

Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems

Khyati C. Kanani ¹, , Gopi B. Sanghani ¹

1. Computer Science & Engineering, Darshan University, Rajkot, IND

Received: April 08, 2026 | Review began: May 05, 2026 | Review ended: June 23, 2026 | Published: June 28, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Intent detection is the basis for real-world dialogue systems such as virtual assistants, conversational agents, and automated customer support, enabling these applications to understand the user's intent. While deploying a single unified classification model across multiple domains is difficult because of the variety of linguistic structures and annotation systems in the numerous datasets (ATIS, SNIPS, and TOPv2) available, the resulting domain variability can lead to a lack of generalizability and stability in the unified intent classification model. To solve this problem, we present Heterogeneity Aware Transformer for Intent Detection (HyTra-ID), a robust transformer-based framework that addresses the inconsistencies that arise from varying domains through a unified preprocessing pipeline, domain-aware dataset embedding modules, and an enhanced classification architecture specifically designed for large-scale multi-domain intent detection. We learn a single-fused representation that combines semantic sentence features with dataset-specific embedding cues needed for HyTra-ID to adapt to various sources of conversation. Our extensive experimental and ablation study results in a combined benchmark of 102 unified intent classes. HyTra-ID shows robust performance across the combined ATIS, SNIPS, and TOPv2 benchmark datasets and in producing accurate predictions of 96.34% for overall accuracy on HyTra-ID unified test set and outperforming all simplified ablation variants tested. HyTra-ID provides a scalable, domain-robust framework for real-world dialogue systems with consistent intent understanding across multiple domains.

Categories: Artificial Intelligence and Machine Learning in the Cloud, Machine Learning (ML), Deep Learning

Keywords: intent detection, cross-domain natural language understanding, multi-domain dialogue systems, multi-task learning, domain adaptation

Introduction

The evolution of Conversational Artificial Intelligence (AI) has made it a core technology for intelligent virtual assistants, automated customer support, task-oriented dialogue systems, and decision-support platforms. In such systems, Natural Language Understanding (NLU) enables the interpretation of user utterances and supports the selection of appropriate system actions. A central component of NLU is intent recognition, which guides downstream processes such as dialogue state tracking, slot filling, and response generation. As a result, inaccurate intent prediction can propagate through the dialogue pipeline, leading to inappropriate system behavior and reduced user confidence.

The development of deep learning, particularly innovations in transformer architectures, has enabled the accuracy of an intent classifier to greatly increase through the training of models to provide rich representations of contextual aspects of language [1,2]. These types of models have been shown to perform well on popular publicly available benchmark datasets such as the ATIS, SNIPS, and TOPv2 datasets, which indicate that model-based intent classification using

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

Transformers should be the preferred method of designing intent classifiers [3]. Many prior studies on intent classification evaluate models under in-domain settings, where training and test samples are drawn from the same dataset distribution, which may not fully reflect real-world multi-domain deployment conditions.

In real-world conversational systems, user requests often span multiple functional domains, such as navigation, scheduling, weather, and purchasing. In this work, we define cross-domain intent variability as the variation in intent expression and intent-label structure that arises when semantically related user requests are drawn from different datasets or domains with distinct annotation schemes, linguistic styles, and class granularities. Previous research has shown that dataset bias and class imbalance can negatively affect model performance in single-domain settings [4]. Likewise, although joint intent-semantic models provide richer contextual representations, they are typically developed under homogeneous training assumptions and do not explicitly address dataset-level heterogeneity across domains [5,6].

To address the problem of domain shifts, quite a few research works have examined domain adaptation and multi-domain learning techniques [7,8]. The purpose of these techniques is to make the feature representation of domains as consistent as possible through the use of Fine-Tuning and Shared Latent Spaces. However, despite their potential success, these techniques are usually designed under the assumption that differences between domains represent noise, whereas it is likely that differences represent useful information (signals), and therefore should be modeled. Likewise, Open-Set Intent Detection frameworks are designed to identify unseen intents [9] but do not solve the problem of producing a single representation for multiple datasets that each have their own unique characteristics. Consequently, most of the current techniques will show varying degrees of performance across the different domains, which limits their ability to scale in a real-world dialogue system.

Due to the limitations identified, this document presents Heterogeneity-Aware Transformer for Intent Detection (HyTra-ID), which is a robust transformer-based framework that was explicitly developed to address variability in intent across multiple domains. The core idea of HyTra-ID is to augment contextual sentence representations with an explicit dataset-identity embedding during training. Rather than assuming that all utterances arise from a single homogeneous distribution, the proposed framework conditions intent prediction on both semantic content and dataset-specific context. This design allows the model to account for differences in annotation schemes, linguistic style, and intent granularity across datasets, thereby distinguishing it from conventional transformer-based intent classification approaches that rely solely on shared textual representations [10].

Extensive experiments conducted on the unified dataset created from ATIS, SNIPS, and TOPv2, which contains 102 distinct classes, showed that HyTra-ID achieved an overall classification accuracy of 96.34% across the combined benchmark, with detailed statistical evaluation reported in the "Results" section. In addition to demonstrating how well HyTra-ID performs as a unified model using transformer methods, systematic models show the importance of dataset-aware modeling, the depth of transformers, and the ways to design classifiers as being the key factors contributing towards achieving high levels of performance.

This report is structured as follows: The section "Related Work" contains a literature review discussing the current state of research regarding intent detection, cross-domain learning, and transformer-based methods. The data sources used in this study and the data-preprocessing pipeline utilized in the processing of those data sources will be described in the section "Dataset & Preprocessing". The section "Proposed Methodology" includes a description and specification of the newly developed HyTra-ID framework. The experimental design and evaluation methodology will be discussed in the section "Experiment Setup". The section "Ablation Study" contains an overview of ablation studies conducted to evaluate several of the important architectural components of the HyTra-ID framework. The section "Results" contains information pertaining to the overall performance of the framework. The section "Discussion" discusses the important implications of the results; and section "Conclusions" summarizes this study and identifies potential avenues for future research.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. *Cureus J Comput Sci* 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

Materials And Methods

Related work

Intent detection has evolved from rule-based and shallow learning methods to deep neural and transformer-based architectures, with each stage improving the ability of dialogue systems to capture semantic context and support more accurate intent prediction. However, these advances have largely been evaluated under homogeneous training conditions, where training and test data are drawn from the same dataset or domain. As a result, while prior approaches have shown strong in-domain performance, their ability to remain reliable across heterogeneous datasets with different annotation schemes, linguistic styles, and intent granularities remains limited. This limitation is especially important in real-world dialogue systems, where utterances often span multiple domains and cannot be assumed to follow a single unified distribution. Accordingly, the key research gap is not only improving intent classification accuracy but developing models that can remain stable under cross-domain variability.

Joint Intent Detection and Semantic Modeling

Extensive research has been dedicated to developing a unified method of identifying intents while extracting slots from text since it has been demonstrated that these two systems depend heavily on one another semantically. The development of neural network architectures capable of producing common representations for both tasks improved the performance of the entire system; thus, many recent transformer-based systems have been built to learn from an integrated framework. Muhammad et al. [5] developed a unified architecture for joint intent classification and semantic parsing, reporting better performance than training the two tasks separately. Likewise, Park et al. [6] proposed a joint transformer framework with opv slot recognition within unified modeling frameworks.

Shifat et al. [12] introduced the PENTA Framework, which combines semantic information from multiple perspectives to improve robustness against linguistic variability and task-specific semantic ambiguity.

Discussion of Ablation Results

By conducting an ablation study, we have shown that the performance improvements of HyTra-ID cannot be solely explained by any one of its architectural choices; instead, they arise from the combined effect of using dataset-aware embeddings, utilizing a full transformer architecture, and applying a non-linearly activated classification layer. The most significant drop in performance (A1, dataset embedding removed) is due to the loss of explicit dataset identity, especially in regard to resolving conflicting annotations in the SNIPS dataset. The second largest drop in performance (A2, Capacity of transformer reduced) is caused by the inherent utterance variability. Paran et al. [13] created an approach to synchronize semantic cues for aligning intent relevant signals across segments of an utterance, while Du et al. [14] used a method of using representations aware of the trajectory to model the evolution of intent. Even though these approaches provide improved contextual understanding within their domains, they rely on dataset-specific semantic patterns and do not provide a solution to the problem of heterogeneity in datasets, such as ATIS, SNIPS, and TOPv2.

In addition to standard token embeddings, each utterance is augmented with an explicit dataset identifier embedding. The dataset index s_i is mapped to a learnable embedding vector $es_i \in R^d$ via a dataset embedding lookup table, where d denotes the hidden dimension of the transformer encoder. This embedding enables the model to condition its internal representations on dataset-specific characteristics such as intent granularity, linguistic style, and label imbalance.

Transformer and Large Language Model-Based Approaches

Transformer-based architectures have become a predominant approach for intent detection because they have the ability to model long-range contextual dependencies. As demonstrated by Madan et al. [1] and Li et al. [2], pretrained transformers outperform traditional Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN)-based models significantly on a wide variety of intent datasets. Furthermore, Assayed et al. [15] provided additional robustness to transformers through representation refinement techniques. On the other hand, Peng et al. [16] aimed to create more efficient deployment of transformers by focusing on lightweight transformer models.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

In terms of creating an architecture to learn the context better, there has been some recent research. DeB3RTA created by Pires et al. [10] improves the way to learn context by using better-suited attention layers, which made the quality of the learning objective better; however, the majority of transformer models are still dataset-agnostic, meaning that they will assume that the representation space of a single model will suffice for creating heterogeneous distributions of intentions.

The rise of Large Language Model (LLM) technology has broadened the range of techniques available for intent detection. In the work of Lin et al. [17], a generating and then classifying method for inferring intent through generative reasoning was presented. Asif et al. [18] and Chow et al. [19] both looked at the use of foundation models to give zero-shot and few-shot intent detections. While there is promise in using LLMs to detect intent, the dependence on carefully constructed prompts to produce accurate classification results, combined with expensive computational requirements and non-deterministic behavior, severely limits their use for large-scale, time-critical, and multi-dataset applications.

Cross-Domain, Multi-Dataset, and Open-Set Intent Detection

In recent years, the research community has focused more and more on how to effectively deal with the problem of Domain Shift and Dataset Heterogeneity. For example, Atuhurra et al. [7] conducted a study on Domain Adaptive Intent Detection through the use of a common and shared feature representation, while Ouaddi et al. [8] examined the ability of transformer models to generalize across datasets. As well, Perdana et al. [20] presented a new type of Adaptive Multi Domain Intent Classifier, and Iqbal et al. [21] investigated transformer depth optimization to make a model more domain efficient.

Detecting open-set intent is becoming increasingly popular through methods that are capable of processing previously undetected intents. The methods proposed by Dong et al. [9] provide a way to reject unknown intents; this will assist in improving the overall robustness of the intent classification systems. In addition, although some approaches to domain adaptation and open-set procedures have been developed by researchers to employ these techniques. However, many domain adaptation approaches primarily aim to reduce inter-domain discrepancy by learning more domain-invariant representations. While this can improve transferability, it may also underutilize dataset-specific characteristics that are informative for intent modeling in heterogeneous benchmark settings [7,8].

Multimodal and Human-Centric Intent Understanding

Multimodal and Human-Centered Intent Modeling has been explored in several studies beyond traditional text-based approaches. In contrast, Das et al. [22] focused on semantic and context-aware text classification for intent understanding, highlighting the importance of contextual feature optimization in improving intent prediction accuracy. The researchers introduced the SpeechVerse model to allow for Understanding of user Intent across Speech and Text. Belardinelli et al. [23] used a Focus-based approach in Gaze with an Attempt to develop Intent-based models. Hajizadeh et al. [24] proposed Types of Human-Centered Representations (HCRs) based on facial gaze analysis. Other authors, Bodonheli et al. [25] and Yee et al. [26], have conducted work to understand User Behavior Modeling.

While these methodologies help to broaden out the representation of user intents, they often require additional hardware (sensors), annotated data or interaction logs to do so, thus limiting their ability to be utilized at large scale. In addition, they do not work well with most text-based conversation systems that are utilized in multiple application domains.

Surveys, Benchmarks, and Domain-Specific Analyses

Multiple benchmark studies across different areas have examined the current state of intent detection with regard to how models are created and evaluated. Ghosh et al. [27], Pimparkhede et al. [28], and Benayas et al. [3] performed extensive empirical comparisons between existing models, while Wang et al. [4] discussed some of the less obvious challenges associated with intent detection.

How to cite this article:

Numerous studies that are domain-specific have come out with respect to Intent Detection, such as Financial Intent Detection [29] and Intent Detection in Low-Resource Settings [30], Reinforcement Learning-based Intent Modeling [31]. Integrating Domain Knowledge into Intent Detection Models [31]; and Analyzing the Intent of Social Media Users [32,33,34]. In addition, application-driven research on Intent Detection includes work by Yin et al. [35], Devi et al. [36], Zhang et al. [37], Aijaz et al. [38], Pugazhenthil et al. [39], and other studies [40,41]. Although these studies help broaden the application of Different Applications of Intent Detection, they still operate under datasets assumptions when studying the different areas of intent detection.

Comparative Analysis and Research Gap

In Table 1, you can see a comparison of some of the most relevant ways to identify intents from different sources and types of data. As you can see from the table, all current methods only focus on optimizing for one domain, lack the explicit dataset-awareness of how data are being used, and most require additional supervision and modalities. Despite recent progress in transformer-based intent detection, existing methods do not explicitly model dataset-level heterogeneity when learning from multiple benchmark sources with different annotation schemes, linguistic styles, and intent granularities. This creates a clear research gap for unified intent classification under cross-domain conditions. HyTra-ID addresses this gap by incorporating dataset-aware embeddings into a unified transformer framework.

In summary, while research to date has considerably improved intent detection capabilities, advancements in the modeling of intent across diverse datasets via a common and complete solution are still lacking. As a result, the development of the HyTra-ID framework is motivated by this challenge to develop robust methods for cross-domain intent variability using dataset-related transformer representations.

Category	Representative works	Datasets	Key limitations
Joint Intent–Slot Models	[5,6,12]	ATIS, SNIPS	Limited cross-domain robustness
Transformer-Based Models	[1,10]	SNIPS, TOPv2	Dataset-agnostic representations
Cross-Domain Adaptation	[7,8]	Multi-domain	No explicit dataset encoding
Multimodal Methods	[22,23]	Custom	High deployment complexity
Surveys & Benchmarks	[3,27]	Multiple	Identify gaps without solutions

TABLE 1: Comparison of representative intent detection approaches.

Dataset & preprocessing

The performance of intent recognition systems relies heavily on the ability of an intent detection model to generalize across different types of datasets. This can be challenging because intent detection datasets often represent multiple domains with varying intent granularities, as well as use different annotation schemes and linguistic structures. This paper attempts to create a best-practice, cross-domain intent detection benchmark to quantify and evaluate the generalizability of intent detection models across domains. A unified benchmark is created by combining data from the three most frequently used intent detection datasets ATIS, SNIPS, and TOPv2, which include examples from the following domains: Airline Reservations, Personal Assistants, and Task- Oriented Dialog Systems. A unified, real-world application-

based benchmark created by combining ATIS, SNIPS, and TOPv2 demonstrates the extent to which intention detection models perform across domains and how well they perform relative to each other. The dataset characterization and challenges faced when developing a unified cross-domain intent detection benchmark, as well as how to create a pre-processing and encoding pipeline for the framework HyTra-ID are described in this section.

Dataset Description

In this subsection, we will discuss three datasets, ATIS, SNIPS, and TOPv2, which we are using for our research. We will discuss these three benchmark datasets in detail to know how that particular dataset behaves for our model.

(1) ATIS Dataset: The ATIS Dataset [42] is a standard data set to test intent detection and spoken language processing in airline-based conversational agents. The ATIS Dataset consists of user requests related to requesting information about commercial flight reservations, flight scheduling, flight costs, airports, and airline services. In the study, 5,473 total utterances were collected from users where all were assigned to one of 26 intent classifications based on user intent. Although the ATIS dataset exhibited substantial class imbalance, no explicit resampling or class-reweighting strategy was applied. Instead, this imbalance was accounted for during evaluation by reporting weighted precision, weighted recall, and weighted F1-score, which provide a more reliable summary under skewed label distributions.

The training, validation, and test data splits contained 3,823, 819, and 831, samples respectively. ATIS utterances tend to be shorter than other types of datasets, averaging 11.37 words per ATIS utterance with a maximum of 46 words and an average character length of 63.89 characters. While the overall number of training instances in the ATIS dataset is small relative to other datasets, there are multiple intent labels associated with each ATIS utterance, as well as extreme imbalances in the number of samples for each intent label. This creates additional difficulties for models built from ATIS data because they must learn how to generalize the decision boundaries between the various intent categories.

(2) SNIPS Dataset: The personal assistant oriented SNIPS Dataset [43] includes 14,257 labeled utterances that classify into categories like music, weather, restaurants, and search. In addition to this, the SNIPS Dataset has also been shown to be a balanced set of seven intents/labels across the whole corpus of utterances.

The total 12,256 utterances are 9,979 used for training, 2,139 used for validation, and 2,139 used for testing. SNIPS utterances are linguistically clean and quite short in terms of character and word count; the average SNIPS utterances contain 9.14 words and 46.34 characters. Because of the balanced nature of the data and the unique intent semantics of each utterance in SNIPS, many natural language processing systems have achieved good baseline performance; however, the structural simplicity of SNIPS is very different from the complexity of other datasets such as TOPv2, which has created difficulty in cross dataset generalization.

(3) TOPv2 Dataset: The TOPv2 dataset [44] was created for the purpose of supporting dialogue systems that perform various functions in a variety of tasks. The dataset consists of 150,205 utterances across 69 different intents and eight domains (navigation, alarms, reminders, weather, music, messaging, events, and timers).

The dataset consists of 105,147 training examples, 22,529 validation examples, and 22,529 test examples. Linguistically, TOPv2 presents a high degree of diversity on both an intent and syntactic level with an average utterance having approximately 7.72 words and a maximum of 51 words per utterance. Additionally, there is a significant long-tail distribution within TOPv2, which contains multiple very rare intents; thus, this dataset causes difficulties in training a single-intent model due to the imbalance within the dataset itself.

Challenges in Cross-Dataset Unification

To reduce redundancy and prevent data leakage, duplicate and empty samples were removed during preprocessing before dataset splitting. Each dataset was cleaned independently and then divided into train, validation, and test partitions using stratified splitting so that label distributions were preserved while preventing overlap of duplicated samples across partitions. Furthermore, the cleaned splits from ATIS, SNIPS, and TOPv2 were merged into the unified benchmark.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. *Cureus J Comput Sci* 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

In addition, the differences in utterance lengths and vocabulary by domain lead to linguistic biases specific to each dataset. As such, traditional transformer-based intent detectors which were trained on a large, aggregated dataset will tend to overfit the most dominant datasets and not be able to generalize well to datasets with less representation. These issues highlight the need for developing dataset-aware modeling approaches that include dataset identity information throughout the training process, as in HyTra-ID.

Preprocessing Pipeline

In order to keep our datasets consistent, we created a single preprocessing method to be used on all datasets before training them on a model. This process included four steps: text normalization, duplication removal, intent label sanitization, and harmonizing the datasets. All utterances were converted to lower case letters, and any unnecessary characters were removed as well as misshaped samples from the data set were filtered. To construct the unified intent space, the intent labels from ATIS, SNIPS, and TOPv2 were collected into a single global label vocabulary after preprocessing and label standardization. Each unique intent label was assigned a unified label index, resulting in 102 intent classes across the merged benchmark. This approach allowed all utterances to be represented within one common classification space while preserving the dataset-specific intent definitions.

Utterance length analysis showed that a maximum sequence length of 64 tokens covers more than 99% of the samples in the unified benchmark. This value was therefore selected as a practical trade-off between preserving nearly all input content and limiting computational cost. The overall preprocessing pipeline is illustrated in Figure 1.

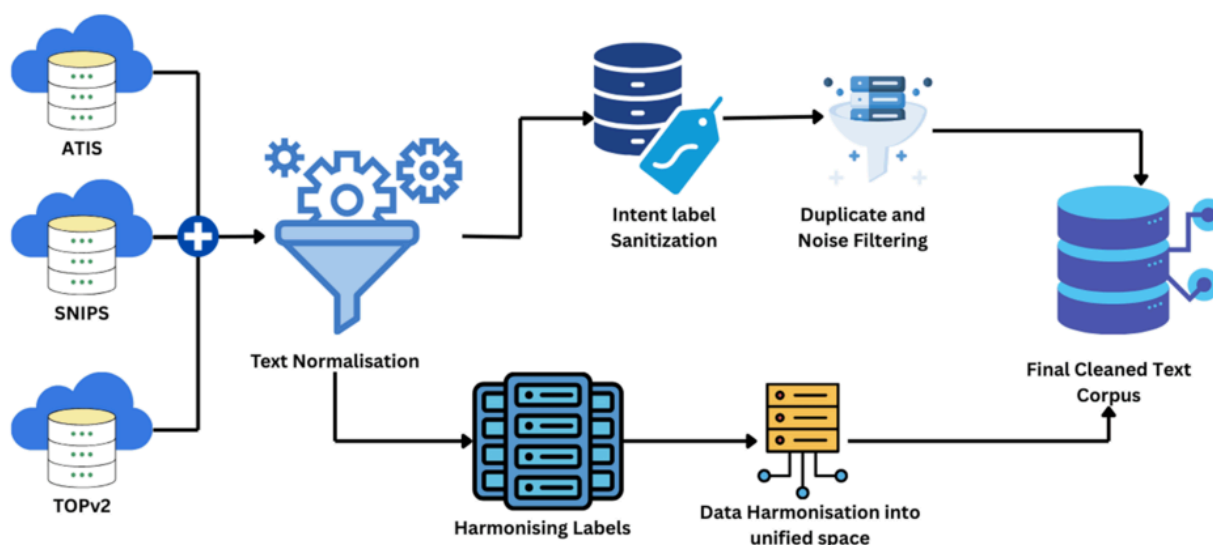


FIGURE 1: Unified preprocessing pipeline applied to ATIS, SNIPS, and TOPv2 datasets.

Encoding Pipeline

By processing data before using it in a machine learning (ML) model, ML models can convert and simplify the way text from different sources is represented. For example, all texts processed into ML models will be encoded in the same way, while maintaining an explicit record of the source of the texts, so that any variability in the intent for each text can easily be determined because it is also clearly indicated in the source.

$D = (x_i, y_i, s_i)$ for $i = 1$ to N denote the unified corpus, where x_i is the i -th utterance, y_i is its intent label, and $s_i \in \{1, \dots, S\}$ represents the dataset identifier corresponding to ATIS, SNIPS, or TOPv2. Before entering any inputs, they will be split into tokens with the help of a pretrained tokenizer BERT-base-uncased, which uses both WordPiece

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

segmentation and the special tokens [CLS] and [SEP]. All tokenized inputs are then either padded or truncated to $L = 64$, which is defined as the empirical's maximum length; this means that all token sequences must have the same dimensions, so that the end result will be able to pass through the model without issue.

In addition to standard token embeddings, each utterance is augmented with an explicit dataset identifier embedding. The dataset index s_i is mapped to a learnable embedding vector $\mathbf{e}_{s_i} \in \mathbb{R}^d$ via a dataset embedding lookup table, where d denotes the hidden dimension of the transformer encoder. This embedding enables the model to condition its internal representations on dataset-specific characteristics such as intent granularity, linguistic style, and label imbalance.

The complete dataset-aware encoding procedure is formally summarized in Algorithm 1.

Algorithm 1: Dataset-Aware Encoding Pipeline

Require: Preprocessed corpus $D = (\mathbf{x}_i, \mathbf{y}_i, s_i)$ for $i = 1$ to N , tokenizer T , maximum sequence length L

Ensure: Encoded dataset $E = \{(\mathbf{t}_i, \mathbf{m}_i, \mathbf{d}_i, \mathbf{y}_i)\}$ for $i = 1$ to N

```

1: for each utterance  $(\mathbf{x}_i, \mathbf{y}_i, s_i)$  in  $D$  do
2:   Tokenize utterance:  $\mathbf{t}_i \leftarrow T(\mathbf{x}_i)$ 
3:   Insert special tokens [CLS], [SEP]
4:   Pad or truncate  $\mathbf{t}_i$  to length  $L$ 
5:   Generate attention mask  $\mathbf{m}_i$  in  $\{0, 1\}$ 
6:   Map dataset identifier  $s_i$  to dataset index  $\mathbf{d}_i$  in  $\{1, \dots, S\}$ 
7:   Store encoded tuple  $(\mathbf{t}_i, \mathbf{m}_i, \mathbf{d}_i, \mathbf{y}_i)$ 
8: end for
9: return Encoded dataset  $E$ 

```

Here, $\mathbf{t}_i$ in N^L denotes the token-id sequence of fixed length L , $\mathbf{m}_i$ in $0, 1^L$ is the corresponding attention mask, $\mathbf{d}_i$ is the dataset index associated with the source dataset of the utterance, and $\mathbf{y}_i$ is the target intent label. During model training, the token sequence and attention mask are passed through the transformer encoder to produce a contextual representation $\mathbf{h}_i^{CLS}$ in $\mathbb{R}^H$, where H is the hidden size of the encoder. The dataset index $\mathbf{d}_i$ is then mapped to a trainable dataset embedding vector in $\mathbb{R}^H$, and the two H -dimensional vectors are concatenated to form a fused representation in $\mathbb{R}(2H)$ for downstream classification.

The dataset-aware encoding ability reduces the impact of negative transfer and improves performance on a variety of heterogeneous intent-detection test sets.

Unified Dataset Statistics and Splits

Table 2 shows the statistics for each dataset and the combined intent space for all datasets in this benchmark. The final benchmark contains a total of 169,935 utterances with 102 distinct intent labels, making it an extensive and challenging dataset for testing methods for cross-domain intent detection.

How to cite this article:

Dataset	Train	Validation	Test	Total samples	Intents
ATIS	3,823	819	831	5,473	26
SNIPS	9,979	2,139	2,139	14,257	7
TOPv2	105,147	22,529	22,529	150,205	69
Unified	118,949	25,487	25,499	169,935	102

TABLE 2: Dataset statistics and unified intent space.

The scale, diversity, and unevenness of the unified benchmark provides a rich and extensive resource for evaluating the strength and breadth of the HyTra-ID across many real-world cross-domain deployment situations.

Proposed methodology

This segment outlines the suggested HyTra-ID Framework, which tackles the challenges of cross-domain intent variations for Dialogue systems in the Real World. HyTra-ID also provides a Transformer based on datasets that utilizes Domain-Specific distribution difference modeling and the Pretrained Language Models (PTLM) for maintaining robust contextual representations. An illustration of the proposed architecture can be found in Figure 2 as well.

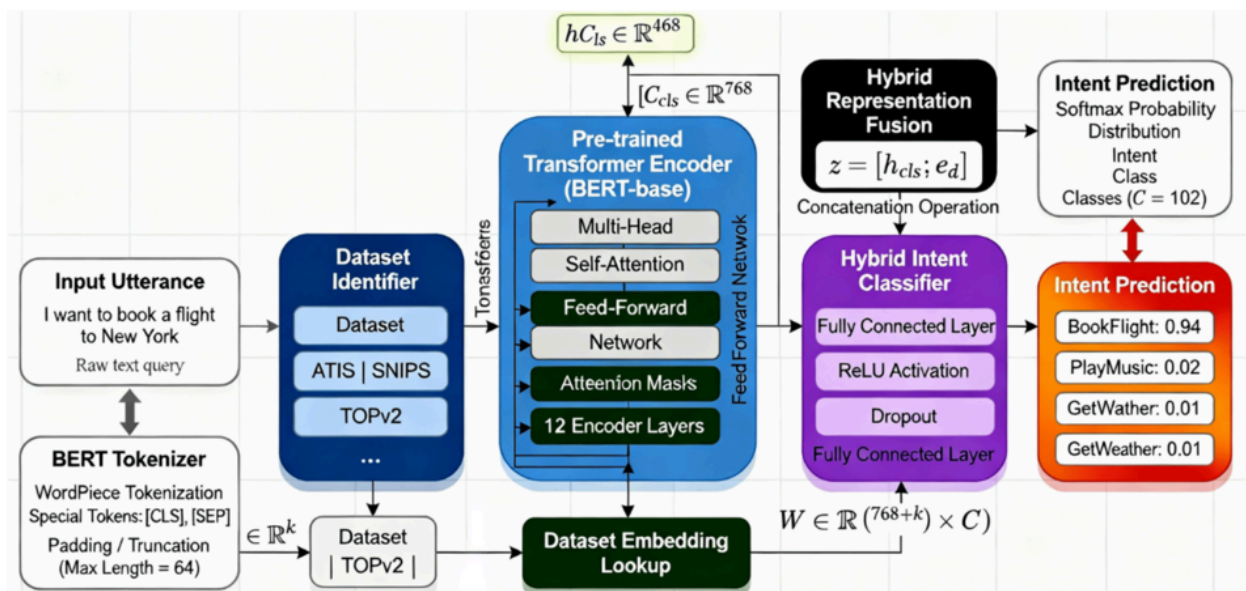


FIGURE 2: Overall architecture of the proposed HyTra-ID framework.

Dataset-aware embeddings are fused with contextual transformer representations to enable robust cross-domain intent detection. HyTra-ID, Heterogeneity Aware Transformer for Intent Detection

Problem Formulation

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

Let $D = (\mathbf{x}_i, \mathbf{y}_i, \mathbf{s}_i)$ for $i = 1$ to N denote the multi-domain intent detection dataset, where $\mathbf{x}_i$ is the input utterance, $\mathbf{y}_i \in \mathbf{Y}$ is the corresponding intent label, and $\mathbf{s}_i \in \mathbf{S}$ denotes the dataset identifier (ATIS, SNIPS, or TOPv2). The objective is to learn a mapping function $f_{\theta}(\mathbf{x}_i, \mathbf{s}_i) \rightarrow \mathbf{y}_i$, where θ denotes the model parameters.

$$f_{\theta}(\mathbf{x}_i, \mathbf{s}_i) \rightarrow \mathbf{y}_i \quad (1)$$

where $\mathbf{x}_i$ represents the input utterance, $\mathbf{s}_i$ denotes the dataset identifier, $\mathbf{y}_i$ is the predicted intent label, and θ represents the model parameters.

where θ represents the model parameters are represented in such a way that the method's robust intent prediction remains intact over heterogeneous datasets with varying intent granularities, linguistic styles, and class distribution; and $\mathbf{d}_{s_i}$ represents the dataset-specific embedding vector and d denotes the hidden dimension of the transformer.

Overall Architecture of HyTra-ID

A summary of HyTra-ID architecture is shown in Figure 2, including the following four major components: (I) token embeddings and contextual embeddings, (II) dataset aware embedded token fusion, (III) hybrid classifier heads, and (IV) intent classification layer.

Contextual Token Encoding

Each utterance $\mathbf{x}_i$ is tokenized using the BERT-base-uncased tokenizer and converted into a sequence of token embeddings:

$$\mathbf{E}_i = [e_{[CLS]}, e_1, \dots, e_L] \quad (2)$$

where $\mathbf{E}_i$ represents the token embedding sequence, $e_{[CLS]}$ is the special classification token, and L denotes the maximum sequence length.

where L denotes the maximum fixed sequence length. The embeddings are then passed to a multi-layer transformer encoder made up of K self-attention layers. In layer k , the hidden representation is revised as follows:

$$\mathbf{H}(k) = \text{TransformerLayer}(k)(\mathbf{H}(k-1)) \quad (3)$$

where $\mathbf{H}(k)$ is the hidden representation at layer k and TransformerLayer denotes the transformation applied at each encoder layer.

with $\mathbf{H}(0) = \mathbf{E}_i$.

The final contextual representation corresponding to the [CLS] token is defined as

$$\mathbf{h}_i^{CLS} = \mathbf{H}_0(K) \quad (4)$$

where $\mathbf{h}_i^{CLS}$ represents the final contextual embedding corresponding to the [CLS] token and serves as a global semantic summary of the utterance.

Dataset-Aware Embedding Fusion

HyTra-ID represents an innovative approach to encoding information about the identity of datasets. A dataset will be represented as a learnable embedding vector:

$$\mathbf{d}_{s_i} \in \mathbb{R}^d \quad (5)$$

How to cite this article:

For each utterance x_i with dataset identity s_i , the transformer encoder produces a contextual representation h_i^{CLS} from the [CLS] token, while a trainable embedding d_{s_i} is used to encode dataset-specific information. These two representations are fused by concatenation to form the final feature vector used for intent classification:

$$z_i = h_i^{CLS} || d_{s_i} \quad (6)$$

where d_{s_i} represents the dataset-specific embedding and d denotes the hidden dimension.

In this study, concatenation was adopted as a simple and effective fusion mechanism for combining contextual and dataset-specific information. While this design performed strongly in the present benchmark, alternative fusion strategies, such as additive fusion, gated fusion, or attention-based fusion, may also be explored in future work to further investigate how dataset-aware signals can be integrated more adaptively.

Hybrid Classification Head

The term hybrid in HyTra-ID refers to the integration of contextual transformer representations with dataset-aware embeddings, rather than to the use of a specialized hybrid classifier structure.

The fused representation z_i is passed through a hybrid classification head composed of stacked fully connected layers:

$$u_i = \sigma(W_1 z_i + b_1) \quad (7)$$

$$o_i = W_2 u_i + b_2 \quad (8)$$

where W_1 and W_2 are weight matrices, b_1 and b_2 are bias terms, and σ represents the ReLU activation function.

Intent Prediction and Loss Function

The final intent probabilities are computed using a softmax function:

$$P(y_i = c | x_i, s_i) = \dots \quad (9)$$

where $o_{i,c}$ is the output score for class c .

The model is trained by minimizing the cross-entropy loss:

$$L = -\sum \log P(y_i | x_i, s_i) \quad (10)$$

where L represents the loss function used to optimize the model, which encourages correct intent classification while implicitly learning dataset-sensitive decision boundaries.

Computational Efficiency and Design Rationale

The design of HyTra-ID aims to improve cross-domain intent classification while keeping the architectural extension lightweight. In particular, the dataset-aware embedding introduces only a small additional component on top of the transformer encoder, and the ablation variants illustrate how different architectural simplifications affect predictive performance. These results suggest that HyTra-ID provides a practical unified framework for intent classification across heterogeneous dialogue datasets, although detailed latency and memory benchmarking remain important directions for future work.

Experiment setup

The purpose of this section is to outline the approach taken to develop and implement an evaluative experimental design that was both transparent and replicable for testing the HyTra-ID framework. Here, we provide specifications for all aspects of our research, including: training environment, model specifications, optimization procedure, and evaluation metrics and protocols that were employed on the different types of dataset used and in the different ablation studies conducted for each dataset.

How to cite this article:

Experimental Environment

To improve interpretability of the later ablation results, we explicitly summarize the key characteristics of the unified benchmark in terms of dataset composition, intent-space diversity, class imbalance, and inter-dataset variability. These properties are important because they directly influence the difficulty of unified cross-domain intent modeling.

Model Configuration

A BERT transformer encoder is the main element for processing data in all experiments. The BERT classifier is built using the BERT-base-uncased with a total of 12 encoder layers, 12 head-size self-attention heads per layer, and a 768-dimensional hidden representation. Although the maximum permitted input sequence length was set to 64 tokens across all datasets, this was based on data distribution analysis of dataset lengths.

To enhance the performance of the HyTra-ID architecture, a hybrid contextual representation was created by combining the contextual representation of the [CLS] token with a set of trainable dataset embeddings for each of the respective source datasets (ATIS, SNIPS, TOPv2). A multi-layer classifier predicted the intent based on the hybrid contextual representation. For conducting ablation studies, the experimental group contained only the individual components (dataset embeddings, transformer depth, or classifier complexity) that had been isolated from each other but maintained constant the remaining system configurations.

Training Protocol

The AdamW Optimiser was used for all of the Models along with an Initial Learning Rate of 2×10^{-5} and Weight Decay of 0.01. Batch Size was set at 32 for Training/Validation/Testing. Training and preventing exploding Gradients were accomplished through Gradient Clipping at 1.0.

The three-epoch training schedule was selected based on repeated pilot experiments, which showed that the validation performance stabilized by the third epoch and that further training did not produce consistent improvement. This indicates that the model had effectively reached a stable local optimum under the chosen optimization setting. To maintain consistency across the ablation experiments, the same training schedule was applied to all models. Early stopping was not used so that each variant could be compared under an identical training protocol. A dropout rate of 0.3 was applied in the classifier head to reduce overfitting.

Evaluation Metrics

Standard classification metrics were used to evaluate the performance of the models on the combined test set (accuracy, weighted precision, recall, and F1-score) while taking into consideration the class imbalance in the datasets (ATIS and TOPv2), so that weighted metrics could provide more accurate measures of overall performance.

In addition to the aggregate evaluation of the combined test set, the datasets were assessed separately for each of the datasets (ATIS, SNIPS, and TOPv2) to provide additional insights into the cross-domain generalization performance and robustness of the models on specific datasets.

Experimental Protocol and Baselines

To ensure a fair and controlled evaluation, all experiments followed a unified protocol:

- Identical train, validation, and test splits were used across all models.
- The same tokenizer, maximum sequence length, optimizer, and training schedule were applied.
- Hyperparameters were fixed across experiments and not tuned separately for individual datasets.

The experimental evaluation consisted of the following configurations:

- **Ablation-1:** Transformer-based intent classifier without dataset embeddings.
- **Ablation-2:** Reduced transformer depth via partial layer freezing.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. *Cureus J Comput Sci* 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

- **Ablation-3:** Shallow classification head without inter-mediate nonlinear transformations.
- **Proposed HyTra-ID:** Full dataset-aware hybrid trans-former architecture.

Through a systematic approach to isolating architectural contributions, this study presents a strong empirical foundation for the benefit of dataset-informed modeling for cross-domain intent classification.

Reproducibility

Every aspect of collecting and transforming the data, developing an encoding system for the data, creating a model, and training the model was done in a modular and repeatable way, from the start of the process through to the end of the process.

Fixed random seed values were used to separate datasets and initialize models. The full set of configuration files and the trained models will be made available publicly to help other researchers replicate these efforts and expand upon them as needed.

Dataset	Metric	A1	A2	A3	A4	HyTra-ID
ATIS	Accuracy	0.9495	0.9374	0.9603	0.9639	0.9639
	Precision (W)	0.9396	0.9201	0.9540	0.9583	0.9520
	Recall (W)	0.9495	0.9374	0.9603	0.9639	0.9639
	F1-score (W)	0.9423	0.9260	0.9557	0.9605	0.9565
SNIPS	Accuracy	0.9374	0.9317	0.9603	0.9411	0.9874
	Precision (W)	0.9928	0.9905	0.9901	0.9906	0.9873
	Recall (W)	0.9374	0.9317	0.9603	0.9411	0.9874
	F1-score (W)	0.9631	0.9587	0.9743	0.9639	0.9873
TOPv2	Accuracy	0.9551	0.9529	0.9507	0.9544	0.9612
	Precision (W)	0.9571	0.9558	0.9572	0.9576	0.9597
	Recall (W)	0.9551	0.9529	0.9507	0.9544	0.9612
	F1-score (W)	0.9554	0.9532	0.9531	0.9546	0.9600

TABLE 3: Dataset-wise ablation comparison of HyTra-ID and its variants.

HyTra-ID, Heterogeneity Aware Transformer for Intent Detection

How to cite this article:

Ablation study

The results shown here demonstrate how much influence each component of the proposed HyTra-ID framework contributes to the overall performance of the model through a systematic ablation analysis. We therefore report both individual dataset results and combined dataset results to evaluate performance across diverse intent distributions and heterogeneous domains. All models were tested under the same experimental protocol to ensure fair comparison. In addition, statistical validation was incorporated to assess whether the observed performance differences are meaningful, including confidence-based uncertainty analysis and comparative significance evaluation across the ablation variants.

Dataset-Wise Ablation Analysis

Table 3 summarizes the performance of ATIS, SNIPS, and TOPv2 using accuracy and weighted metrics at the dataset level. By decomposing the results in this way, we can examine how each model variant behaves under different dataset conditions, particularly with respect to imbalance, intent-space diversity, and domain-specific variation, while also reducing dominance effects from the much larger datasets.

Overall Performance Comparison

For a comprehensive assessment of overall system behavior, Table 4 reports the aggregated results on the combined test set. These results should be interpreted together with the dataset characteristics described earlier, since the unified benchmark reflects substantial heterogeneity in class balance, dataset scale, and intent structure.

Model	Accuracy	Precision (W)	Recall (W)	F1-score (W)
A1	0.9534	0.9524	0.9534	0.9521
A2	0.9506	0.9482	0.9506	0.9483
A3	0.9518	0.9513	0.9518	0.9508
A4	0.9536	0.9522	0.9536	0.9516
HyTra-ID	0.9634	0.9618	0.9634	0.9622

TABLE 4: Overall ablation comparison on the combined test set.

HyTra-ID, Heterogeneity Aware Transformer for Intent Detection

Figure 3 presents a comparative visualization of overall metrics, highlighting the consistent margin achieved by HyTra-ID.

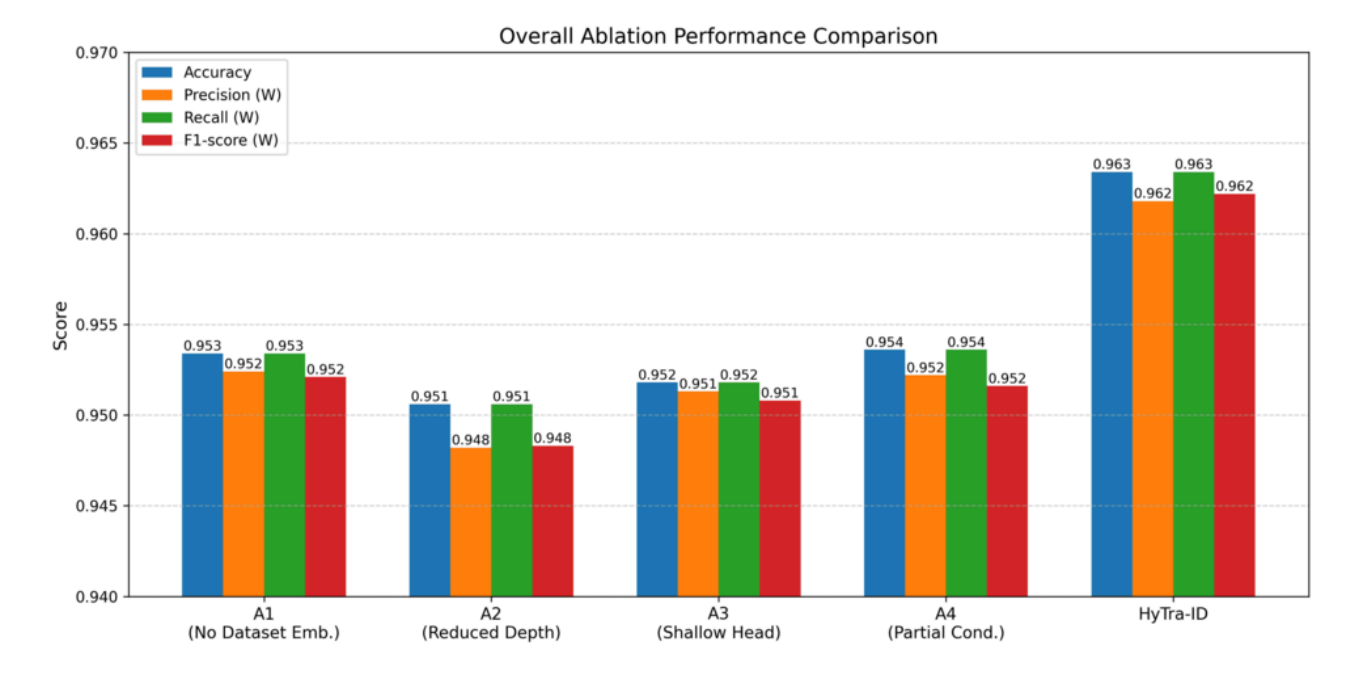


FIGURE 3: Overall performance comparison across ablation variants.

Discussion of Ablation Results

The ablation results suggest that the performance improvements of HyTra-ID are not attributable to a single architectural component alone but instead reflect the combined contribution of dataset-aware embeddings, full transformer capacity, and a non-linear classification head. The largest performance reduction is observed in A1 (without dataset embedding), indicating that explicit dataset identity likely plays an important role in the unified benchmark setting. Likewise, the weaker performance of A2 suggests that reduced transformer capacity limits the model’s ability to capture contextual variation across datasets. The performance drop in A3 further suggests that a shallower classifier may be less effective in separating intent boundaries under heterogeneous domain conditions.

Although using the added training dataset and similar DASH, recorded data in real time will show only moderate performance improvement and you will see significant differences between performance on datasets, DASH is required to give any kind of realistic performance estimate. The performance of the full HyTra-ID model is 96.34%, indicating that it has good robustness and would work well in actual environments with a variety of applications.

Results And Discussion

Results

The quantitative study conducted on the proposed HyTra-ID framework revealed many interesting findings. These findings were examined from four complementary perspectives, which are as follows: (1) the aggregate performance and datasets, (2) errors based on the Confusion Matrix, (3) the Class-wise performance for all Categories of Intent, and (4) a comparison of our method with Representative State-of-the-Art (SOTA) Intent Detection Methods to see how our method compares to other methods regarding performance, robustness, generalization, and scalability in Cross-Domain Dialogue Systems in the real world.

Overall and Dataset-Wise Performance

HyTra-ID achieved an overall accuracy of 96.34% on the unified benchmark. Table 5 presents the overall and dataset-wise performance across ATIS, SNIPS, and TOPv2, showing that performance varies across datasets with different class distributions, intent granularities, and levels of imbalance. In particular, the stronger performance on SNIPS and the comparatively more challenging behavior on ATIS and TOPv2 reflect differences in dataset complexity and heterogeneity. These results demonstrate strong performance on the combined benchmark, although broader claims regarding generalization beyond the evaluated datasets require further validation through unseen-domain or out-of-distribution testing.

Evaluation scope	Accuracy (%)	Precision (Weighted)	Recall (Weighted)	F1-score (Weighted)
Overall (all datasets)	96.34	96.18	96.34	96.22
ATIS	96.39	95.20	96.39	95.65
SNIPS	98.74	98.73	98.74	98.73
TOPv2	96.12	95.97	96.12	96.00

TABLE 5: Overall and dataset-wise performance of HyTra-ID.
HyTra-ID, Heterogeneity Aware Transformer for Intent Detection

There are three major findings from the Results section: First, the near-perfect accuracy on SNIPS provides evidence of good discrimination among balanced intent datasets. Second, strong scores on TOPv2 support our claim of scalability to large, multi-domain environments with 69 intents and very large class imbalances. Third, stable accuracy results for ATIS are indicative of robustness even when extreme label composition and low-frequency intent labels exist. The findings validate the ability of HyTra-ID’s datasets-based models to achieve good accuracies.

B. Confusion Matrix Analysis

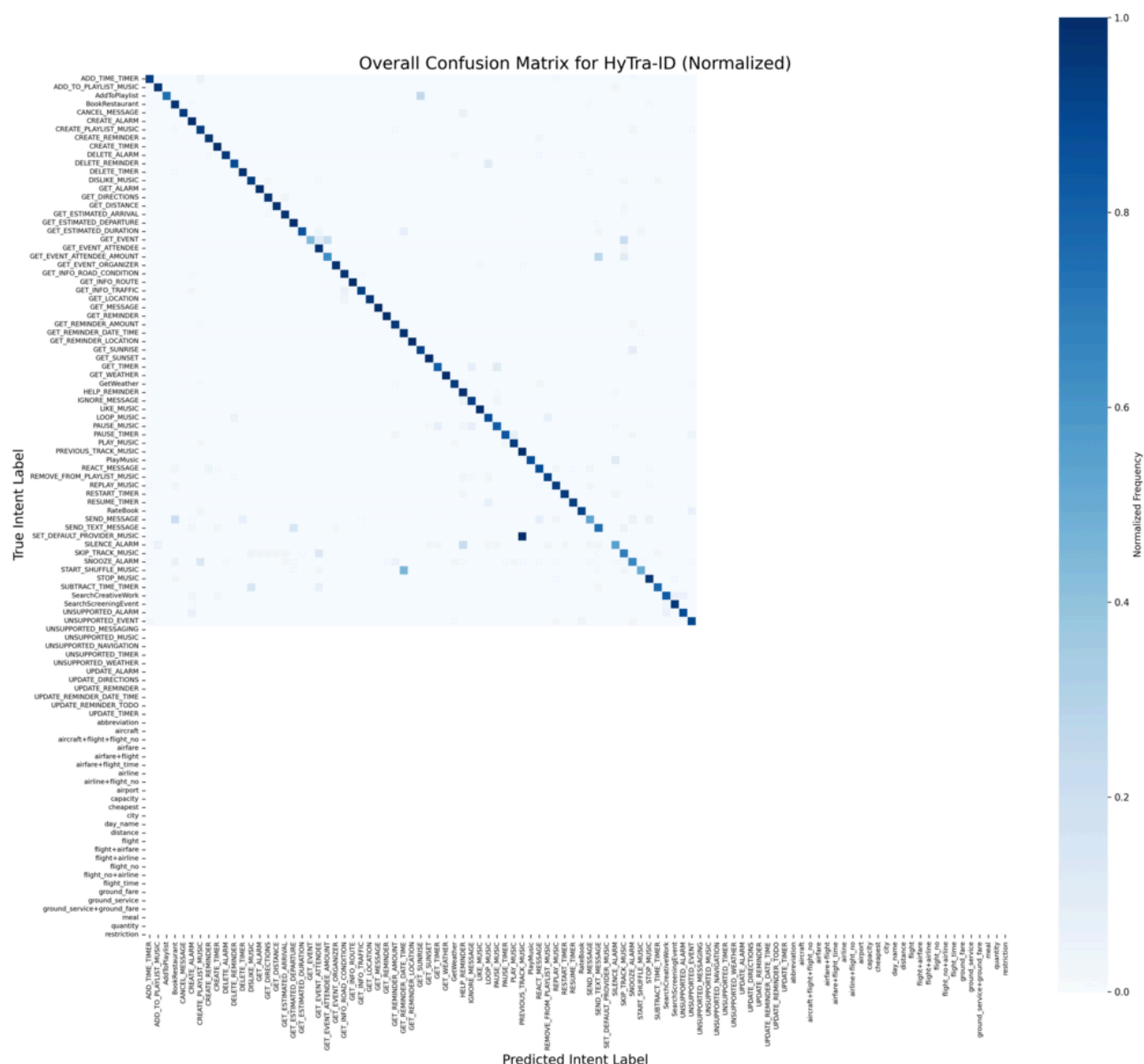


FIGURE 4: Overall confusion matrix across all datasets.

A confusion matrix will be used to assess and evaluate predictive accuracy within both the overall test datasets and each of the individual test datasets so that researchers can comprehend how the system's predictions are behaving. The overall confusion matrix, as seen in Figure 4, shows a great deal of diagonal dominance; this indicates that predictions for different intents are experiencing a very low degree of cross-intent confusion across over one hundred different intent types.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

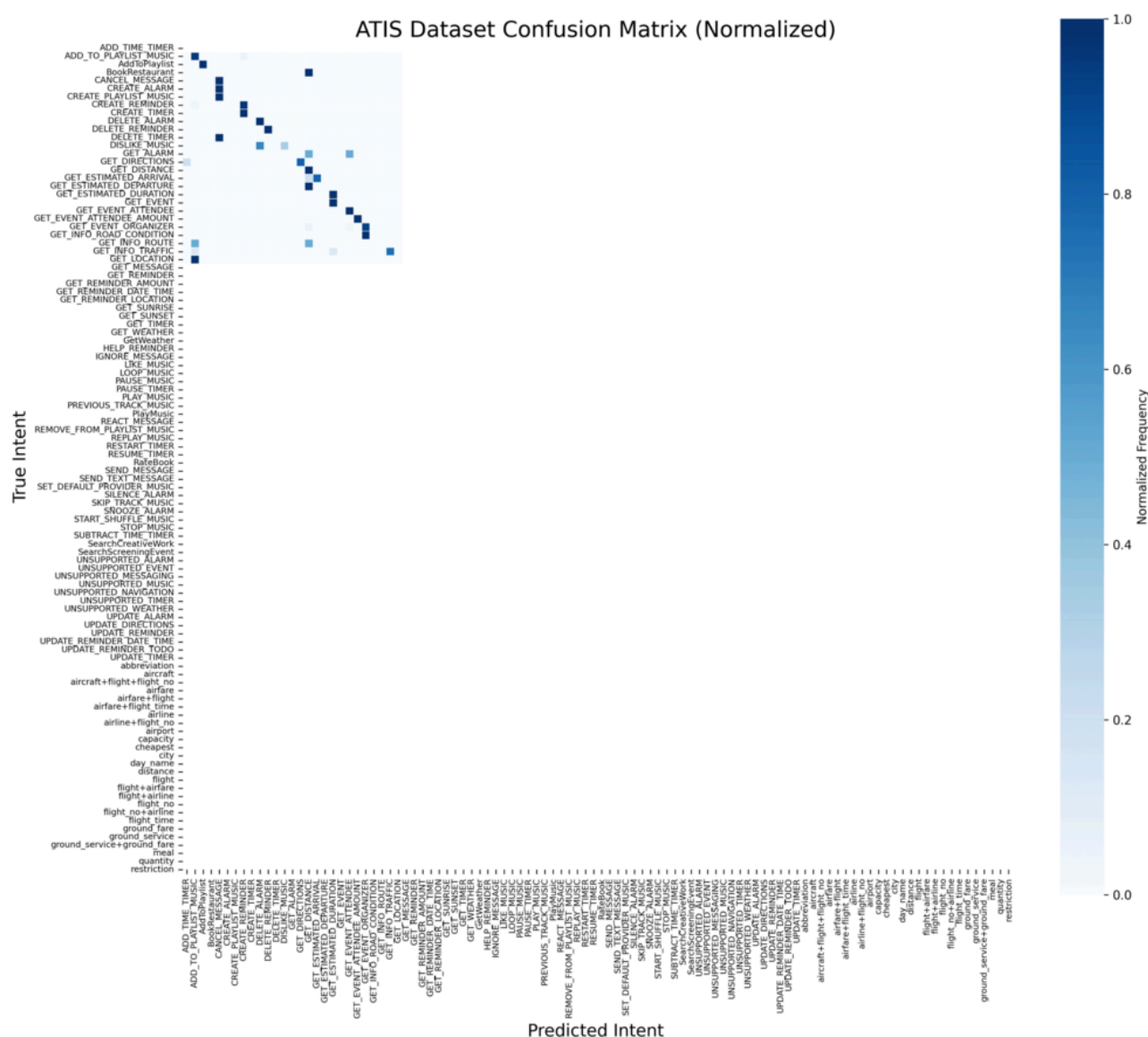


FIGURE 5: Confusion matrix for ATIS dataset.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

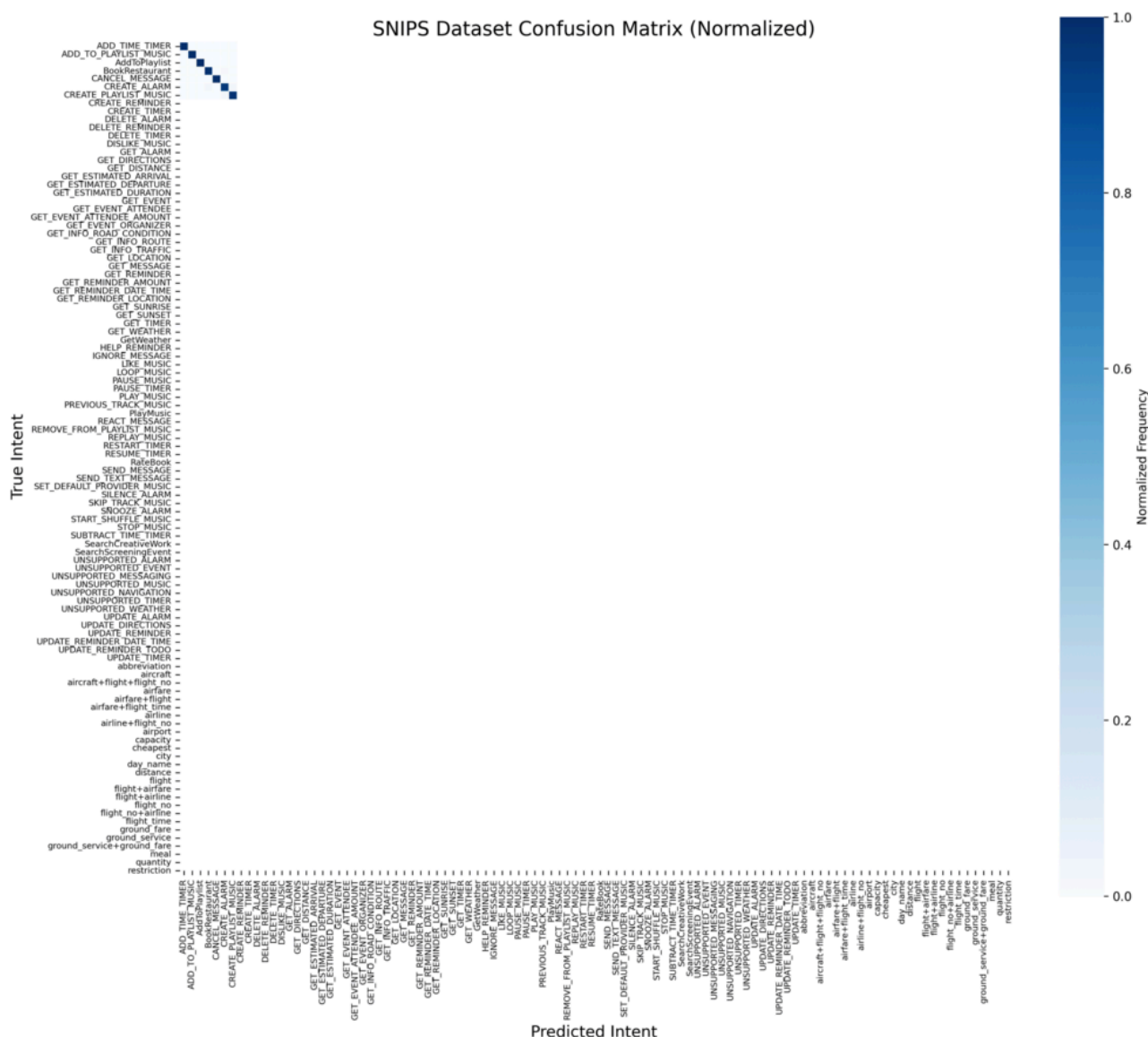


FIGURE 6: Confusion matrix for SNIPS dataset.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

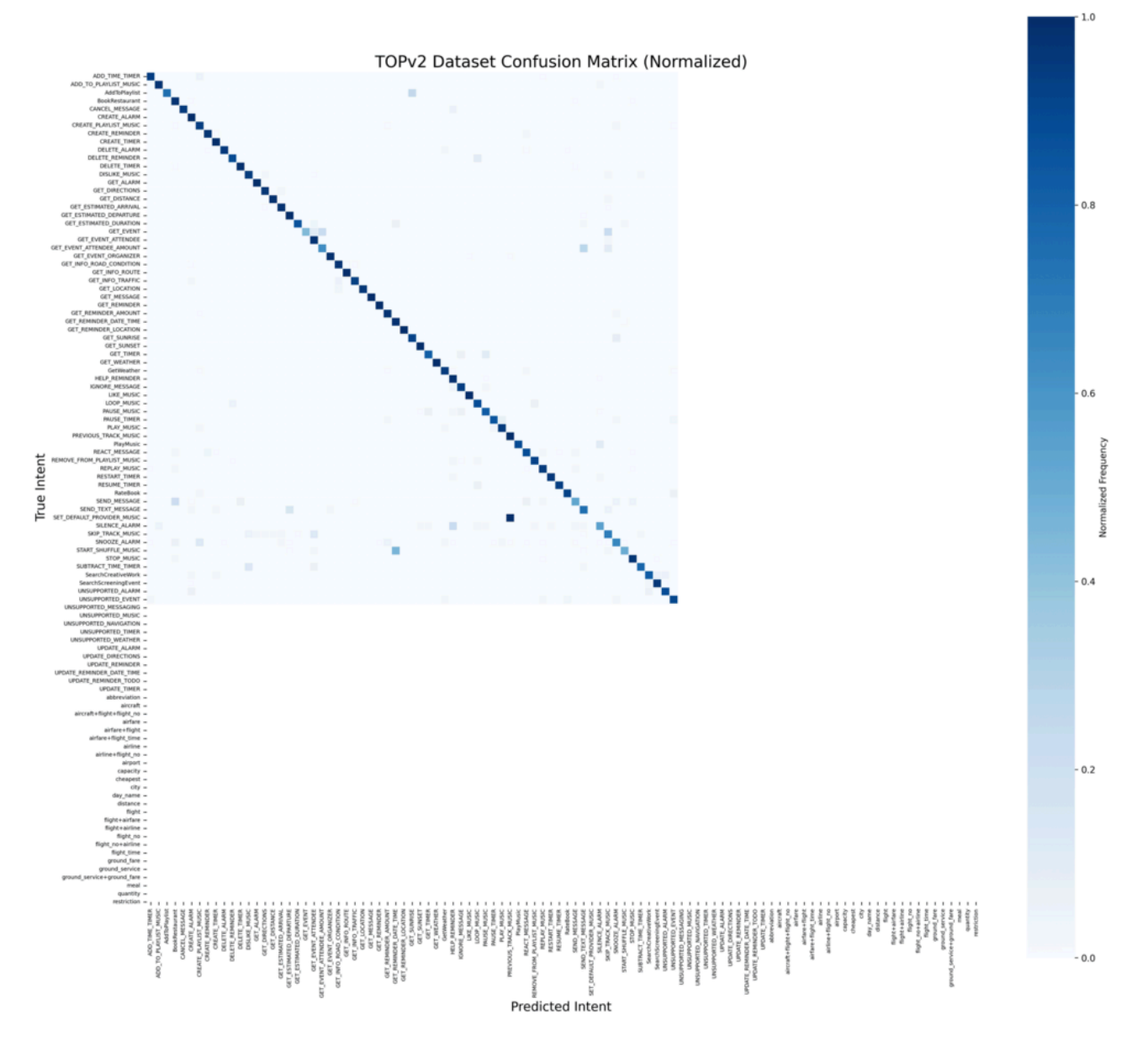


FIGURE 7: Confusion matrix for TOPv2 dataset.

The confusion matrices for each dataset can be found in Figure 5, Figure 6, and Figure 7, respectively. The ATIS dataset appears to exhibit a pattern of misclassified examples by populating errors around flights and airfares classified as separate categories due to their semantic inconsistency as opposed to error during the implementation of the model.

The SNIPS dataset seems to separate classes very well, as shown by almost no examples being misclassified and the vast majority of misclassified examples being in close proximity to the correct classes on the confusion matrix. The TOPv2 dataset contains a greater proportion of errors between semantically similar classes; however, it appears that this occurs predominantly between navigation and reminder or timer intent.

Overall, these matrices confirm that HyTra-ID maintains consistent decision boundaries across datasets with varying scale, complexity, and intent granularity.

Class-Wise Performance Analysis

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

Table 6 reports class-wise precision, recall, F1-score, and support, allowing performance to be interpreted in the context of class frequency. The stronger results observed for several high-support intent categories are consistent with their larger representation in the unified benchmark, whereas lower-support and structurally rarer intent classes are naturally more challenging under heterogeneous cross-domain conditions. These class-wise differences should therefore be interpreted together with the imbalance and inter-dataset variability described earlier in the section "Data & Preprocessing" and Table 2, As uniform evidence of equivalent model behavior across all intent categories.

HyTra-ID performance is adversely affected when training data contains unsupported intents (support size 5) as in the case of the intent UNSUPPORTED_MESSAGING. This is an expected result in supervised environments and consistent with what the literature indicates. Additionally, when examining aggregate or weighted overall metrics, it becomes clear that rare analogue class distributions do not negatively impact on HyTra-ID output, demonstrating HyTra-ID's reliable performance despite imbalanced real-world intent distributions.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

Intent class	Precision	Recall	F1-score	Support
GET_WEATHER	0.97	0.99	0.98	3374
CREATE_ALARM	0.97	0.98	0.98	2144
CREATE_REMINDER	0.99	0.99	0.99	2077
GET_INFO_TRAFFIC	0.97	0.98	0.97	1829
GET_EVENT	0.99	0.99	0.99	1825
SEND_MESSAGE	1.00	0.99	0.99	1318
PLAY_MUSIC	0.98	0.98	0.98	1282
GET_ESTIMATED_DURATION	0.97	0.98	0.97	877
DELETE_REMINDER	0.99	1.00	0.99	616
CREATE_TIMER	0.92	0.94	0.93	538
GET_REMINDER	0.98	0.97	0.98	512
GET_DIRECTIONS	0.93	0.94	0.94	405
GET_ALARM	0.98	0.98	0.98	407
DELETE_ALARM	0.98	0.97	0.97	382
ADD_TIME_TIMER	0.97	0.94	0.95	135
ADD_TO_PLAYLIST_MUSIC	0.91	0.97	0.94	33
DELETE_TIMER	0.98	0.95	0.97	103
GET_DISTANCE	0.98	0.99	0.99	368
GET_MESSAGE	0.99	0.99	0.99	226
GET_TIMER	0.93	0.97	0.95	294
PAUSE_TIMER	0.93	0.95	0.94	116
REACT_MESSAGE	0.99	1.00	1.00	131
SNOOZE_ALARM	0.94	0.95	0.94	172

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

SILENCE_ALARM	0.87	0.88	0.87	168
SKIP_TRACK_MUSIC	0.93	0.92	0.93	92
START_SHUFFLE_MUSIC	0.97	0.95	0.96	41
STOP_MUSIC	0.97	0.94	0.95	31
UNSUPPORTED_ALARM	0.67	0.53	0.59	91
UNSUPPORTED_EVENT	0.70	0.75	0.72	64
UNSUPPORTED_MUSIC	0.72	0.55	0.62	60
UNSUPPORTED_NAVIGATION	0.82	0.70	0.75	325
UNSUPPORTED_TIMER	0.81	0.66	0.73	244
UNSUPPORTED_WEATHER	0.68	0.49	0.57	174
CANCEL_MESSAGE	1.00	0.75	0.86	4
GET_INFO_ROUTE	0.80	0.44	0.57	9
GET_LOCATION	0.70	0.64	0.67	11

TABLE 6: Class-wise performance of HyTra-ID on high-frequency intent classes.

HyTra-ID, Heterogeneity Aware Transformer for Intent Detection

Comparison with SOTA Methods

The last section of the research compares and contrasts the results of HyTra-ID experiments with the current SOTA intent detection algorithms across classical, attention, and recently proposed domain-aware models. Comparable accuracy and F1 scores have been tabulated in Table 7 for each of these categories of models.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

Method	Accuracy (%)	F1-score (%)
GFIDF [45]	74.61	72.06
SDIF-DA [45]	72.36	68.29
RNN-LSTM [46]	94.60	89.40
BiRNN + Attention [47]	94.40	95.78
BiLSTM + Attention [48]	95.70	95.60
HyTra-ID (Proposed)	96.34	96.22

TABLE 7: Comparison with state-of-the-art intent detection methods.

HyTra-ID, Heterogeneity Aware Transformer for Intent Detection; RNN, Recurrent Neural Network; LSTM, Long Short-Term Memory; BiLSTM, Bidirectional Long Short-Term Memory; GFIDF, Gradual Fusion Intent Detection Framework; SIDF, Shallow-to-Deep Interaction Framework

Table 7 shows summary of the various trends. In terms of performance, HyTra-ID is always better than previous methods of investigation. On the other hand, previous methodologies have been developed to evaluate datasets in isolation and were not designed to operate in a combined cross-domain environment. Therefore, HyTra-ID provides an appropriate performance level for applicable dialogue systems, as its results are valid across multiple datasets without requiring fine-tuning for individual datasets.

Discussion

The findings from the experiments indicate that HyTra-ID performs strongly across heterogeneous intent detection benchmarks. A key reason for this behavior is that the model does not treat inter-dataset variation purely as a nuisance factor to be removed; instead, it explicitly encodes dataset identity as an additional conditioning signal during training. This allows the model to jointly learn shared semantic patterns across utterances while preserving dataset-specific differences in annotation structure, linguistic style, and intent granularity. In contrast to approaches that rely only on domain-invariant transformer representations, this design provides a more direct mechanism for handling heterogeneity in unified cross-domain intent classification. The following discussion interprets the empirical results in light of these architectural choices and their implications for real-world dialogue systems.

Effectiveness of Dataset-Aware Representation Learning

Incorporating dataset identity before model training is the main design concept behind HyTra-ID. By explicitly incorporating the identity of each dataset during the learning phase, HyTra-ID has an overall accuracy that is greater than 96% and a consistent weighted F1 score across all datasets. The results from the evaluation of HyTra-ID show that the dataset identity incorporated into HyTra-ID mitigates the effects of dataset-related distributional shift. Traditional transformer-based intent models assume that the training data are homogeneous; HyTra-ID does not. Instead, HyTra-ID uses dataset embeddings to model the unique idiosyncrasies in the language, annotation conventions, and granularity of intent found within each of the different datasets.

How to cite this article:

The dataset-wise results show the best demonstration of this phenomenon. For example, despite SNIPS performing nearly optimally because it has a very uniform distribution of intents with little semantic interaction between intents, HyTra-ID achieves a very high degree of accuracy for TOPv2, which has an order of magnitude more intents than SNIPS and is heavily imbalanced across classes. Since both systems operate on very different extremes with regard to dataset size and class imbalance, this result supports the notion that in addition to building a model specifically for respect to how they will be used, dataset-aware conditionality allows generalization from one dataset to another without being overfitted to any single dataset.

Robustness Across Intent Frequency Spectrum

The class-wise evaluation indicates that HyTra-ID performs consistently across both frequent and relatively low-support intent categories within the unified benchmark. High-frequency intents such as GET_WEATHER, CREATE_REMINDER, and PLAY_MUSIC achieve very strong precision and recall, reflecting clear separability under the evaluated setting. The model also maintains reasonable performance on lower-support classes, although these categories remain more challenging because of limited training representation. The confusion-matrix analysis further suggests that rare classes are not catastrophically misclassified within the tested benchmark. However, broader robustness claims would require additional validation under unseen-domain, out-of-distribution, or zero-shot evaluation settings, which remain important directions for future work.

Insights from Ablation Studies

The results of the ablation study revealed important information about how much of the effect on performance can be attributed to components of HyTra-ID as opposed to how much of the effect is the result of interaction between components. For example, when all dataset embeddings are removed from HyTra-ID, there is a significant drop in performance across every dataset tested, with the largest drop observed in performance testing on TOPv2. This demonstrates that employing dataset-aware conditioning is critical to successfully learn from large-scale multi-domain datasets. Additionally, as we reduce the depth of the attention mechanism or the number of levels in the attention hierarchy within HyTra-ID, we see a decrease in the ability of HyTra-ID to represent data. This indicates that hierarchical contextual modeling is key to being able to successfully model large multi-domain datasets. Despite retaining high-accuracy scores on smaller datasets like ATIS as compared to other methods, there are limits on how effectively these architectures can adapt toward more complicated real-world scenarios. The results support the rationale for HyTra-ID balanced approach to architectural design.

Comparative Perspective with SOTA Methods

Compared to recent SOTA techniques for intent detection and also compared to traditional statistical techniques, HyTra-ID has a significant advantage over both in terms of accuracy and robustness. Current methods of intent detection rely heavily upon robust statistical classification and features and exhibit extremely low levels of adaptability to changes between domains; older neural approaches continue to have issues when trying to address very large intent spaces resulting in under-performance of these approaches when evaluated against HyTra-ID or even transformer-based baseline models (all of which have decreased performance and stability when trained together on datasets with a single source/data type).

HyTra-ID improves its overall performance through uniform treatment of domain diversity, variable intent and the representation learning of each of these different contexts. Because of this, it is an effective replacement for model specialization or the process of fine-tuning a specific dataset for the purpose of constructing a model that operates within the specified domain. Both avenues create operational overhead in production environments.

Practical Implications for Real-World Deployment

There are several advantages to HyTra-ID from a deployment standpoint. First, because it utilizes one unified training approach, it removes the necessity to have multiple models for each individual domain or data source. This helps to reduce the amount of maintenance required for model upkeep as there is only one model to maintain. Second, with

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. *Cureus J Comput Sci* 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

HyTra dataset-aware design, it allows you to gradually integrate additional datasets into the training of your model, without having to retrain your entire model from the beginning.

In addition, the model is very robust to both label imbalance and domain variation, which makes it an excellent candidate for real-world conversational agents. This means that the user queries, which are diverse and can be rapidly changing, can be well supported by this model. The importance of these attributes is especially evident in customer service systems, large-scale dialogue platforms, and multi-skill virtual assistants.

Limitations and Open Challenges

Although HyTra-ID has shown significant success in providing strong outputs over a wide variety of intent detection scenarios, it does have limitations. For example, HyTra-ID has been limited in its ability to detect rare intents because of the limited number of supervised input examples available for training. In these cases, HyTra-ID performance on low-support classes can be enhanced with additional regularization or data augmentation techniques. In addition, using dataset embeddings in HyTra-ID provides greater robustness, but creates the need to also rely on the metadata associated with the dataset being used to build HyTra-ID, which may not always be available when working in an open-world environment.

Noting that there is potential for further enhancement through features such as dynamic dataset inference, continuous learning, and the explicit modeling of open set intent, we can now continue to work towards improving the utility of HyTra-ID in conversational settings where the topics being discussed are changing and may not be explicitly defined or known prior to conversation.

In summary, the results reported in this study suggest that HyTra-ID is a promising framework for unified cross-domain intent detection across heterogeneous benchmark datasets. The model shows strong empirical performance and demonstrates the potential value of combining contextual transformer representations with dataset-aware embeddings in a single classification framework. At the same time, the present findings should be interpreted within the scope of the evaluated benchmark, and further work involving stronger statistical validation, deeper dataset analysis, broader baseline comparison, and additional robustness experiments would help strengthen the conclusions.

Conclusions

Conclusion and future work

This work introduced HyTra-ID, a robust transformer-based framework for cross-domain intent detection that explicitly addresses dataset-induced variability in real-world dialogue systems. By integrating dataset aware embeddings within a unified transformer architecture, HyTra-ID effectively captures both linguistic semantics and dataset specific distributional characteristics, enabling stable learning across heterogeneous corpora. Extensive evaluations on ATIS, SNIPS, and TOPv2 demonstrate that the proposed model achieves strong and consistent performance, attaining an overall accuracy of 96.34% while maintaining robustness across datasets with diverse intent granularities and imbalance levels. Ablation studies and class-wise analyses further confirm the importance of dataset-aware conditioning and architectural design choices, validating HyTra-ID as a scalable and practical solution for multi-domain conversational AI systems.

While the present results demonstrate the promise of HyTra-ID for unified cross-domain intent detection, several important directions remain for further strengthening and extending the framework. Future work should include stronger statistical validation, broader comparison against more recent transformer baselines, and more rigorous robustness evaluation under unseen-domain, out-of-distribution, and zero-shot settings. In addition, deeper exploratory analysis of dataset heterogeneity and class imbalance would further improve the interpretability of model behavior across domains. Beyond evaluation, the framework may also be extended to support rare and emerging intents through few-shot or continual learning, as well as open-set intent detection and incremental domain expansion. From a deployment perspective, lightweight and parameter-efficient transformer variants may further improve practical applicability in resource-constrained environments.

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. *Cureus J Comput Sci* 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Khyati C. Kanani, Gopi B. Sanghani

Acquisition, analysis, or interpretation of data: Khyati C. Kanani, Gopi B. Sanghani

Drafting of the manuscript: Khyati C. Kanani, Gopi B. Sanghani

Critical review of the manuscript for important intellectual content: Khyati C. Kanani, Gopi B. Sanghani

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

All data generated or analyzed during this study are included in this published article and/or its appendices.

Acknowledgements

The authors would like to acknowledge Khyati C. Kanani and Gopi B. Sanghani.

References

1. Madan S, Lentzen M, Brandt J, Rueckert D, Hofmann-Apitius M, Fröhlich H: [Transformer models in biomedicine](#). BMC Medical Informatics and Decision Making. 2024, 24:1-22. [10.1186/s12911-024-02600-5](#)
2. Li D, Wang S, Zhao B, Ma Z, Li L: [A novel model based on a transformer for intent detection and slot filling](#). Urban Informatics. 2024, 3:1-9. [10.1007/s44212-024-00056-6](#)
3. Benayas A, Sicilia MA, Mora-Cantalops M: [A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: navigating the trade-offs in model size and performance](#). Language Resources and Evaluation. 2024, 59:2007-2030. [10.1007/s10579-024-09796-y](#)
4. Wang X, Koneru S, Venkit PN, Frischmann B, Rajtmajer S: [The unappreciated role of intent in algorithmic moderation of social media content](#). Harvard Kennedy School Misinformation Review. 2024, 6:1-47. [10.37016/mr-2020-180](#)
5. Muhammad YI, Salim N, Zainal A: [Joint intent detection and slot filling with syntactic and semantic features using multichannel CNN-BiLSTM](#). PeerJ Computer Science. 2024, 10:1-26. [10.7717/peerj-cs.2346](#)
6. Park S, Menassa CC, Kamat VR: [Joint bert model for intent classification and slot filling analysis of natural language instructions in co-robotic field construction work](#). Computing in Civil Engineering 2023. 2024, 453-460. [10.1061/9780784485224.055](#)
7. Atuhurra J, Kamigaito H, Watanabe T, Nichols E: [Domain adaptation in intent classification systems: a review](#). ArXiv, abs/2404.14415. 2024, 1-16. [10.48550/arXiv.2404.14415](#)
8. Ouaddi C, Benaddi L, Bouziane EM, Naimi L, Rahouti M, Jakimi A, Saadane R: [Assessing the effectiveness of large language models for intent detection in tourism chatbots: a comparative analysis and performance evaluation](#). Scientific African. 2025, 28:1-12. [10.1016/j.sciaf.2025.e02649](#)

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

9. Dong Y, Wang Z, Wu T: [An open intent detection model optimized for datasets based on the Bert large model](#). Applied and Computational Engineering. 2024, 47:225-231. [10.54254/2755-2721/47/20241381](#)
10. Pires H, Paucar L, Carvalho JP: [DeB3RTa: a transformer-based model for the Portuguese financial domain](#). Big Data and Cognitive Computing. 2025, 9:1-30. [10.3390/bdcc9030051](#)
11. Kwon N, Yoo Y, Lee B: [Novel curriculum learning strategy using class-based TF-IDF for enhancing personality detection in text](#). IEEE Access: Practical Innovations, Open Solutions. 2024, 12:1-10. [10.1109/access.2024.3417180](#)
12. Shifat FT, Haider F, Sourove MSUR, Barua DD, Ishmam MF, Fahim M, Bhuiyan MA: [Penta-nlp at EXIST 2024 task 1-3: sexism identification, source intention, sexism categorization in tweets](#). CLEF (Working Notes). 2024, 1214-1227.
13. Paran A, Shohan S, Hossain M, Hossain J, Ahsan S, Hoque MM: [Semanticcuetysync at arafinnlp2024: classification of cross-dialect intent in the banking domain using transformers](#). Proceedings of the Second Arabic Natural Language Processing Conference. 2024, 422-427. [10.18653/v1/2024.arabicnlp-1.38](#)
14. Du J, Lu D, Li F, Liu K, Qiu X: [Trajectory prediction and intention recognition based on CNN-GRU](#). IEEE Access: Practical Innovations, Open Solutions. 2025, 13:26945-26957. [10.1109/access.2025.3539931](#)
15. Assayed SK, Alkhatib M, Shaalan K: [Enhancing student services: machine learning chatbot intent recognition for high school inquiries](#). BUID Doctoral Research Conference 2023. Lecture Notes in Civil Engineering. Al Marri K, Mir FA, David SA, Al-Emran M (ed): Springer, Cham; 2024. 473:243-254. [10.1007/978-3-031-56121-4_24](#)
16. Peng M, Guo X, Chen X, Chen K, Zhu M, Chen L, Wang FY: [LC-LLM: explainable lane-change intention and trajectory predictions with large language models](#). Communications in Transportation Research. 2025, 5:1-13. [10.1016/j.commtr.2025.100170](#)
17. Lin IF, Hasibi F, Verberne S: [Generate then refine: data augmentation for zero-shot intent detection](#). Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, 13138-13146. [10.18653/v1/2024.findings-emnlp.768](#)
18. Asif M, Khan TA, Song WC: [Leveraging cognitive machine reasoning and NLP for automated intent-based networking and e2e service orchestration](#). IEEE Access: Practical Innovations, Open Solutions. 2025, 13:19456-19468. [10.1109/access.2025.3534282](#)
19. Chow R, Suen KY, Lam AYS: [On leveraging large language models for multilingual intent discovery](#). ACM Transactions on Management Information Systems (TMIS). 2025, 16:1-17. [10.1145/3688400](#)
20. Perdana RS, Adikara PP: [Multi-task learning for named entity recognition and intent classification in natural language understanding applications](#). Journal of Information Systems Engineering and Business Intelligence. 2025, 11:1-16. [10.20473/jisebi.11.1.1-16](#)
21. Das N, Dingliwal S, Ronanki S, et al.: [Speechverse: a large-scale generalizable audio language model](#). arXiv. 2025, 1-19.
22. Iqbal A, Shahid A, Roman M, Afzal MT, Hassan UU: [Optimising window size of semantic of classification model for identification of in-text citations based on context and intent..](#) PLOS One. 2025, 20:1-28. [10.1371/journal.pone.0309862](#)
23. Belardinelli A: [Gaze-based intention estimation: principles, methodologies, and applications in HRI](#). ACM Transactions on Human-Robot Interaction. 2024, 13:1-30. [10.1145/3656376](#)
24. Saffar AH, Sitbon L, Hoogstrate M, Abbas A, Roomkham S, Miller D: [Human and large language model intent detection in image-based self-expression of people with intellectual disability](#). CHIIR '24: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval. 2024, 199-208. [10.1145/3627508.3638347](#)
25. Bodonhelyi A, Bozkir E, Yang S, Kasneci E, Kasneci G: [User intent recognition and satisfaction with large language models: a user study with chatgpt](#). arXiv. 2024, 1-22.
26. Yee SSS, Soe KM: [Exploring Bert-based encoders for sequence classification and multi-tasking learning in dialogue acts and joint intent-slot filling](#). Indian Journal of Computer Science and Engineering. 2024, 15:317-325.
27. Ghosh N, Santoni D, Saha I, Felici G: [A review on the applications of transformer-based language models for nucleotide sequence analysis](#). Computational and Structural Biotechnology Journal. 2025, 27:1244-1254. [10.1016/j.csbj.2025.03.024](#)
28. Pimparkhede P, Bhattacharyya P: [Main predicate and their arguments as explanation signals for intent classification](#). Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025, 10778-10789. [10.18653/v1/2025.naacl-long.539](#)
29. Lichouri M, Lounnas K, Amziane MZ: [dzFinNlp at AraFinNlp: improving intent detection in financial conversational agents](#). Proceedings of the Second Arabic Natural Language Processing Conference. 2024, 450-455. [10.18653/v1/2024.arabicnlp-1.43](#)

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>

30. Jbene M, Chehri A, Saadane R, Tigani S, Jeon G: [Intent detection for task-oriented conversational agents: a comparative study of recurrent neural networks and transformer models](#). Expert Systems. 2024, 42:1-20. [10.1111/exsy.13712](#)
31. Karim AR, Uzuner O: [Leveraging machine-generated data for joint intent detection and slot filling in Bangla: a resource-efficient approach](#). Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025). 2025, 208-216.
32. Noviyanti P, Gudiato C, Horhoruw LFM: [Tweet sentiment analysis to detect suicide intentions through a machine learning approach to big data using the naive Bayes method](#). Jurnal Teknologi Terapan G-Tech. 2025, 9:381-389. [10.70609/gtech.v9i1.6154](#)
33. Lunagaria M, Katkar V, Vaghela K: [Covid-19 and pneumonia infection detection from chest X-ray images using U-Net, EfficientNetB1, XGBoost and recursive feature elimination](#). International Journal of Advanced Computer Science and Applications. 2022, 13:1-9. [10.14569/ijacsa.2022.0130954](#)
34. Chauhan P, Lunagaria M, Verma DK, Vaghela K, Tejani GG, Sharma SK, Khan AR: [PBViT: a patch-based vision transformer for enhanced brain tumor detection](#). IEEE Access: Practical Innovations, Open Solutions. 2025, 13:13015-13029. [10.1109/access.2024.3521002](#)
35. Yin Y, Hwang E, Carenini G: [SWI: speaking with intent in large language models](#). Proceedings of the 18th International Natural Language Generation Conference. 2025, 684-698.
36. Devi VS, Kannimuthu S, Madasamy AK: [The effect of phrase vector embedding in explainable hierarchical attention-based Tamil code-mixed hate speech and intent detection](#). IEEE Access: Practical Innovations, Open Solutions. 2024, 12:11316-11329. [10.1109/access.2024.3349958](#)
37. Zhang T, Norouzian A, Mohan A, Ducatelle F: [A new approach for fine-tuning sentence transformers for intent classification and out-of-scope detection tasks](#). Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. 2024, 910-919. [10.18653/v1/2024.emnlp-industry.68](#)
38. Aijaz A, Vakaj E, Tabaku E: [Intent discovery pipeline using Z-Bert-A](#). Journal of Advanced Natural Sciences and Engineering Researches. 2025, 9:78-84.
39. Pugazhenth VJ, Pandey G, Jeyarajan B, Murugan A: [AI-driven voice inputs for speech engine testing in conversational systems](#). IEEE (SoutheastCon 2025). 2025, 700-706. [10.1109/SoutheastCon56624.2025.10971485](#)
40. Sivakumar N, Khan AR, Umar S, et al.: [A hybrid brain tumor classification using FL with FedAvg and FedProx for privacy and robustness across heterogeneous data sources](#). IEEE Access: Practical Innovations, Open Solutions. 2025, 13:57705-57719. [10.1109/access.2025.3549440](#)
41. Lunagariya M, Katkar V: [Light weight approach for COVID-19, pneumonia detection from X-ray images using deep feature extraction and XGBoost](#). 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA). 2022, 1-5. [10.1109/ICCSEA54677.2022.9936194](#)
42. Hemphill CT, Godfrey JJ, Doddington GR: [The ATIS spoken language systems pilot corpus](#). Speech and Natural Language: Proceedings of a Workshop. 1990, 1-6.
43. Coucke A, Saade A, Ball A, et al.: [Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces](#). arXiv.org. 2018, 1-29.
44. Chen X, Ghoshal A, Mehdad Y, Zettlemoyer L, Gupta S: [Low-resource domain adaptation for compositional task-oriented semantic parsing](#). Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020, 5090-5100. [10.18653/v1/2020.emnlp-main.413](#)
45. Yang Q, Liu Y, Lu L, Liu L: [GFIDF: gradual fusion intent detection framework](#). The Journal of Supercomputing. 2024, 81:1-29. [10.1007/s11227-024-06708-3](#)
46. Hakkani-Tür D, Tur G, Celikyilmaz A, Chen YN, Gao J, Deng L, Wang YY: [Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM](#). Proceedings of Interspeech. 2016, 715-719. [10.21437/Interspeech.2016-402](#)
47. Liu B, Lane I: [Attention-based recurrent neural network models for joint intent detection and slot filling](#). Proceedings of Interspeech. 2016, 685-689. [10.21437/Interspeech.2016-1352](#)
48. Chao W, Ke Y, Xiaofei W: [Pos scaling attention model for joint slot filling and intent classification](#). 2020 IEEE 20th International Conference on Communication Technology (ICCT). 2020, 1483-1487. [10.1109/ICCT50939.2020.9295901](#)

How to cite this article:

Kanani K C, Sanghani G B (June 28, 2026) Tackling Cross-Domain Intent Variability With HyTra-ID: A Robust Transformer-Based Framework for Real-World Dialogue Systems. Cureus J Comput Sci 3 : es44389-026-00137-4. DOI <https://doi.org/10.7759/s44389-026-00137-4>