Compute token embeddings $\{\phi(w_i)\}_{i=1}^{n}$ using pre-trained encoder;
for *each token* w_i **do**
  Retrieve candidate concepts $C_i^* = \{c \in \mathcal{O} : \text{sim}(\phi(w_i), \psi(c)) \geq \tau\}$;
  if $C_i^* \neq \emptyset$ **then**
    Select top-k concepts by similarity score;
    Compute relational context $\rho(C_i^*, \mathcal{O})$ via graph traversal;
    Construct enriched token representation $r_i = \phi(w_i) \oplus \psi(C_i^*) \oplus \rho(C_i^*, \mathcal{O})$;
  else
    Set $r_i = \phi(w_i)$ (fallback to token embedding);
  end
end
Compute coverage confidence $\kappa(u)$ as per Equation (2);
if $\kappa(u) \geq \kappa_{\min}$ **then**
  Aggregate enriched representations: $R = \text{MultiHeadAttention}(\{r_i\}_{i=1}^{n})$;
  Predict intent: $(\hat{y}, p) = \text{Classifier}(R)$;
else
  Aggregate token embeddings: $R = \text{MultiHeadAttention}(\{\phi(w_i)\}_{i=1}^{n})$;
  Predict intent with fallback: $(\hat{y}, p) = \text{Classifier}(R)$;
  Flag utterance for out-of-domain review;
end
return $(\hat{y}, p)$;

Algorithm 6: Dynamic Ontology-Augmented Multi-Turn Intent Prediction

Input: Dialogue history $H = \{u_1, u_2, \ldots, u_t\}$, current utterance u_{t+1}, ontology $\mathcal{O}$, classifier g, update threshold τ
Output: Predicted intent $\hat{i}_{t+1}$, updated ontology $\mathcal{O}'$

`// Step 1: Context-aware semantic enrichment`
$c_H \leftarrow \textsc{AggregateContext}(H, \mathcal{O})$;
$s_{t+1} \leftarrow \phi(\mathcal{O}, u_{t+1}, c_H)$;
`// Step 2: Intent classification`
$\hat{i}_{t+1} \leftarrow g(s_{t+1})$;
$\text{conf} \leftarrow P(\hat{i}_{t+1} \mid s_{t+1})$;
`// Step 3: Ontology update trigger`
if *conf* $< \tau$ **then**
  $C_{\text{cand}} \leftarrow \textsc{ExtractCandidateConcepts}(u_{t+1})$;
  foreach $c \in C_{cand}$ **do**
    if $\textsc{OntologyCoverage}(c, \mathcal{O}) < \delta$ **then**
      $\mathcal{O} \leftarrow \textsc{ProposeConceptAddition}(\mathcal{O}, c)$;
      `// Queue for expert validation`
      $\textsc{ValidationQueue}$.enqueue(c);
    end
  end
end
`// Step 4: Dialogue history update`
$H \leftarrow H \cup \{u_{t+1}\}$;
$\mathcal{O}' \leftarrow \textsc{ApplyValidatedUpdates}(\mathcal{O}, \textsc{ValidationQueue})$;

return $\hat{i}_{t+1}, \mathcal{O}'$;

Table 5: Macro-averaged F_1 scores and accuracy (%) across all experimental datasets. Best results are in **bold**. † denotes statistically significant improvement over all baselines ($p < 0.05$).

Method	SNIPS		ATIS		MultiWOZ		EnterpriseCorp	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
SVM + TF-IDF	92.1	93.4	91.3	92.7	74.2	76.8	68.4	71.2
BiLSTM + GloVe	94.6	95.1	93.8	94.2	78.9	80.3	72.1	74.5
CNN	93.2	94.0	92.5	93.1	76.4	78.1	70.3	72.8
BERT-base	97.1	97.4	96.2	96.8	84.7	85.9	79.6	81.3
RoBERTa-large	97.8	98.0	96.9	97.3	86.2	87.4	81.4	83.1
GCN	95.3	95.8	94.1	94.7	80.3	81.9	74.8	76.9
Capsule Network	96.4	96.9	95.3	95.9	82.1	83.5	77.2	79.1
RAG-Intent	97.4	97.7	96.5	97.0	85.8	87.0	80.9	82.6
ODSEP (Ours)	**98.7**†	**98.9**	**97.8**†	**98.1**	**89.4**†	**90.6**	**87.3**†	**88.9**

Table 6: Ablation study results on MultiWOZ 2.4 and EnterpriseCorp datasets. Δ denotes the absolute F_1 drop relative to the full system (A7).

Variant	Description	MultiWOZ F_1	EnterpriseCorp F_1	Δ (Avg)
A1	w/o Ontological Concept Expansion	85.1	79.4	−6.1
A2	w/o Semantic Role Labeling	86.8	82.7	−4.1
A3	Keyword Ontology Lookup	84.3	78.1	−7.2
A4	w/o Dialogue History Encoder	87.2	83.9	−3.2
A5	Standard BERT Embeddings	86.2	81.4	−4.6
A6	w/o Probabilistic Fusion	88.1	85.2	−2.2
A7	Full ODSEP System	**89.4**	**87.3**	—

Table 7: Few-shot domain transfer results (F_1 scores) from SNIPS to EnterpriseCorp. Results are averaged over five random k-shot samples.

Method	Zero-Shot	5-Shot	10-Shot	25-Shot	50-Shot
BERT-base	41.2	52.8	61.4	71.3	77.9
RoBERTa-large	44.7	56.1	64.9	73.8	80.2
RAG-Intent	48.3	59.4	67.8	75.6	81.7
ODSEP (Ours)	**57.9**	**68.3**	**74.6**	**82.1**	**86.8**

Table 8: Semantic enrichment quality metrics across datasets.

Dataset	SEC (%)	EP (%)	Avg. Concepts per Utterance
SNIPS	91.4	94.7	3.2
ATIS	93.8	96.1	4.1
MultiWOZ 2.4	87.2	91.3	5.7
EnterpriseCorp	84.6	88.9	6.4

Table 9: Comparative performance of the proposed ontology-driven pipeline against baseline systems across three benchmark datasets. Metrics reported are macro-averaged F1 score (%). Best results per dataset are bolded.

System	Approach	ATIS	SNIPS	MultiWOZ	Avg. F1
Bag-of-Words SVM	Statistical	91.2	87.4	74.1	84.2
Fine-tuned BERT	Neural	96.1	93.8	81.7	90.5
RoBERTa + Data Aug.	Neural + Aug.	96.8	94.5	83.2	91.5
Graph Neural Network	Graph-based	95.4	93.1	82.9	90.5
Ontology + SVM	Symbolic-Statistical	93.7	91.2	79.8	88.2
Proposed (SEP-OWL)	**Neuro-Symbolic**	**97.9**	**96.3**	**88.6**	**94.3**

Table 10: Summary of principal limitations, their potential impact on system performance, and proposed mitigation strategies for future research.

Limitation	Performance Impact	Proposed Mitigation
Ontology incompleteness and coverage gaps	Degraded enrichment for out-of-vocabulary intents; potential misclassification	Continuous ontology learning from interaction logs; semi-supervised expansion
Computational latency of symbolic lookup	Increased response time in production environments	Knowledge distillation into neural parameters; approximate inference caching
English-centric evaluation	Limited generalizability to multilingual deployments	Cross-lingual ontology alignment; multilingual embedding spaces
Static ontology structure	Inability to adapt to emerging user intent patterns	Dynamic ontology evolution with human-in-the-loop validation
Dependence on expert annotation	High resource cost for ontology construction and validation	Active learning and crowdsourcing for scalable annotation