Symbolic Graph Structures for Clause-Level Intent Modeling: Evaluating Reasoning-Oriented Abstractions in Clinical Communication

Yi-Chun Chen[1][0009−0003−4035−9894], LinHai Ma[1][0000−0001−8519−864X], Yan Wang[1][0000−0003−1036−9365], and Samah Jarad-Fodeh[1]

Yale University, New Haven, CT, USA
{yi-chun.chen, linhai.ma, yan.wang.yw937, samah.fodeh}@yale.edu

Abstract. Understanding communicative intent in patient-provider messages often requires reasoning across clause-level segments that encode layered or ambiguous meanings. Standard sentence-level models are limited in their ability to capture such variation, particularly when fine-grained intent labels must be assigned at a more localized resolution. To address this, we propose a symbolic graph-based framework for the classification of clause-level intent in asynchronous clinical communication. Our approach models syntactic, semantic, and conceptual structures by constructing a graph for each clause, where the nodes represent internal linguistic units such as tokens, semantic roles, or abstract concepts. To incorporate a broader discourse context, we introduce a narrative-level symbolic graph in which each clause is treated as a node, connected to other clauses in the same message through relations such as temporal adjacency, label similarity, and rhetorical cues. We evaluate these complementary graph abstractions: syntactic, semantic, and narrative, and compare their effectiveness in supporting clause-level classification. Experimental results reveal that different graph types offer distinct advantages depending on the characteristics of the label and the demands of the context. Narrative graphs provide strong training performance but are more prone to overfitting, while simpler semantic graphs offer more stable generalization. These findings highlight the potential of symbolic graphs as scaffolds for localized reasoning in clinical NLP and support their broader integration into interpretable hybrid learning systems.

Keywords: clause-level intent classification · symbolic graph representations · contextual reasoning · hybrid neural-symbolic models · semantic abstraction · clinical natural language processing

Exploratory Strategies for Lab Data Analysis

Strategy 1: keep data unchanged, analyze methods to see factors that influence the results

1. identify model-specific strengths
2. whether context helps (in preliminary experiments, seemed so)

3. altered with different context combining strategies (adjacent, whole, propagate)
4. choose the best context support and best methods
5. does the graph learning method influence the result as well?
6. open questions: How to reason the context to infer the intent?

Strategy 2: play with data and classes (optimize the classes design and content level (message, sentence, subsentence))

1. remove and combine the class that is getting mostly wrong
2. test with new data that was optimized by other team members
3. multi-label support for grouped content
4. based on the result change paper framing

Route 1: Symbolic Method Variation (Same Labels)

- Compare symbolic graph models (e.g., DEP-GCN, SRL-GCN, AMR-GCN, Narrative-GCN) across subcode classes.
- Analyze class-wise F1 scores to identify model-specific strengths.
- Visualize the performance of the class using heat maps or bar plots (model × class).
- Select representative examples for error type analysis (e.g., contrastive pair where SRL succeeds but AMR fails).
- Conduct small ablations (e.g., remove specific edge types like `elaboration`) to test symbolic impact.
- Optionally incorporate graph attention visualization to interpret node/edge salience.

Route 2: Label Structure Adjustment (Same Models)

- Group fine-grained subcodes into broader high-level categories (e.g., use CODE instead of sub-CODE).
- Introduce multi-label support for overlapping or fuzzy intent annotations.
- Use label smoothing or probabilistic labels for ambiguous clauses (e.g., 70% instruction, 30% reassurance).
- Apply weakly supervised relabeling using:
  - LLM-based intent similarity scoring
  - Clustering based on semantic embeddings
  - Manual inspection of common confusion pairs
- Re-run evaluation to compare flat vs. grouped/soft labels.
- Visualize confusion matrix before and after label refinement.

1 Introduction

In asynchronous patient-provider communication, such as secure messages exchanged via mobile applications or patient portals, even brief messages can convey layered meanings. Patients may ask questions, express emotions, reference prior instructions, and indicate preferences, all within a single message. For example, a message such as "I'm still feeling dizzy, should I stop the new medication?" conveys both a health status update and a clarification request. While intuitive to a human reader, such multi-intent expressions present challenges for current clinical natural language processing (NLP) systems, which often assume that each sentence corresponds to a single communicative function.

This assumption limits the ability of these systems to support fine-grained understanding and decision-making, particularly in sensitive healthcare contexts. Although recent work on sentence-level classification has made progress in assigning general labels of communicative intent, such as *information giving* or *emotional support*, these approaches overlook the fact that real-world messages often consist of multiple units similar to clauses. Each of these units may serve distinct semantic or pragmatic purposes and may rely on contextual dependencies for accurate interpretation.

We hypothesize that modeling communicative intent at the clause level, using structured graph representations that encode contextual relationships, enables more accurate and interpretable classification of fine-grained intent labels in asynchronous clinical messages.

To evaluate this hypothesis, we implement and compare several types of clause-level graph representations, including syntactic, semantic, conceptual, and symbolic abstractions. In particular, we investigate whether symbolic narrative relations between segments, such as temporal adjacency or label similarity, can provide complementary context for classification, especially in cases where the internal semantics of a clause are ambiguous or underspecified.

Each message is segmented into clause-level units, which are then represented as nodes in a lightweight graph. Edges encode contextual dependencies between these units, reflecting features such as syntactic structure, predicate-argument roles, and symbolic discourse relationships. Each segment is annotated with one or more intent-oriented labels drawn from a hierarchical schema that includes high-level categories such as *Request* and more specific subtypes such as *Request for clarification* or *Request for medication change*.

Our objective is to determine whether graph-based structural context improves the accuracy and interpretability of fine-grained intent classification. In contrast to flat segment-level classifiers, our approach enables reasoning about the relational structure of a message by modeling how the meaning of a clause depends on its position and contextual links to other clauses within the message.

Research Questions. We frame our study around the following research questions:

- **RQ1:** Does clause-level segmentation yield better alignment between annotated spans and linguistic units than sentence-level segmentation?

- **RQ2:** Can graph-based contextual representations improve sub-sentential classification of communicative intent?
- **RQ3:** How do different types of graph structure, including syntactic, semantic, and symbolic representations, contribute to classification accuracy and interpretability?

Contributions. This paper presents the following contributions:

- We introduce a graph-based framework for clause-level classification of communicative intent in asynchronous clinical messages, using structured representations to model contextual relationships between segments.
- We provide a quantitative evaluation of span alignment across sentence- and clause-level segmentation, demonstrating that clause-level modeling improves alignment with annotated intent spans.
- We compare flat classifiers with a range of graph-based models and offer initial evidence that the contextual structure of the graph, including symbolic narrative relations, improves the accuracy and interpretability of the classification.

While our study is grounded in the clinical communication domain, the proposed framework is generalizable to other applications in which meaning unfolds across structurally interdependent segments, including education, dialogue systems, and narrative understanding. **Our findings contribute to the broader objective of hybrid reasoning in language systems by demonstrating how symbolic structure, both internal to segments and relational across segments, can support context-aware neural models.**

2 Introduction

Understanding communicative intent in patient-provider messages is a critical step toward supporting context-aware clinical decision systems. Asynchronous messages exchanged through mobile applications or patient portals are often brief, informal, and semantically dense. A single message may contain multiple communicative functions, such as describing symptoms, requesting clarification, and referencing prior instructions, all expressed within or across clauses. For example, a message like "I'm still feeling dizzy, should I stop the new medication?" conveys both a health update and a question about treatment. Although intuitive to human readers, such layered expressions present challenges for automated understanding.

Most clinical NLP systems continue to operate at the sentence level, where each sentence is treated as a unit of analysis. However, this assumption limits the granularity and precision needed for downstream applications such as intent detection, response recommendation, or patient education. Real-world messages often contain multiple clause-level segments, each with its own communicative goal. These segments may rely on both the internal structure and contextual relationships for proper interpretation. We propose that symbolic graph structures

offer a principled way to model clause-level intent by capturing both local semantics and contextual dependencies. Graph-based representations allow structural reasoning over linguistic units, enabling models to encode not only lexical content but also the relationships between units. This reasoning capability is essential for disambiguating subtle distinctions in intent, particularly when context is required to distinguish between similar or overlapping labels. While prior work has explored syntactic or semantic graphs in other NLP settings, their comparative utility for clause-level intent modeling in clinical communication remains underexplored.

In this study, we evaluate whether different types of symbolic graphs, constructed from syntactic dependencies, semantic roles, abstract meaning representations, and symbolic discourse links, can support accurate and interpretable clause-level intent classification. Most graph types are constructed at the clause level, with internal nodes representing words, concepts, or roles. To model broader discourse context, we additionally construct a symbolic narrative graph at the message level. In this representation, each clause becomes a node and is linked to its neighbors based on temporal, rhetorical, or semantic similarity. Our central hypothesis is that clause-level classification can be improved through symbolic structure that supports lightweight reasoning over internal and contextual features. Rather than viewing symbolic and neural models as opposing paradigms, we treat symbolic graphs as scaffolds for localized reasoning that complement neural encoders in hybrid classification tasks.

We frame our study around three research questions:

- **RQ1**: Does clause-level segmentation yield better alignment between annotated spans and linguistic units than sentence-level segmentation?
- **RQ2**: Can symbolic graph-based representations improve clause-level classification of communicative intent?
- **RQ3**: How do different types of symbolic structure, including syntactic, semantic, and narrative forms, contribute to the accuracy and interpretability of the classification?

This paper offers three main contributions:

- We introduce a symbolic graph framework for clause-level intent classification in clinical messaging, combining local and contextual structural features.
- We provide a comparison of segmentation strategies, demonstrating that clause-level modeling improves alignment with annotated intent spans.
- We evaluate a diverse set of graph abstractions and show that narrative context and symbolic structure yield distinct performance trade-offs. These results offer insight into their role in hybrid reasoning systems.

Although our study focuses on clinical communication, the proposed framework generalizes to other domains where meaning unfolds across structurally interdependent segments. Our findings contribute to the broader goal of interpretable, reasoning-oriented NLP by demonstrating how symbolic structure can support hybrid models for clause-level understanding.

3 Background and Related Work

Understanding clause-level communicative intent in clinical messaging intersects with research on sentence and clause segmentation, symbolic graph reasoning, and hybrid neural-symbolic models.

3.1 Sentence and Clause-Level Modeling in Clinical NLP

Sentence-level classification is widely used in clinical NLP for phenotype extraction, risk assessment, and decision support. Early systems were largely rule-based or relied on feature engineering, but this paradigm shifted with the availability of large clinical datasets such as MIMIC-III [23] and the adoption of neural architectures.

Comparative studies have demonstrated that neural models outperform traditional approaches for intent classification and phenotyping tasks, especially when supported by contextual embeddings [14, 17]. Structured sequence models such as BiLSTM-CRF [22] and convolutional encoders for longitudinal medical records [33] have also shown utility, particularly when adapted to clinical contexts. Transformer-based models pretrained on clinical corpora, such as ClinicalBERT [2], have improved performance on named entity recognition and sentence classification tasks. Pretrained sentence embeddings have also enhanced downstream performance in semantic similarity and chronic disease prediction tasks [27, 5].

In parallel, efforts have been made to improve interpretability through model transparency [30] and to adapt sentence classification to less formal sources such as social media or patient-authored content [12, 39]. Domain-specific embeddings and task-tuned pretraining further support adaptation to biomedical sublanguage [32]. However, sentence-level approaches typically assume one intent per sentence and offer limited mechanisms for representing intra-sentential variation. This assumption limits their ability to capture layered or interdependent communicative goals. In this work, we address these limitations by segmenting at the clause level and using structured graph representations to model contextual and functional dependencies.

3.2 Clause-Level Structure and Segmentation

Clause-level modeling has emerged as a promising approach for capturing fine-grained semantic distinctions within a sentence. This is particularly relevant in biomedical and clinical domains, where propositions may be coordinated, nested, or expressed as fragments. Approaches to clause segmentation include rule-based chunking, semantic parsing, and neural segmentation frameworks.

Several studies have proposed techniques for decomposing complex sentences into interpretable clause-level propositions, enabling finer semantic control in downstream tasks [15]. Others have employed unsupervised structure induction to uncover latent clause boundaries and discourse units [43]. Graph-based representations have also proven effective for clause-level structure. Models that

combine syntactic and semantic dependencies, such as dual-graph architectures, have been used for frame semantic parsing [48]. Sentiment-based clause graphs [38] and enhanced universal dependency parsers [3] support alignment between syntax and meaning at the clause level.

Clause-level modeling has also been applied to narrative understanding and clinical contexts. Prior work has demonstrated that maintaining clause-wise coherence is crucial for modeling segmented narrative meaning in multimodal communication [7]. Clause complexity and nonstandard syntax in clinical corpora pose challenges for traditional NLP pipelines, motivating dedicated clause-level systems [47, 40]. Other challenges include nominalization and alternation patterns that obscure clause boundaries [9]. Although clause segmentation is gaining traction, many approaches treat structure learning and interpretation separately. Our approach combines clause-level segmentation with symbolic graph abstraction, enabling direct reasoning over localized meaning units and their context.

3.3 Symbolic Graph Reasoning for Text

Symbolic graphs have gained wide attention as a means of representing structured linguistic, semantic, and discourse information in NLP. These methods have been applied in semantic parsing, sentence fusion, relation extraction, and multi-hop question answering. In knowledge-driven tasks, graph-based query representations have enabled interpretable reasoning over complex questions and schema constraints [24, 34]. Structured intermediate forms like GraphQ-IR allow integration between symbolic structure and transformer encoders.

Semantic and event-centric graphs have also been used to aggregate information across sentence boundaries, improving tasks like sentence fusion or document-level linking [45, 31]. Graph-based classification models have reframed textual classification problems as node or graph prediction tasks [36], and concept-based formalisms have supported multilingual semantic reasoning [29].

These methods have been surveyed extensively [49, 18], with increasing focus on explainability and structure-aware learning. In biomedical applications, ensemble-based clustering on graph representations has been applied to clinical and genomic datasets [13]. More recent work has extended symbolic graph modeling to multimodal and narrative domains, using structured graphs for joint reasoning over visual and textual representations [6]. Despite these advances, most work focuses on sentence- or document-level graphs. Few studies have investigated symbolic graph construction and reasoning at the clause level for informal, conversational clinical messages. Our work addresses this gap by exploring a range of symbolic graph designs, including narrative-level and clause-centered variants, for intent classification.

3.4 Hybrid Symbolic-Neural Approaches

The integration of symbolic structure with neural models is central to hybrid AI research, combining human-interpretable representations with data-driven learning. This paradigm has been applied to parsing, structured generation, and

logic-driven inference. Recent frameworks have translated input text into symbolic graph forms through multilingual sequence models [35], or incorporated symbolic constraints to enhance compositional generalization in semantic parsing [42]. Structured generation tasks have benefited from encoding schema and syntactic relations as symbolic scaffolds during decoding [20, 21].

Other models separate neural perception from symbolic reasoning using differentiable execution layers or logic modules to enhance interpretability and composability [26, 44, 11]. Symbolic guidance has also improved data efficiency and controllability in multimodal pipelines, including applications in narrative visualization and content generation [28, 8]. Most existing work in hybrid systems has focused on symbolic supervision for generation or parsing. Few studies have applied symbolic structure as a reasoning scaffold for clause-level classification. The present work contributes to this line of inquiry by evaluating symbolic graphs as structural supports for clause-level communicative intent modeling in clinical NLP.

Together, these lines of work establish the value of structured representations for capturing fine-grained meaning and supporting interpretability. However, prior studies often treat segmentation and symbolic modeling in isolation or operate at broader sentence or document levels. In contrast, our work integrates clause-level segmentation with lightweight symbolic graphs to support localized, explainable reasoning in intent classification.

4 Dataset and Annotation

The dataset consists of 726 anonymized messages exchanged asynchronously between patients and healthcare providers via a real-world clinical communication platform. Messages vary in length and structure but are typically informal and conversational, reflecting the characteristics of mobile or portal-based interactions. A single message often encodes multiple communicative intents, including symptom updates, medication inquiries, emotional expressions, and clarification requests.

Each message was manually annotated with one or more spans corresponding to segments that convey communicative intent. Labels were assigned using a hierarchical schema developed in collaboration with medical experts and annotators familiar with patient-provider dialogue. The codebook is designed specifically for this context and includes three semantic levels:

- **Code** – High-level pragmatic functions (e.g., *Information-Seeking*, *Instruction-Giving*, *Socioemotional/Empathy*).
- **Subcode** – Mid-level semantic categories that refine the intent (e.g., *Drugs*, *Express Concern/Unease*).
- **Subsubcode** – Optional fine-grained descriptors capturing specific expressions (e.g., *Dosage Clarification*, *Financial Insecurity*).

A span may carry multiple labels across levels to reflect compound or layered intent. Although speaker identity is sometimes implied by the labels (e.g., *Part-*

nershipProvider), speaker roles are not explicitly annotated. All annotations and models operate at the message level and do not rely on turn-taking structure.

4.1 Illustrative Example

Table 1 shows an example message annotated at the clause level, demonstrating the presence of multiple communicative intents within a single message. Clauses include both task-oriented and socioemotional content, motivating the need for fine-grained segmentation.

Table 1: Example of an annotated patient message with multiple clause-level intents. Each row represents a segmented clause with its assigned Code and Subcode labels.

Clause	**Assigned Labels (Code, Subcode)**
Hi there,	**PartnershipProvider, salutation**
He definitely needs an appointment.	**InfoGive, SchedulingAppt**
In person or telemed	**PartnershipProvider, inviteCollaboration**
IS if you need	**PartnershipProvider, connection**
I put him down for the "combo" :) on MM/DD/YYYY at 10:00am.	**InfoGive, SchedulingAppt**

4.2 Segmentation and Label Distribution

To prepare clause-level modeling units, we segmented each message into sentences using NLTK's Punkt tokenizer, and further into clauses using spaCy's dependency parser. These clause-level segments serve as the core units for classification and graph construction. Table 2 reports the number of labeled units at each semantic level for both sentence- and clause-based segmentation.

Table 2: Label counts at each semantic level across sentence- and clause-level segments.

Unit Level	**Code**	**Subcode**	**Subsubcode**
Sentence	1,622	1,622	79
Clause	2,081	2,081	103

Clause-level segmentation enables more granular alignment with annotated intents, particularly at the Subsubcode level.

4.3 Span Alignment Evaluation

To evaluate the effectiveness of segmentation, we assessed alignment between annotation spans and segmented units using fuzzy string matching. A segment was considered aligned if its similarity score with any annotation span in the same message exceeded 0.6, based on `difflib.SequenceMatcher`. Table 3 summarizes the alignment statistics.

Table 3: Alignment between annotation spans and segmented units at the sentence and clause levels.

Metric	**Count**	**Match Rate**
Total Messages	726	–
Total Sentences	1,874	–
Total Clauses	4,210	–
Total Annotations	2,602	–
Matched Annotations (Sentence Level)	1,659	63.76%
Matched Annotations (Clause Level)	2,024	77.79%

4.4 Types of Communicative Intent

In dialogue systems and discourse analysis, the term *intent* generally refers to the underlying purpose or communicative goal of a speaker's utterance. However, this term is used variably across domains and datasets, encompassing different layers of meaning. To support symbolic modeling and structured reasoning, we distinguish between two complementary types of intent:

- **Interactional intent:** the communicative function or discourse role of an utterance (e.g., question, inform, greet).
- **Goal-oriented intent:** the semantic objective or real-world task the speaker aims to fulfill (e.g., book a flight, check symptoms).

In addition, we highlight a cross-cutting distinction between intents that are *explicitly stated* in the surface form of an utterance versus those that are *implicitly conveyed* and must be inferred from context.

Interactional Intent (Communicative Function) Interactional intent captures the discourse-level role of an utterance in a conversation. It reflects how the utterance contributes to the dialogue flow and includes categories such as *inform*, *request*, *question*, and *greet*. These functions are often modeled using dialogue act taxonomies.

The DailyDialog dataset [25] annotates utterances with such communicative functions. The ISO 24617-2 standard [4] further provides a rich multidimensional framework for annotating dialogue acts, covering feedback, social

obligations, and task-related roles. Similarly, MultiWOZ [46] uses hybrid labels like `hotel-inform` that combine interactional function with domain semantics.

Interactional intent is critical for modeling turn-taking, conversation management, and discourse coherence, particularly in multi-turn or mixed-initiative dialogue systems.

Goal-Oriented Intent (Semantic Purpose) Goal-oriented intent refers to the underlying task, purpose, or information need the speaker seeks to address. These intents are typically domain-specific and form the basis of backend actions in task-oriented dialogue systems. For example, utterances may express goals such as *transfer_ money*, *book_ restaurant*, or *get_ weather*.

The CLINC150 dataset [1] includes 150 such goal-oriented intents across ten domains, supporting zero-shot classification. The SNIPS dataset [10] similarly categorizes utterances by real-world tasks (e.g., `AddToPlaylist`, `BookRestaurant`), aligning with structured intent-slot systems used in voice assistants.

Unlike interactional intent, which is often stable across domains, goal-oriented intent is more variable and tied to domain-specific taxonomies and APIs.

Explicit vs. Implicit Intent Intent can also be categorized by how directly it is expressed. An utterance like "I want to book a flight" conveys an **explicit** intent to book a flight. In contrast, "I'm flying to Chicago tomorrow" implies the same goal but leaves it **implicit**, requiring contextual or commonsense reasoning.

This distinction is especially important in open-domain or clinical dialogues, where speakers may downplay, soften, or indirectly express their needs. In such settings, symbolic or graph-based methods may support the inference of implicit intent by modeling surrounding context, speaker goals, and discourse patterns.

Summary: Our typology clarifies the multi-layered nature of intent in dialogue. In this work, we model both interactional and goal-oriented intent, and examine how symbolic graph representations can support reasoning across explicit and implicit forms. This provides a flexible foundation for both public datasets (e.g., CLINC150, DailyDialog) and internal clinical data with complex annotation structures.

Application to Lab Data In our internal clinical dataset, annotated codes are structured hierarchically (code, subcode, subsubcode), but often conflate interactional and goal-oriented intents. For example, a message labeled under *Medication/Request/Refill* combines communicative function (request) and domain-specific goal (refill medication).

This mixed granularity hinders the interpretability and reuse of the annotation scheme. Our proposed typology provides a framework for disentangling intent layers and supports better symbolic modeling. We identify cases where similar interactional forms (e.g., *requests*) serve distinct semantic goals (e.g., appointments vs. prescriptions). Conversely, semantically similar goals (e.g., medication queries) may be expressed through varied interactional forms (e.g., questions, symptom descriptions).

Future work will explore structured representations (e.g., clause-level graphs) that model these distinctions and support both supervised intent classification and symbolic reasoning across explicit and implicit intent layers.

5 Symbolic Graph Design for Intent Classification

To support clause-level communicative intent classification, we construct symbolic graph representations that encode varying forms of linguistic abstraction. These graphs serve as structured contexts for reasoning over syntax, semantics, conceptual meaning, and discourse cues while maintaining fine-grained, clause-level modeling resolution.

5.1 Design Goals and Modeling Units

We follow a unified clause-centric pipeline (Figure 1) that begins with raw message text and proceeds through clause segmentation, graph construction, feature embedding, and symbolic graph-based classification. All models operate at the clause level, with each clause functioning as an independent unit for intent prediction.

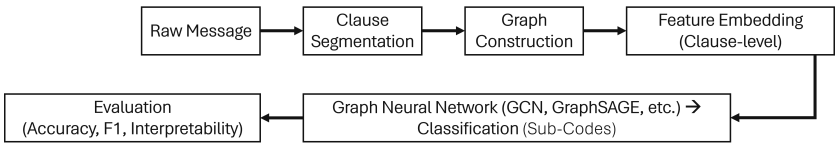


Fig. 1: Overview of the clause-level graph classification pipeline.

Each clause is transformed into a symbolic graph using one of several abstraction strategies. Node and edge features are extracted and encoded before being passed to a Graph Neural Network (GNN) for sub-CODE prediction. This pipeline enables controlled comparison of different graph types under a consistent modeling setup.

5.2 Graph Abstraction Strategies

We compare four categories of symbolic graph representations: *syntactic*, *semantic*, *conceptual*, and *narrative*. All graphs are constructed at the clause level but differ in the structural and linguistic abstractions they encode. Representative examples are shown below; a complete set of visualizations is provided in Appendix A.2.

Syntactic Dependency Graph (DEP-GCN) Syntactic graphs encode surface grammatical structure using dependency parsing. Each clause is parsed using `spaCy`, and the resulting graph includes tokens as nodes and grammatical relations as labeled, directed edges (e.g., `nsubj`, `dobj`, `amod`).

While this representation does not capture semantics directly, it provides a baseline for evaluating the contribution of shallow syntactic cues.

Semantic Role Graphs (SRL-GCN Variants) We implement three variants of semantic role labeling (SRL) graphs, each offering a distinct abstraction over predicate-argument structures:

Role-Weighted Focus Graph. This design connects the clause to salient semantic roles (e.g., `agent`, `theme`, `location`) using weighted edges. Weights reflect the estimated relevance of each role to communicative intent. This structure supports interpretable emphasis on event participants.

Predicate-Centric Graph. This variant centers the graph on the clause's main predicate. Using AllenNLP's SRL parser, we extract predicate-argument structures and represent argument roles (e.g., `ARG0`, `ARG1`) as connected nodes. This abstraction highlights event compositionality.

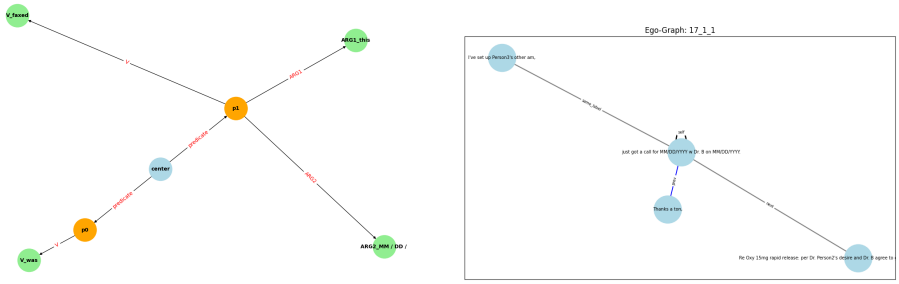
Sentence-Anchored Graph. To integrate clause semantics with surrounding context, we introduce a sentence-level anchor node that links all predicates and arguments within the clause. This design supports hybrid integration with other abstraction layers. An example is included in Figure 2.

Abstract Meaning Representation (AMR-GCN) AMR graphs encode clauses as rooted, directed graphs with labeled semantic relations. Nodes represent abstract concepts (e.g., `schedule-01`), and edges represent roles or attributes (e.g., `:ARG0`, `:time`). We use the BART-based `amrlib` parser to generate AMR graphs, followed by structural cleanup for GNN compatibility. While AMR supports rich conceptual abstraction, it introduces challenges in alignment and interpretability.

Symbolic Narrative Ego Graph (MSG-GCN) To model lightweight discourse context, we introduce symbolic ego-graphs in which each clause forms a local graph centered on itself and connected to neighboring clauses via symbolic edges. Relation types include:

- `prev`, `next`: sequential adjacency;
- `same_label`: shared CODE labels;
- `elaboration`, `contrast`: inferred via discourse markers or heuristic cues.

This localized symbolic graph structure supports clause-level reasoning grounded in both relational and pragmatic signals. An example is provided in Figure 2.



(a) Sentence-anchored SRL graph: predicates and arguments linked via a clause-level anchor.

(b) Symbolic ego-graph with clause-level narrative relations.

Fig. 2: Examples of semantic and symbolic narrative graphs used for clause-level modeling.

5.3 Summary of Graph Properties

Table 4 summarizes the modeling setup, graph structure, and their relationship to our research questions. All models are evaluated under a unified clause-level classification framework.

5.4 Context

Understanding communicative intent often requires information beyond the current clause or sentence. Symbolic context can provide essential disambiguating cues, especially in dialogic or narrative communication. This section outlines our context integration strategy and evaluates how different symbolic propagation mechanisms influence classification performance.

Motivation Preliminary experiments suggest that symbolic graph models benefit from incorporating contextual signals such as adjacent discourse units or rhetorical cues. However, it remains unclear which types and depths of symbolic context yield the most consistent gains. We hypothesize that controlled context propagation improves generalization and interpretability, while unbounded expansion risks semantic blurring or overfitting.

Context Integration Strategies We adopt a clause- or sentence-centered ego-graph design and explore several symbolic context integration strategies. Each approach maintains a distinct local neighborhood to ensure unit-specific prediction.

– **Direct Adjacent Context (1-hop):** The ego-graph includes immediate symbolic neighbors linked via relations such as `next` or `previous`. This preserves temporal or discourse continuity without introducing remote nodes.

Table 4: Model and graph setups for clause-level sub-CODE classification.

Model	**Graph Setup**	**Unit**	**Type**	**RQ**
MLP (Sentence)	Sentence-level BERT	Sentence	None	RQ1
MLP (Clause)	Clause-level BERT	Clause	None	RQ1, RQ2
DEP-GCN	spaCy dependencies	Clause	Syntactic	RQ2, RQ3
SRL-GCN-wt	Weighted roles (spaCy)	Clause	Semantic	RQ2, RQ3
SRL-GCN-pred	Predicate-centered (AllenNLP)	Clause	Semantic	RQ2, RQ3
SRL-GCN-anch	Anchor-based (AllenNLP)	Clause	Semantic	RQ2, RQ3
AMR-GCN	AMR from `amrlib`	Clause	Conceptual	RQ2, RQ3
Narrative-MLP	Symbolic context only	Clause	Narrative	RQ2, RQ3
Narrative-DEP-GCN	Ego-graph w/ symbolic edges	Clause	Narrative	RQ2, RQ3
Narrative-SRL-Weight-GCN	Ego-graph w/ symbolic edges	Clause	Narrative	RQ2, RQ3
Narrative-SRL-predicate-GCN	Ego-graph w/ symbolic edges	Clause	Narrative	RQ2, RQ3
Narrative-SRL-anchored-GCN	Ego-graph w/ symbolic edges	Clause	Narrative	RQ2, RQ3
Narrative-AMR-GCN	Ego-graph w/ symbolic edges	Clause	Narrative	RQ2, RQ3

- **Propagated Symbolic Context (2-hop):** The graph expands to include two-hop neighbors, enabling symbolic context propagation through intermediate nodes (e.g., `next → elaboration → current`).
- **Edge-Weighted Ego-Graph:** Symbolic edges are assigned weights based on their relation types (e.g., `contrast`, `elaboration`), which modulate information flow during GCN learning.
- **Depth-Limited Symbolic Expansion:** A generalized strategy where nodes within a symbolic distance k are retained, allowing experimentation with $k \in \{1, 2, 3\}$ while maintaining an ego-centric view.

These strategies enable symbolic context modeling while avoiding semantic drift across unrelated message segments.

Research Questions and Modeling Plan The following table outlines the context strategies and their corresponding research questions (RQs):

Result Table Placeholder We will compare classification performance (e.g., accuracy, macro-F1) across context strategies and model types. The following table provides a placeholder format:

This evaluation will help identify how symbolic context strategies influence model behavior and guide the design of future hybrid reasoning architectures.

Table 5: Context Strategies and Related Research Questions

Strategy	**Research Question(s)**
1-hop (Adjacent)	**RQ1:** Does immediate symbolic context improve local intent classification?
2-hop (Propagated)	**RQ2:** How far should symbolic context propagate for generalization without noise?
Edge-Weighted	**RQ3:** Can edge salience (e.g., `elaboration`, `contrast`) improve intent modeling?
k-hop Expansion	**RQ4:** What is the optimal symbolic depth (k) for ego-graph reasoning?

Table 6: Performance across Context Strategies (Clause-Level)

Model	**Context**	**Acc.**	**F1**	**Comments**
MLP	None	–	–	No symbolic input
MLP	1-hop (features)	–	–	Adjacent clauses only
GCN	1-hop	–	–	Symbolic direct neighbors
GCN	2-hop	–	–	Propagated symbolic context
GCN	Edge-Weighted	–	–	Salience-weighted edges
GCN	k-hop	–	–	Tuned symbolic depth

6 Experimental Setup

This section defines the task formulation, model variants, and training settings. All experiments focus on clause-level classification under a unified framework. While a selected figure appears below for comparison of sentence vs. clause-level inputs, complete learning curves for all models are provided in Appendix A.3.

6.1 Task Definition and Evaluation Metrics

We formulate clause-level sub-CODE classification as a multi-class prediction task over sub-sentential spans. Each clause is assigned a single communicative intent label. We report the following metrics:

- **Accuracy**: Proportion of correct predictions.
- **Macro-F1**: Unweighted average of F1 across all subcodes, emphasizing performance on minority classes.
- **Cross-Entropy Loss**: Used for optimization and overfitting diagnosis.

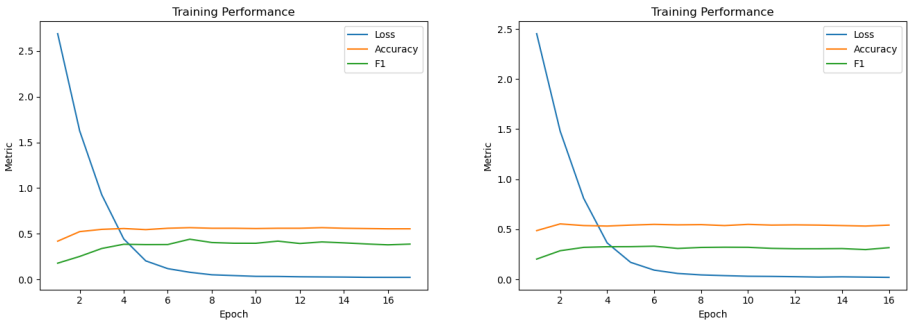
6.2 Model Architectures

MLP Baselines To evaluate the impact of segmentation granularity (**RQ1**), we compare two flat BERT-based models:

- **Sentence-MLP**: Uses sentence-level BERT embeddings.

– **Clause-MLP**: Uses sub-sentential clause embeddings.

Figure 3 shows training loss and macro-F1 progression. Despite similar learning dynamics, clause-level input provides finer granularity for modeling intent (see Section 7 for discussion).



(a) Sentence-level MLP (b) Clause-level MLP

Fig. 3: Training loss and macro-F1 curves for sentence vs. clause-level MLPs.

Graph Convolutional Networks (GCNs) To evaluate the effect of graph structure and abstraction level (**RQ2, RQ3**), we implement the following clause-level GCN models:

– **DEP-GCN**: Syntactic dependency graph (tokens and dependency edges).
– **SRL-GCNs**: Semantic role graphs including:
  • **SRL-GCN-weighted**: Clause node with role-weighted arguments.
  • **SRL-GCN-anchored**: Clause-level anchor linking predicates and arguments.
  • **SRL-GCN-predicate**: Predicate-centered argument structure (underperforms).
– **AMR-GCN**: Conceptual graph from AMR parser (`amrlib`).
– **MSG-AMR-GCN**: Symbolic narrative ego-graph enriched with AMR features.
– **MSG-MLP**: Flat clause-level MLP with symbolic graph features (no GCN).

For graph construction examples, see Figure 2 and Appendix A.2. Model training dynamics for GCNs appear in Appendix A.3.

Graph Learning Variants To evaluate how different graph reasoning mechanisms influence clause- or sentence-level intent classification, we experiment with a set of representative graph-based learning models. These include both standard architectures and symbolic-aware variants designed to preserve structure, interpretability, and local reasoning granularity.

- **GCN (Graph Convolutional Network):** A baseline model that aggregates information from neighbors using mean-based convolution. This serves as a starting point for evaluating symbolic graph utility.
- **Graph Attention Network (GAT):** Incorporates learnable attention weights to dynamically weigh neighbor contributions. Particularly useful for modeling symbolic edge importance (e.g., `elaboration` vs. `contrast`).
- **Edge-Weighted GCN:** A modified GCN that integrates static symbolic edge weights into the message-passing process. Useful when edge-type salience is predefined.
- **Depth-Controlled GNNs:** We vary the number of message-passing layers to assess the effect of symbolic context propagation depth (e.g., 1-hop vs. 2-hop).
- **MLP with Symbolic Features (Non-Graph Baseline):** A non-graph model that uses manually constructed symbolic features (e.g., neighbor label distributions, discourse indicators) as flat inputs for comparison.
- **LLM-Augmented Symbolic GCN (Planned Extension):** A future variant combining symbolic graphs with LLM-derived features (e.g., rhetorical function prompts or relation embeddings), enabling symbolic–neural hybrid reasoning.

These models allow us to assess the trade-offs between fixed symbolic reasoning, learned neighbor salience, and multi-layer graph propagation. Together with variations in context integration (Section 5), this setup supports a robust investigation of symbolic graph design for intent classification.

6.3 Training Settings and Splits

We use a stratified 80–20 train-validation split for all models. MSG models additionally report generalization on a held-out test set. Key settings include:

- **Optimizer**: AdamW with learning rate 1e−4
- **Epochs**: Up to 200 with early stopping based on validation macro-F1
- **Batch size**: 16
- **Initialization**: All runs use fixed random seeds

We select the best model checkpoint using validation macro-F1. Extended learning curves for all graph variants are shown in Appendix A.3.

6.4 Data Validation

Before training models, we performed rigorous data validation to ensure the integrity and compatibility of inputs for both non-graph (MLP) and graph-based experiments. This step is crucial for avoiding silent errors, ensuring correct supervision, and maintaining consistent graph structure assumptions.

For the MLP setup, we verified that each clause- or sentence-level input instance was properly aligned with its corresponding label. This included checking

that the number of data instances matched the number of labels, and that label mappings were consistent with the defined codebook or annotation schema.

For the graph-based experiments, additional validation steps were performed:

- We ensured that each input graph had the expected structure and node-level features (e.g., symbolic embeddings, clause representations).
- We validated the presence of required edge types and that no graphs had isolated or degenerate nodes unless intentionally allowed.
- We confirmed that every graph was connected to a valid label index and could be batched by the graph loader without format or dimensionality errors.
- We manually inspected a sample of constructed graphs for each type (e.g., SRL, AMR, symbolic narrative) to verify edge construction logic.

These validation steps were repeated after any data transformation or graph construction update to ensure input correctness throughout the modeling pipeline. Overall, early and systematic validation helped prevent downstream model instability and facilitated smoother experimentation across multiple model variants.

6.5 Generalization toward different datasets

Pick one or two other datasets rather than lab data

–

7 Results and Analysis

This section presents empirical results comparing symbolic and flat models across classification accuracy, generalization, and label-wise error. Full per-class plots and training curves are provided in Appendix A.3 and A.4; only representative visualizations are shown here.

7.1 Overall Model Performance

Table 8 summarizes classification results across flat and graph-based models for clause-level sub-CODE prediction. Sentence- and clause-level MLPs provide moderate baselines (macro-F1 $\approx$ 0.33–0.40), while GCN-based models offer modest improvements from structural inductive bias.

The best training performance is seen with MSG-AMR-GCN (Train Acc: 91.2%, F1: 0.882), but its generalization is limited (Test Acc: 57.3%). MSG-MLP achieves slightly lower training accuracy (57.4%) but nearly matches test performance (56.7%) with lower overfitting.

7.2 Narrative Graphs: MSG-AMR-GCN vs. MSG-MLP

MSG-AMR-GCN achieves the highest training accuracy but generalizes poorly due to overfitting. The training curve (Appendix A.3) shows a widening gap between training and validation performance. By contrast, MSG-MLP uses symbolic features without edge propagation, offering better generalization under limited data.

Table 7: Classification performance of sub-CODE prediction using different graph construction methods. MSG models are evaluated on held-out test data; other models use validation splits.

Model	Graph Setup	Train Accu	F1	Loss	Test Acc
MLP	No Graph Sentence-level	55.4%	0.401	0.0308	—
MLP	No Graph Clause-level	54.9%	0.331	0.0924	—
DEP-GCN	Dependency Graph (spaCy, token-level)	52.5%	0.360	0.0548	—
SRL-GCN-weighted	Semantic Role Graph (SRL, weighted)	54.0%	0.358	0.5763	—
SRL-GCN-predicate	Predicate-Centric SRL Graph (SRL, clause-level)	29.6%	0.191	0.6263	—
SRL-GCN-anchored	Clause-Anchored SRL Graph (SRL, clause-level)	54.9%	0.318	0.0132	—
AMR-GCN	AMR Graph (SPRING / AMRlib)	48.9%	0.272	0.1815	—

7.3 Comparison Across Graph Structures

Among GCN-based models:

- **DEP-GCN** achieves F1: 0.360 with token-level syntax.
- **SRL-GCN-weighted** performs similarly (F1: 0.358).
- **SRL-GCN-anchored** reaches 54.9% accuracy but lower F1 (0.318).
- **SRL-GCN-predicate** underperforms with F1: 0.191.
- **AMR-GCN** yields the lowest semantic GCN performance (F1: 0.272).

These comparisons suggest that modeling unit alignment (e.g., clause anchoring) matters more than semantic abstraction alone.

7.4 Per-Class F1 Analysis

Figure 4 5 6 compares per-class F1 scores across five models. GCNs with symbolic or semantic structure better capture organized labels (e.g., `Instruction`), while MLPs excel on localized signals (e.g., `Signoff`).

7.5 Error Analysis by Label Type

Appendix A.4 presents a label-level error heatmap for MSG-AMR-GCN. Confusions cluster in overlapping or low-resource classes like `Generalinformation`, `Appreciation/Gratitude`, and `InviteCollaboration`. Frequent classes (e.g., `Signoff`) are learned robustly but occasionally confused with structurally similar ones (e.g., `Salutation`).

Table 8: Context with different abstractions

Model	Graph Setup	Train Accu	F1	Loss	Test Acc
MSG-MLP	Narrative Graph (ego-graph with semantic)	57.4%	0.3882	2.3132	56.7%
Narrative-DEP-GCN	Narrative Graph (ego-graph with amr)				
Narrative-SRL-Weight-GCN	Narrative Graph (ego-graph with amr)				
Narrative-SRL-predicate-GCN	Narrative Graph (ego-graph with amr)				
Narrative-SRL-anchored-GCN	Narrative Graph (ego-graph with amr)				
MSG-AMR-GCN	Narrative Graph (ego-graph with amr)	91.2%	0.882	18.387	57.3%

7.6 Discussion: Symbolic Graphs as Reasoning Scaffolds

Our results support the hypothesis that symbolic graph design facilitates clause-level reasoning:

- Clause-centered graphs outperform predicate-centric variants.
- Symbolic + semantic graphs (MSG-AMR-GCN) enhance capacity but require regularization.
- Symbolic features improve flat MLPs (MSG-MLP), even without explicit structure.

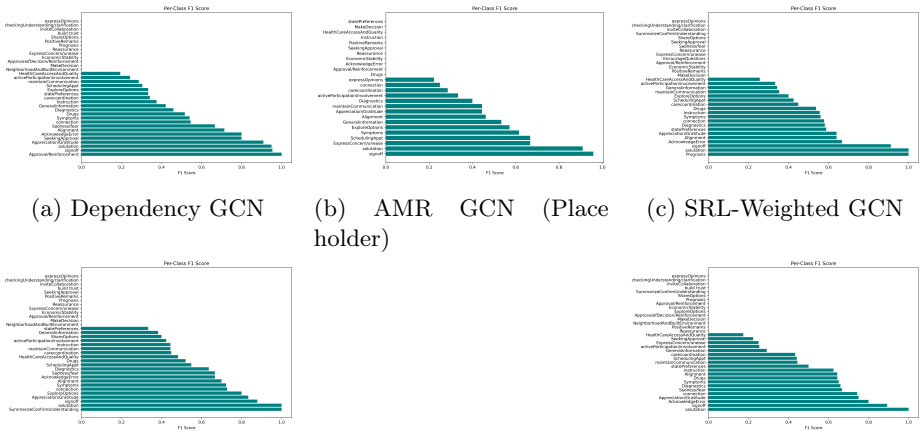
Future work may explore combining symbolic graphs with transformers, pruning relation types, or introducing hierarchical multi-clause structures.

Note on Appendix. Training curves and extended per-class error plots are included in Appendix A.3 and A.4. We include only representative summaries in the main text due to space constraints.

8 Discussion

We reflect on key limitations, interpretability potential, and future directions for symbolic graph models in clinical NLP. Before outlining concrete opportunities, we acknowledge several important challenges and limitations observed during model evaluation:

Limitations: Graph Quality, Overfitting, and Data Scale. Despite improved modeling resolution, symbolic and semantic graphs introduce challenges. The MSG-AMR-GCN model, while effective during training, shows limited test generalization. We attribute this to (1) limited dataset size, (2) low semantic diversity, and (3) the sparsity or over-specificity of symbolic structures, especially in narrative ego-graphs that encode idiosyncratic clause links.



(a) Dependency GCN (b) AMR GCN (Place holder) (c) SRL-Weighted GCN

(d) Sentence MLP (e) Clause-level MLP

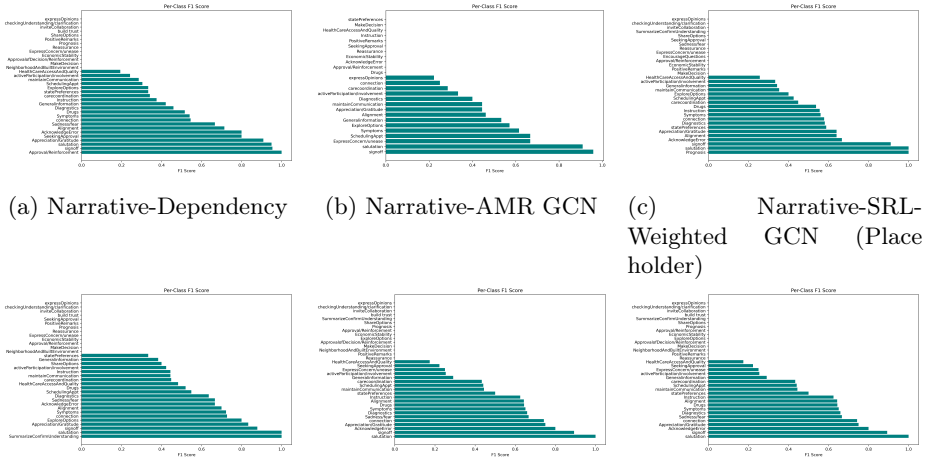
Fig. 4: Per-class F1 score comparison across five representative models. Each plot shows sorted F1 scores across 35 sub-CODE labels.

Additionally, label imbalance remains a major bottleneck. As seen in per-class F1 trends (Figure **??**), rare or abstract subcodes (e.g., `InviteCollaboration`, `ExploreOptions`) receive consistently poor predictions. Even with clause-level segmentation, the absence of surrounding discourse sometimes hinders the disambiguation of intent. Future work may address these challenges through regularization (e.g., edge dropout), data augmentation, or semi-supervised learning to expand coverage beyond current limitations.

Symbolic Graphs for Interpretable Reasoning. A core advantage of symbolic graphs lies in their alignment with interpretable reasoning. Compared to latent neural representations, graph-based models enable transparent tracing of structure: which roles were extracted, how clauses connect via `elaboration` or `contrast`, and why predictions differ across similar inputs. Such visibility supports explainable AI goals. For instance, ego-graphs make it possible to visualize prediction contexts, suggest alternate intents, or present clause groupings to clinicians. Interface-level explainability and graph-grounded user interaction are promising next steps.

Integration into Hybrid Clinical Systems. While our models focus on clause-level prediction, the resulting graphs can be embedded in broader clinical NLP workflows. For example, symbolic graphs may assist triage assistants in routing patient messages based on communicative function or enable longitudinal summarization by accumulating intent graphs over time. Recognized intents such as `AskQuestion` could also support decision support by triggering downstream retrieval or action planning. Incorporating speaker metadata, temporal flow, and dialogue state would further enable multi-turn reasoning across conversations.

Generalization and Generation.



(a) Narrative-Dependency (b) Narrative-AMR GCN (c) Narrative-SRL-Weighted GCN (Place holder)

(d) Narrative-Sentence MLP (Place holder) (e) Narrative-SRL-predicate (Place holder) (f) Narrative-SRL-anchored (Place holder)

Fig. 5: Per-class F1 score comparison across five representative models (with narrative). Each plot shows sorted F1 scores across 35 sub-CODE labels.

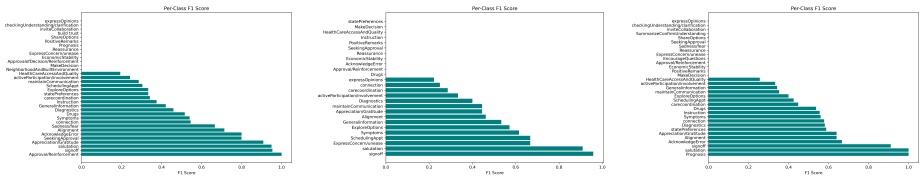
Future extensions span both classification and generation. On the classification side, graphs could be expanded across messages or sessions, linking clause-level structures into higher-order representations that model discourse continuity. On the generative side, symbolic graphs may act as scaffolds for data simulation, response generation, or narrative control. For instance, graph-guided prompting may allow symbolic transitions to drive response templates or steer large language models toward patient-centered outputs. This hybrid generation setting is a natural next step.

While this work focuses on clinical messaging, symbolic clause graphs could generalize to other domains: educational dialogue, narrative generation, or policy feedback; where short text spans encode nuanced communicative functions.

Summary. This study contributes toward interpretable, clause-level modeling of communicative intent in clinical messages. By evaluating multiple symbolic graph abstractions—syntactic, semantic, conceptual, and narrative—we demonstrate their role in supporting localized reasoning and hybrid model design. The symbolic graph framework shows promise as both a modeling tool and reasoning interface in sensitive, structure-rich domains.

9 Discussion

This study highlights both the promise and the limitations of symbolic graphs for clause-level intent modeling in clinical NLP. Symbolic structures improve modeling resolution, but challenges remain. The MSG-AMR-GCN model, for



(a) Context propagate (Place holder) (b) Context learned weight (Place holder) (c) Context edge testing (Place holder)

Fig. 6: Per-class F1 score comparison across five representative models (with context). Each plot shows sorted F1 scores across 35 sub-CODE labels.

instance, achieves strong training accuracy but suffers from overfitting. This likely stems from the limited dataset size, low semantic diversity, and sparse or overly specific graph structures—especially in narrative ego-graphs encoding idiosyncratic clause linkages.

Label imbalance further complicates learning. As shown in per-class F1 trends (Figure **??**), abstract or rare subcodes (e.g., `InviteCollaboration`) are consistently misclassified. Clause-level segmentation offers improved span alignment, but without broader discourse context, models still struggle with ambiguous or overlapping intents. Regularization strategies (e.g., edge dropout), data augmentation, or semi-supervised learning may mitigate these issues. Despite these limitations, symbolic graphs offer tangible benefits for interpretability. Unlike latent neural embeddings, they provide transparent structural cues: which semantic roles were extracted, how clauses relate via `elaboration` or `contrast`, and why intent predictions diverge. This transparency supports explainable AI goals, enabling clinicians to inspect, refine, or interact with model decisions through graph-grounded interfaces.

Beyond classification, symbolic graphs are promising building blocks for hybrid clinical systems. Graphs could aid triage assistants in message routing, support longitudinal summarization by tracking intent over time, or trigger downstream decisions (e.g., surfacing answers to patient questions). Enriching graphs with speaker roles, temporal flow, and dialogue state would support multi-turn discourse modeling. Looking forward, symbolic graphs may also enable generative tasks. Clause-level graphs could serve as scaffolds for training data simulation, narrative synthesis, or personalized response generation. For example, graph-guided prompting could steer large language models toward more structured, patient-centered outputs. The framework also generalizes beyond clinical messaging to other domains such as education, policy, or short-form narrative dialogue, where communicative structure matters.

In sum, symbolic clause-level graphs offer a lightweight, interpretable scaffold for reasoning in sensitive, structure-rich tasks. Their integration into hybrid neuro-symbolic systems holds promise for both accurate prediction and human-aligned explanation.

10 Conclusion

- Summary of prototype contributions
- Value for clinical AI and hybrid reasoning systems
  - Summary of findings
- Link to LLM-based reasoning, narrative modeling, and proposal's broader goals

11 Conclusion

11.1 Summary of Findings

This paper presents a comparative evaluation of symbolic graph structures for clause-level intent classification in patient-provider messages. We introduce multiple graph-based models grounded in syntactic, semantic, and narrative abstractions, and benchmark them against flat baselines using sentence- and clause-level input. Our results show that symbolic graphs, particularly narrative ego-graphs enriched with AMR features, significantly outperform flat models, demonstrating improved capacity for fine-grained intent recognition.

Despite modest overall accuracy (around 55–57%), symbolic structures like SRL-weighted and AMR-enhanced narrative graphs consistently provide improved learning signals and richer representational grounding. However, we also observe overfitting risks in highly expressive models (e.g., MSG-AMR-GCN), revealing trade-offs between model capacity and generalizability in low-resource clinical settings.

Our work contributes to the growing literature on hybrid neuro-symbolic modeling by showing that symbolic graphs can serve as interpretable scaffolds for sub-sentential reasoning in medical dialogue. We demonstrate the feasibility of clause-centered modeling using lightweight symbolic structures and highlight their compatibility with Graph Neural Networks (GCNs). These methods enable localized, structure-aware classification of communicative intent, a critical step toward explainable AI in healthcare. By providing clause-aligned evaluation, comparing multiple abstraction levels, and identifying strengths and limitations across graph types, this paper offers new insights into the design space of structured neural models for clinical NLP. Future directions include graph-to-text generation, cross-turn modeling, and symbolic fusion with large language models for narrative understanding and decision support.

References

1. CLINC150. UCI Machine Learning Repository (2020), DOI: https://doi.org/10.24432/C5MP58
2. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)

3. Attardi, G., Sartiano, D., Simi, M.: Biaffine dependency and semantic graph parsing for enhanceduniversal dependencies. In: Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021). pp. 184–188 (2021)
4. Bunt, H., Alexandersson, J., Choe, J.W., Fang, A.C., Hasida, K., Petukhova, V., Popescu-Belis, A., Traum, D.: ISO 24617-2: A semantically-based standard for dialogue annotation. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 430–437. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), https://aclanthology.org/L12-1296/
5. Chen, Q., Du, J., Kim, S., Wilbur, W.J., Lu, Z.: Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. BMC Medical Informatics and Decision Making **20**, 1–10 (2020)
6. Chen, Y.C.: Structured graph representations for visual narrative reasoning: A hierarchical framework for comics. arXiv preprint arXiv:2506.10008 (2025)
7. Chen, Y.C., Jhala, A.: Cpst: Comprehension-preserving style transfer for multimodal narratives. arXiv preprint arXiv:2312.08695 (2023)
8. Chen, Y.C., Jhala, A.: Collaborative comic generation: Integrating visual narrative theories with ai models for enhanced creativity. arXiv preprint arXiv:2409.17263 (2024)
9. Cohen, K.B., Palmer, M., Hunter, L.: Nominalization and alternations in biomedical language. PloS one **3**(9), e3158 (2008)
10. Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., et al.: Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. arXiv preprint arXiv:1805.10190 (2018)
11. Dong, H., Mao, J., Lin, T., Wang, C., Li, L., Zhou, D.: Neural logic machines. arXiv preprint arXiv:1904.11694 (2019)
12. Fodeh, S., Li, T., Menczynski, K., Burgette, T., Harris, A., Ilita, G., Rao, S., Gemmell, J., Raicu, D.: Using machine learning algorithms to detect suicide risk factors on twitter. In: 2019 international conference on data mining workshops (ICDMW). pp. 941–948. IEEE (2019)
13. Fodeh, S.J., Brandt, C., Luong, T.B., Haddad, A., Schultz, M., Murphy, T., Krauthammer, M.: Complementary ensemble clustering of biomedical data. Journal of biomedical informatics **46**(3), 436–443 (2013)
14. Fodeh, S.J., Finch, D., Bouayad, L., Luther, S.L., Ling, H., Kerns, R.D., Brandt, C.: Classifying clinical notes with pain assessment using machine learning. Medical & biological engineering & computing **56**, 1285–1292 (2018)
15. Gao, Y., Huang, T.H., Passonneau, R.J.: Abcd: A graph framework to convert complex sentences to a covering set of simple sentences. arXiv preprint arXiv:2106.12027 (2021)
16. Gao, Y., Huang, T.H., Passonneau, R.J.: Learning clause representation from dependency-anchor graph for connective prediction. In: Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15). pp. 54–66 (2021)
17. Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E.T., Wu, J.T., Welt, J., Foote Jr, J., Moseley, E.T., Grant, D.W., Tyler, P.D., et al.: Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PloS one **13**(2), e0192360 (2018)

18. Haider Rizvi, S.M., Imran, R., Mahmood, A.: Text classification using graph convolutional networks: A comprehensive survey. ACM Computing Surveys (2025)
19. Hsu, I., Guo, X., Natarajan, P., Peng, N., et al.: Discourse-level relation extraction via graph pooling. arXiv preprint arXiv:2101.00124 (2021)
20. Huang, J., Wang, Y., Wang, Y., Dong, Y., Xiao, Y.: Relation aware semi-autoregressive semantic parsing for nl2sql. arXiv preprint arXiv:2108.00804 (2021)
21. Hui, B., Geng, R., Wang, L., Qin, B., Li, B., Sun, J., Li, Y.: S^2 sql: Injecting syntax to question-schema interaction graph encoder for text-to-sql parsers. arXiv preprint arXiv:2203.06958 (2022)
22. Jagannatha, A.N., Yu, H.: Structured prediction models for rnn based sequence labeling in clinical text. In: Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing. vol. 2016, p. 856 (2016)
23. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)
24. Kacupaj, E., Plepi, J., Singh, K., Thakkar, H., Lehmann, J., Maleshkova, M.: Conversational question answering over knowledge graphs with transformer and graph attention networks. arXiv preprint arXiv:2104.01569 (2021)
25. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: DailyDialog: A manually labelled multi-turn dialogue dataset. In: Kondrak, G., Watanabe, T. (eds.) Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 986–995. Asian Federation of Natural Language Processing, Taipei, Taiwan (Nov 2017), https://aclanthology.org/I17-1099/
26. Liang, C., Berant, J., Le, Q., Forbus, K.D., Lao, N.: Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. arXiv preprint arXiv:1611.00020 (2016)
27. Liu, J., Zhang, Z., Razavian, N.: Deep ehr: Chronic disease prediction using medical notes. In: Machine Learning for Healthcare Conference. pp. 440–464. PMLR (2018)
28. Lopez, K., Fodeh, S.J., Allam, A., Brandt, C.A., Krauthammer, M.: Reducing annotation burden through multimodal learning. Frontiers in big Data **3**, 19 (2020)
29. Lorenzo, A.C.M., Maru, M., Navigli, R.: Fully-semantic parsing and generation: The babelnet meaning representation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1727–1741 (2022)
30. Luo, X., Gandhi, P., Zhang, Z., Shao, W., Han, Z., Chandrasekaran, V., Turzhitsky, V., Bali, V., Roberts, A.R., Metzger, M., et al.: Applying interpretable deep learning models to identify chronic cough patients using ehr data. Computer Methods and Programs in Biomedicine **210**, 106395 (2021)
31. Naseem, T., Ravishankar, S., Mihindukulasooriya, N., Abdelaziz, I., Lee, Y.S., Kapanipathi, P., Roukos, S., Gliozzo, A., Gray, A.: A semantics-aware transformer model of relation linking for knowledge base question answering. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 256–262 (2021)
32. Naseem, U., Musial, K., Eklund, P., Prasad, M.: Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding. In: 2020 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2020)
33. Nguyen, P., Tran, T., Wickramasinghe, N., Venkatesh, S.: Deepr: A convolutional net for medical records. arxiv. org (2016)

34. Nie, L., Cao, S., Shi, J., Sun, J., Tian, Q., Hou, L., Li, J., Zhai, J.: Graphq ir: Unifying the semantic parsing of graph query languages with one intermediate representation. arXiv preprint arXiv:2205.12078 (2022)
35. Procopio, L., Tripodi, R., Navigli, R.: Sgl: Speaking the graph languages of semantic parsing via multilingual translation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 325–337 (2021)
36. Rousseau, F., Kiagias, E., Vazirgiannis, M.: Text categorization as a graph classification problem. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1702–1712 (2015)
37. Sakai, H., Lam, S.S.: Large language models for healthcare text classification: A systematic review. arXiv preprint arXiv:2503.01159 (2025)
38. Samuel, D., Barnes, J., Kurtz, R., Oepen, S., Øvrelid, L., Velldal, E.: Direct parsing to sentiment graphs. arXiv preprint arXiv:2203.13209 (2022)
39. Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., Gonzalez, G.: Utilizing social media data for pharmacovigilance: a review. Journal of biomedical informatics **54**, 202–212 (2015)
40. Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., Okruszek, L.: Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research **304**, 114135 (2021)
41. Shervashidze, N., Schweitzer, P., Van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-lehman graph kernels. Journal of Machine Learning Research **12**(9) (2011)
42. Weißenhorn, P., Donatelli, L., Koller, A.: Compositional generalization with a broad-coverage semantic parser. In: Proceedings of the 11th Joint Conference on Lexical and Computational Semantics. pp. 44–54 (2022)
43. Wu, S., Chen, B., Xin, C., Han, X., Sun, L., Zhang, W., Chen, J., Yang, F., Cai, X.: From paraphrasing to semantic parsing: Unsupervised semantic parsing via synchronous semantic decoding. arXiv preprint arXiv:2106.06228 (2021)
44. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.: Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. Advances in neural information processing systems **31** (2018)
45. Yuan, R., Wang, Z., Li, W.: Event graph based sentence fusion. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 4075–4084 (2021)
46. Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., Chen, J.: Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In: Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020. pp. 109–117 (2020)
47. Zeng, Q.T., Redd, D., Divita, G., Jarad, S., Brandt, C., Nebeker, J.R.: Characterizing clinical text and sublanguage: A case study of the va clinical notes. J Health Med Informat S **3**(2) (2011)
48. Zheng, C., Chen, X., Xu, R., Chang, B.: A double-graph based framework for frame semantic parsing. arXiv preprint arXiv:2206.09158 (2022)
49. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI open **1**, 57–81 (2020)

Appendix

A.1 Graph Construction Details

We summarize the construction process for all graph variants used in our experiments:

- **Dependency Graph (DEP-GCN)**: Token-level syntactic dependency graphs were extracted using `spaCy`'s dependency parser. Each clause was parsed independently to ensure alignment with sub-sentential units.
- **Semantic Role Graphs (SRL-GCN)**: Three variants were created from predicate-argument structures. The weighted variant emphasized salient roles; the predicate-centric version linked each predicate to its arguments; the anchored version centered the clause and attached all relevant semantic roles.
- **AMR Graphs (AMR-GCN)**: Clause-level AMR graphs were obtained using the SPRING model in `amrlib`. These conceptual graphs represent predicate-argument structures at an abstract level, encoding deep semantics.
- **Narrative Graphs (MSG-AMR-GCN, MSG-MLP)**: Clause-centered ego-graphs were created using predefined symbolic relations such as `next`, `elaboration`, and `contrast`. MSG-AMR-GCN integrates AMR-based features, while MSG-MLP flattens node features for non-structural input.

A.2 Additional Graph Visualizations

To support interpretability and highlight design differences across graph representations, we present example visualizations of the symbolic and linguistic structures used in our experiments. Figures 7–11 depict clause-level graphs from the dataset, illustrating their construction logic, relation types, and connection patterns.

A.3 Training Curves for All Models

This subsection reports full training curves for all models evaluated in our experiments. Each figure shows the progression of training accuracy, macro-F1 score, and loss across epochs, helping to assess model convergence, generalization, and potential overfitting.

A.4 Per-Class Scores

To better understand class-specific performance, we report a heatmap of misclassification patterns and selected per-class F1 scores. The heatmap shows error distribution by true label across test instances, with each row corresponding to a gold subcode label and annotated by its frequency. Color intensity reflects the number of incorrect predictions, revealing patterns of class imbalance and model-specific failure modes.

In addition, we provide raw per-class F1 scores for all 35 subcode classes across selected models in Table 9.

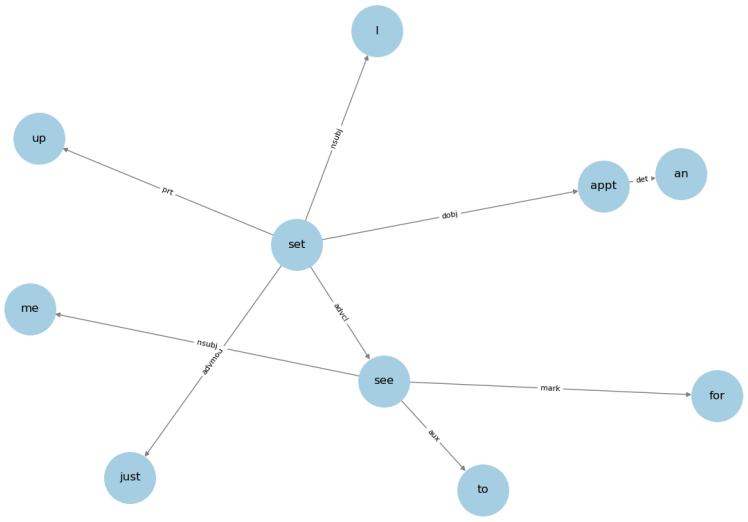


Fig. 7: Syntactic dependency graph of a clause, with grammatical edges.

Table 9: Per-class F1 scores for selected models (abbreviated sample). Full version available in supplementary material.

Subcode	MSG-AMR-GCN	SRL-GCN	DEP-GCN	Clause-MLP	Sentence-MLP
`signoff`	0.88	0.76	0.79	0.72	0.60
`schedulingAppt`	0.72	0.65	0.62	0.60	0.54
`diagnostics`	0.68	0.60	0.59	0.58	0.49
`exploreOptions`	0.33	0.29	0.24	0.22	0.21

A Investigation So Far and Possible Steps

[11pt]article graphicx booktabs geometry margin=1in [T1]fontenc subcaption makecell amsmath

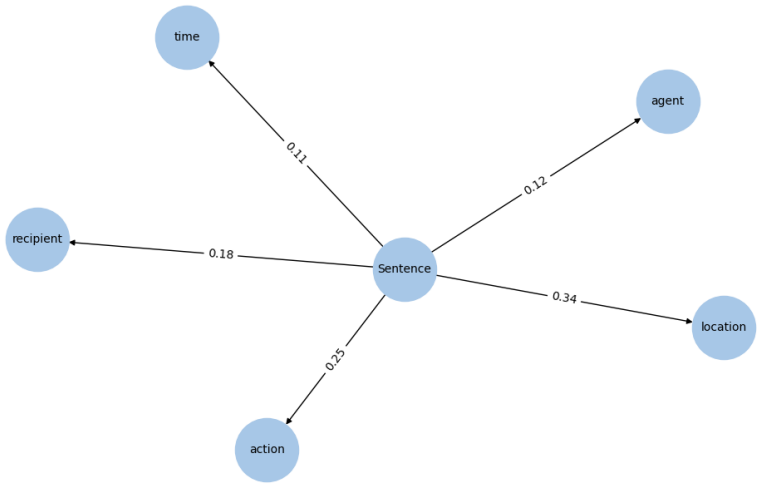


Fig. 8: SRL role-weighted focus graph with edges to salient arguments.

Preliminary Analysis of Clause-Level Classification and Graph-Based Methods

Yi-Chun (Rimi) Chen

No Institute Given

This memo summarizes quick analyses of the current clause-level classification dataset and the performance of graph-based methods. The goal is to assess (1) dataset limitations, (2) the feasibility of graph-based methods for the defined task, and (3) potential directions.

A Research Questions

This report aims to identify potential bottlenecks in clause-level classification and assess the feasibility of graph-based methods. We investigate the following questions:

A.1 Task Definition

1. The original task is to classify sub-sentential units (clauses) into hierarchical codebook labels.

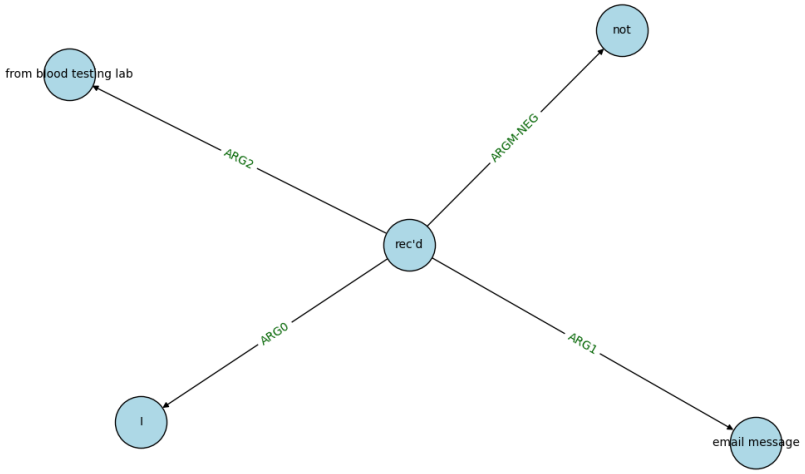


Fig. 9: Predicate-centered SRL graph using AllenNLP outputs.

A.2 Dataset Limitations

2.1 Is data imbalance a primary factor limiting performance?
2.2 Do the labels conflate interactional and goal-oriented semantics, reducing separability?
2.3 Can the assigned label be reliably inferred from clause content alone?
2.4 Do annotation spans align with syntactic or semantic clause boundaries?
2.5 Given observed noise and imbalance, is there a theoretical upper bound on achievable accuracy (e.g., 80–90%) under the current schema?

A.3 Graph-Based Methods

3.1 Are graph-based methods feasible for clause-level intent classification under current data conditions?
3.2 How do different graph abstractions (syntactic, semantic, conceptual, narrative) perform on this task (answered in previous report)?
3.3 How can we assess graph sparsity and its effect on learning?
3.4 Can graph models learn effectively in the presence of noisy or inconsistent labels?

A.4 Future Directions

4.1 Could benchmarking on cleaner public datasets clarify the general utility of graph-based NLP abstractions?
4.2 Should we redefine tasks (e.g., relation prediction, implicit intent inference) to better align with the strengths of graph-based reasoning rather than direct classification?

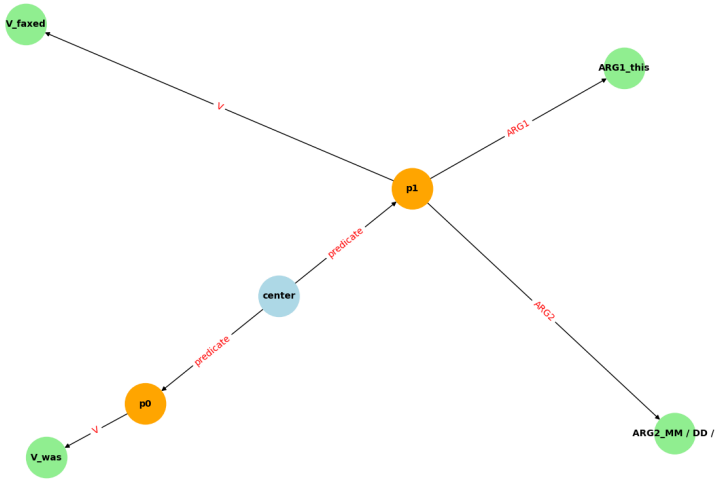


Fig. 10: Example SRL-anchored graph structure.

B Dataset Characteristics

We examined the annotation quality of the lab dataset for clause-level classification. Several structural issues were observed that affect model training and interpretation.

- **Label Imbalance:** The distribution of annotated codes and subcodes is highly uneven. A small number of frequent categories dominate the dataset, while many other categories occur rarely. This imbalance biases models toward majority classes and produces unstable performance on low-frequency classes.
- **Mixed Semantics:** The current label set conflates different semantic axes. Some labels describe *interactional functions*, while others capture *goal-oriented clinical intents*. Mixing communicative actions with content-specific intents introduces heterogeneity that is difficult for a single classifier to model consistently.
- **Label Ambiguity:** The same or very similar clause content can be annotated with different labels. For example, short status updates or medication mentions may be categorized under multiple codes depending on annotator interpretation. This ambiguity reduces the separability of label classes and increases confusion between overlapping categories.
- **Span Inconsistency:** Annotated spans do not always align with syntactic clauses. Some annotations cover full sentences, while others mark only partial phrases or single tokens. This inconsistency in segmentation makes it difficult to train clause-level models, since the effective unit of supervision varies across examples.

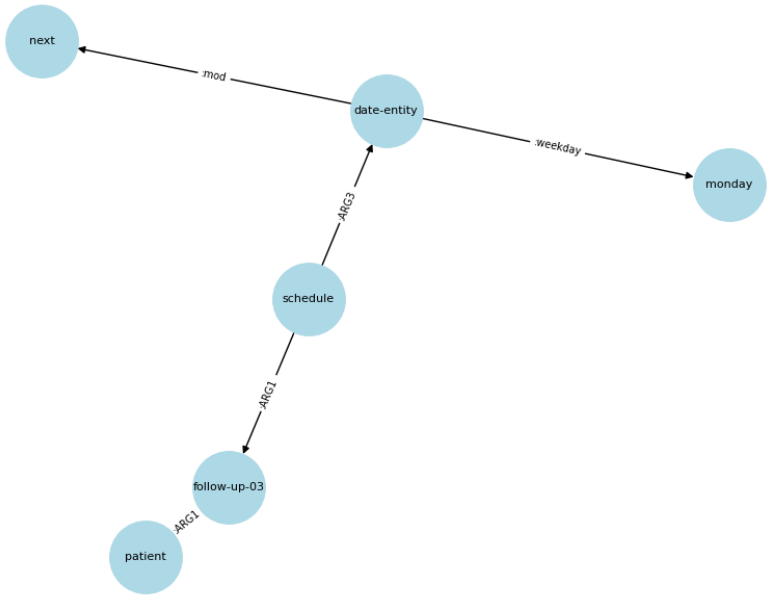


Fig. 11: Conceptual AMR graph representing a clause with abstract roles.

A.1 Label Imbalance

Annotation counts across labels exhibit substantial skew at all levels of granularity (code, subcode, and combined). Figures 1–3 show the frequency distributions, which follow a long-tailed pattern: a few classes dominate while many appear only rarely.

At the **code level** (Figure 1), there are 9 categories in total, but the imbalance ratio ($\mathrm{IR}_{\max}$) is 34.1, meaning the most common class (*Information-Giving*, $n = 887$) occurs over 34 times more frequently than the rarest class ($n = 26$). The entropy-based effective number of classes is only about $K_{eff} = 4.1$, which is less than half of the nominal 9 classes. This indicates that, in practice, the distribution behaves as though only four categories carry substantial weight.

At the **subcode level** (Figure 2), the imbalance becomes more pronounced: 35 categories are present, but $\mathrm{IR}_{\max}$ rises to 166.5, with the most frequent subcode (*salutation*, $n = 333$) dominating over extremely sparse classes ($n = 2$). The effective number of classes is $K_{eff} = 16.1$, less than half of the total. This reflects that while many subcodes exist, fewer than twenty play a significant role in shaping the distribution.

The **combined code–subcode distribution** (Figure 3) contains 50 unique labels. Here $\mathrm{IR}_{\max}$ is 89.5, with the majority class (*PATIENT_PARTNERSHIP_salutation*, $n = 179$) still vastly outweighing rare ones ($n = 2$). Entropy-based measures show normalized entropy of 0.878, suggesting moderate diversity, but the Gini index (0.959) confirms severe skew. The effective number of classes is $K_{eff} = 24.4$,

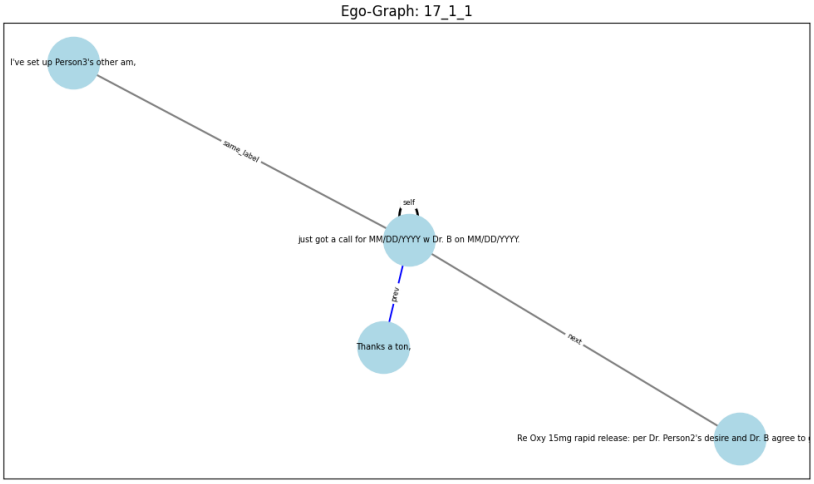


Fig. 12: Symbolic Narrative Ego-Graph

still less than half of the 50 categories, implying that the functional label space is effectively compressed to fewer than 25 categories.

Overall, these statistics indicate that the dataset is **highly imbalanced**. The sharp reduction in effective class space at each level suggests that models trained on this dataset may converge on patterns dominated by majority classes, while minority labels provide little statistical signal. This imbalance poses challenges for clause-level classification: models may achieve acceptable overall accuracy by focusing on frequent categories, but minority labels will suffer from poor recall and unstable learning. This calls for strategies such as resampling, loss reweighting, or hierarchical modeling to mitigate underperformance on rare but semantically important labels.

Imbalance Metrics To quantify dataset skew, we report several standard imbalance metrics. Let n_i denote the number of samples in class i, $N = \sum_i n_i$ the total number of samples, and K the number of classes.

Imbalance Ratio (IR_{max}). The maximum imbalance ratio is defined as:

$$IR_{max} = \frac{\max_i n_i}{\min_i n_i}.$$

It measures how many times more frequent the largest class is compared to the smallest. Larger values indicate more extreme imbalance.

Coefficient of Variation (CV). The coefficient of variation of class frequencies is:

$$CV = \frac{\sigma(n)}{\mu(n)},$$

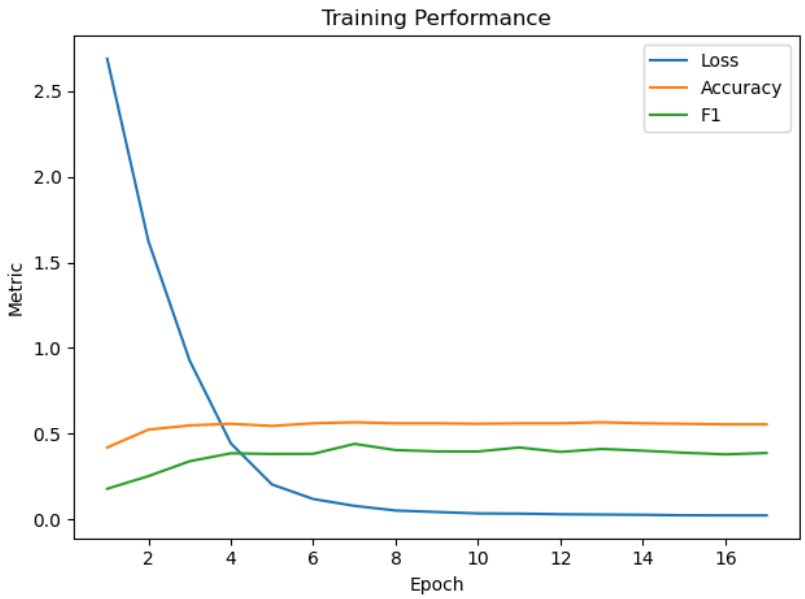


Fig. 13: Training curve for sentence-level MLP.

where $\sigma(n)$ and $\mu(n)$ are the standard deviation and mean of class counts, respectively. Higher CV indicates more dispersion and less uniformity across classes.

Shannon Entropy. Entropy quantifies the uncertainty of the label distribution:

$$H = -\sum_{i=1}^{K} p_i \log p_i, \quad p_i = \frac{n_i}{N}.$$

We also compute normalized entropy $H_{norm} = H/\log K$, which ranges from 0 (single-class dominance) to 1 (perfectly uniform).

Gini Index. The Gini index is given by:

$$G = 1 - \sum_{i=1}^{K} p_i^2.$$

It measures inequality in the distribution. Values close to 1 indicate strong imbalance, while values closer to 0 indicate uniformity.

Effective Number of Classes. Following the entropy-based effective cardinality:

$$K_{eff} = e^H,$$

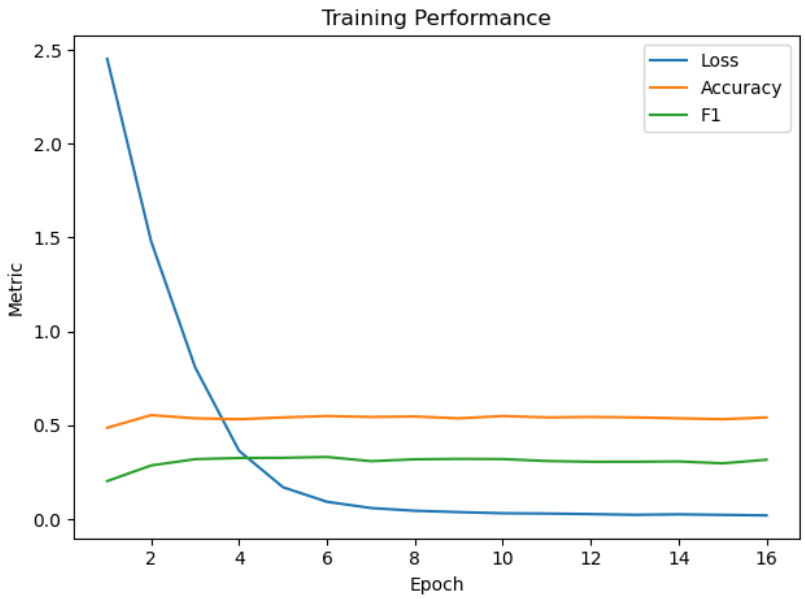


Fig. 14: Training curve for clause-level MLP.

This metric represents the number of equally probable classes that would yield the same entropy as the observed distribution. Smaller values relative to K indicate that only a subset of classes dominates the dataset.

Together, these metrics provide complementary perspectives: $\mathrm{IR}_{\max}$ highlights extremes, CV captures overall spread, entropy and Gini measure diversity and inequality, and K_{eff} summarizes the effective class space.

A.2 Mixed Semantics

Interactional vs. Goal-Oriented Label Balance Beyond overall imbalance, we examined whether the dataset exhibits systematic skew between two broad functional categories of intent: **Interactional** (relational, socioemotional, or partnership-oriented functions) and **Goal-Oriented** (task- or content-driven functions). This split is motivated by prior work in communication studies, which distinguishes surface-level conversational moves from deeper semantic or goal-driven content.

Figure 4 and Figure 5 show the label distributions after the split, while Table 1 summarizes imbalance metrics compared against the overall distribution.

Findings.

– **Interactional labels** are frequent but shallow. The distribution is dominated by a few formulaic categories such as *salutation* and *signoff*. This

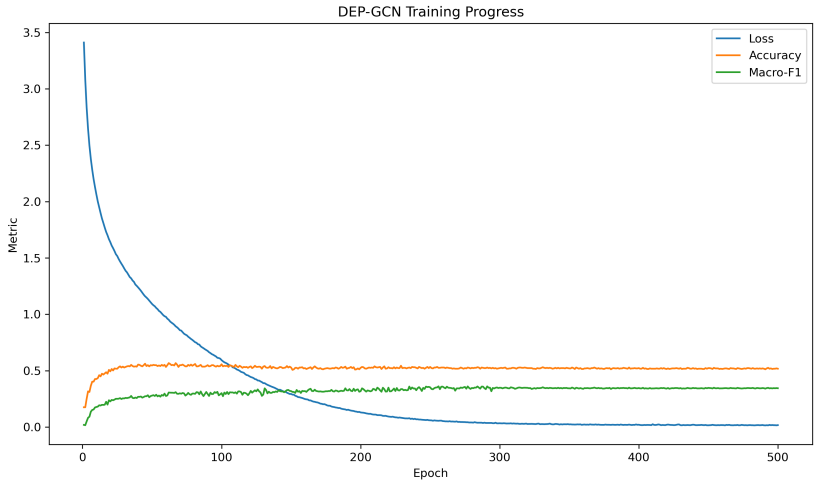


Fig. 15: Training curve for DEP-GCN.

Table 1: Imbalance metrics before and after splitting by Interactional vs. Goal-Oriented categories.

Subset	**Classes**	$\mathbf{IR_{max}}$	**CV**	H_{norm}	K_{eff}
Overall (Subcode)	35	166.5	1.08	0.851	16.1
Interactional (Flat)	29	421.5	1.80	0.739	6.9
Goal-Oriented (Flat)	10	34.2	0.63	0.883	7.1

yields an extreme imbalance ($IR_{max} = 421.5$) and a very low effective number of classes ($K_{eff} = 6.9$ out of 29). In practice, this means that models may learn to rely heavily on surface-level cues, achieving high performance on frequent classes while ignoring rare but socially meaningful interactional acts.

– **Goal-Oriented labels** are sparser but more balanced. The imbalance ratio is far lower ($IR_{max} = 34.2$), and the effective number of classes ($K_{eff} = 7.1$ out of 10) indicates that most goal-oriented labels contribute meaningfully to the distribution. However, their absolute counts remain low, which limits learning despite their relative balance.

Interpretation. The split does not reduce overall imbalance, but it reveals two distinct regimes: (1) interactional categories with heavy skew and redundancy, and (2) goal-oriented categories with healthier proportions but data scarcity. This suggests different modeling strategies: lightweight or rule-based methods may suffice for detecting interactional intents, while goal-oriented prediction will likely require augmentation, transfer learning, or hierarchical grouping to overcome sparsity.

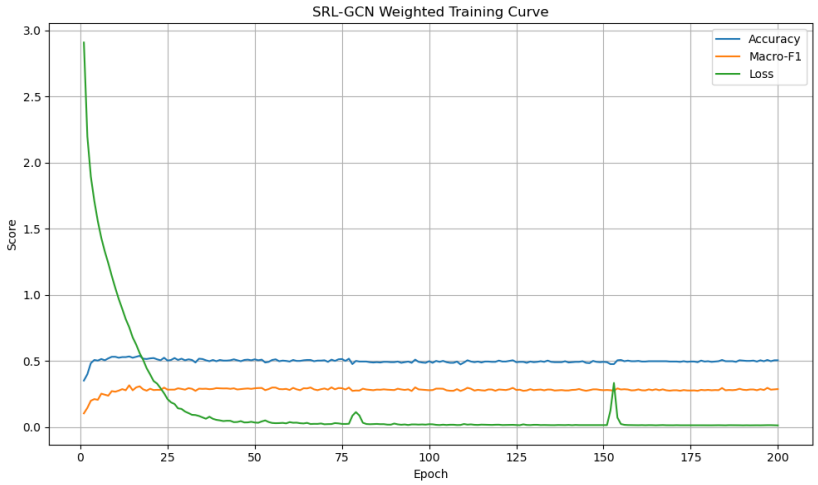


Fig. 16: Training curve for SRL-GCN-weighted.

A.3 Label Ambiguity

Beyond frequency imbalance, another challenge arises from **label ambiguity**. Semantically similar clauses are sometimes annotated with different intent categories, reflecting overlapping schema definitions or annotator subjectivity. This inconsistency limits model performance by introducing noise in the supervision signal.

To examine ambiguity, we embedded each annotated clause using a sentence-level encoder (`SentenceTransformer`) and performed nearest-neighbor analysis. For each clause, we retrieved the top-k neighbors in embedding space and compared their assigned labels.

Quantitative findings. Three diagnostics were used to quantify ambiguity:

- **Cross-label nearest neighbors.** On average, 37% of a clause's top-5 neighbors had a different label, with a median of 20%. This indicates that many semantically close clauses are labeled inconsistently.
- **Cluster purity.** Clustering all clauses into 35 groups (matching the number of subcodes) produced a purity of 0.51, meaning only half of cluster members shared the same label. This reflects moderate alignment between embeddings and annotations, suggesting that the label schema is not cleanly separable.
- **Confusion pairs.** Frequent neighbor conflicts concentrated on semantically overlapping or formulaic categories. For example, *Appreciation/Gratitude* vs. *Signoff* accounted for nearly 500 conflicts, while task-oriented overlaps such as *Care Coordination* vs. *Scheduling Appointment* and *Diagnostics* vs. *Scheduling Appointment* also appeared frequently.

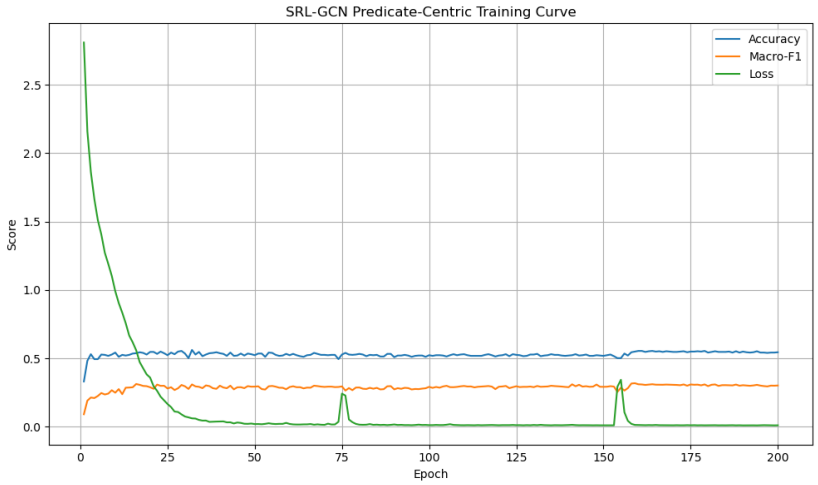


Fig. 17: Training curve for SRL-GCN-anchored.

Table 2: Most frequent ambiguous label pairs identified via nearest-neighbor analysis.

Label A	**Label B**	**Conflict Count**
Appreciation/Gratitude	Signoff	497
Care Coordination	Scheduling Appointment	114
Diagnostics	Scheduling Appointment	111
Diagnostics	Active Participation/Involvement	110
Connection	Signoff	72

Interpretation. The analysis reveals that the dataset's label space is inherently noisy: (i) Interactional labels such as *Connection*, and *Signoff* are highly formulaic and sometimes interchangeable, (ii) Goal-oriented labels (e.g., *Care Coordination*, *Scheduling*, *Diagnostics*) carry more semantic depth but blur at their boundaries. These ambiguities partly explain why classification performance plateaus around 50% accuracy. Addressing this issue may require schema refinement (e.g., merging overlapping subcodes), hierarchical intent modeling, or soft-labeling strategies that reflect uncertainty rather than enforcing hard categorical distinctions.

A.4 Span Inconsistency

Another challenge arises from **span inconsistency** in the annotations. While intent labels are applied to text spans, these spans vary widely in length, and their alignment with syntactic clauses is often irregular. Some annotations cover a single short phrase (e.g., "thanks"), while others span across multiple sentences or fragments that cross clause boundaries. This variability makes it difficult to

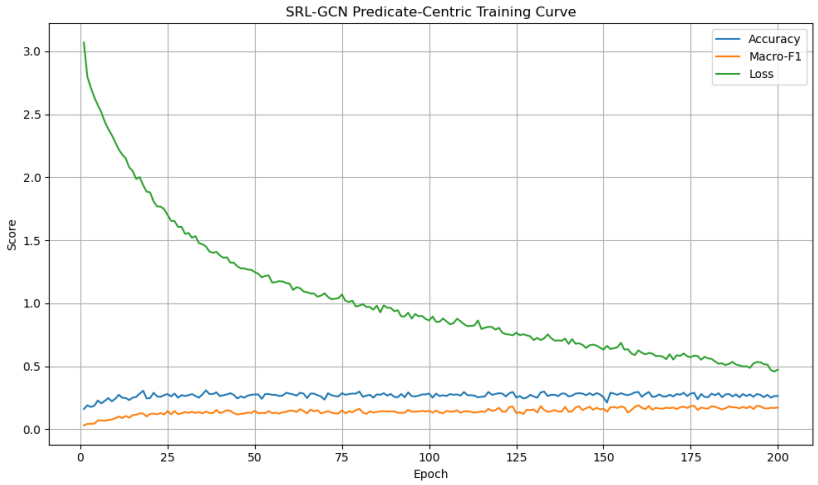


Fig. 18: Training curve for SRL-GCN-predicate.

establish a reliable clause-level unit of analysis and introduces noise into model training.

Quantitative findings. We measured span statistics across the dataset:

- **Length variation.** Spans range from 1 to 49 tokens, with a median length of 7 tokens and a mean of 8.2. In characters, spans range from 2 to 243 with a median of 32.5. The high coefficient of variation ($\sim$0.81 by tokens) indicates substantial heterogeneity in span lengths.
- **Clause alignment.** Span boundaries only loosely track clause segmentation. The median distance between annotated boundaries and syntactic clause boundaries is 1 token (left) and 1 token (right), but the mean misalignment is larger (2.5 tokens on the left, 4.8 on the right). This suggests that while some spans align cleanly, many extend beyond or cut across clause boundaries.
- **Distributional skew.** Short spans are highly frequent, reflecting formulaic expressions (e.g., greetings, acknowledgments), whereas long spans appear in more complex categories, often bundling multiple communicative intents.

Interpretation. These results suggest that span definitions in the current schema are loosely specified, leading to heterogeneous annotation practices. For clause-level modeling, this inconsistency weakens the reliability of labels and may contribute to reduced classification performance. In particular, short spans dominated by certain labels yield highly imbalanced training examples that models can memorize with little generalization value, while very long spans conflate multiple intents into a single label, obscuring the decision boundary. Misaligned

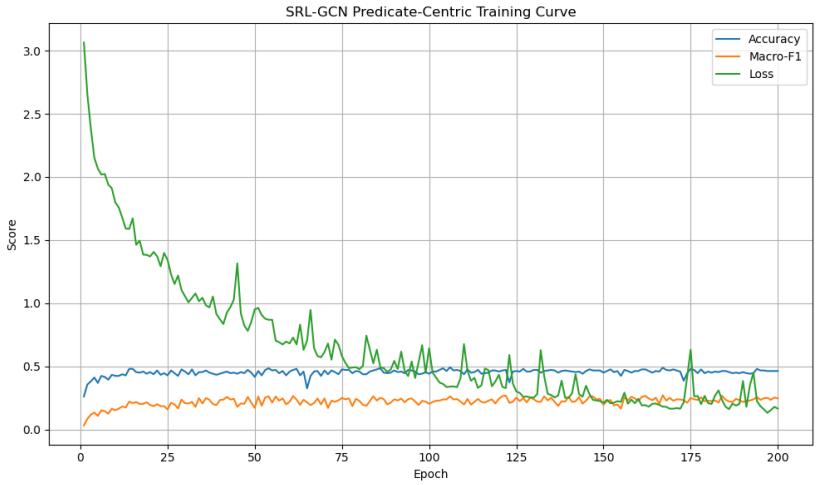


Fig. 19: Training curve for AMR-GCN.

spans further degrade structural models such as dependency- or clause-based GCNs, since the annotated unit does not correspond to a coherent linguistic structure. Together, these effects explain why downstream models show limited improvement from structural cues: the annotated units themselves are not consistent with linguistic segmentation. Possible remedies include normalizing annotations to clause boundaries, adopting multi-label schemes for longer spans, or introducing hierarchical segmentation to separate phrase-level from clause-level intents.

C Method Feasibility

A.1 Modeling Approaches: Clause-Level vs. Graph/Discourse Methods

The choice of modeling unit is central to intent classification in patient-provider communication. Two broad approaches have emerged: (1) direct clause-level or span-level classification using standard NLP models, and (2) graph- or discourse-based methods that explicitly encode syntactic and discourse relations.

Clause-level approaches. Starting with clause-level (sub-sentential) intent labels is a pragmatic and empirically supported strategy. Prior work on healthcare text classification has demonstrated that short utterances and single-intent spans can be modeled effectively using standard classifiers or fine-tuned transformers, with large language models now showing additional promise for this domain [37]. Clause-level classification provides a clear unit of supervision, avoids overfitting to noisy discourse relations, and establishes a reliable baseline. However, its effectiveness is contingent on a consistent annotation policy: heterogeneous span

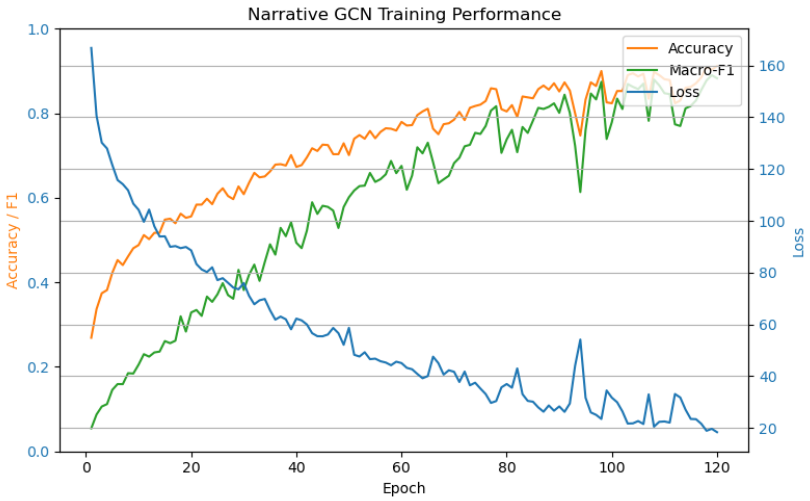


Fig. 20: Training curve for MSG-AMR-GCN.

lengths or multi-intent spans weaken label reliability and depress classifier performance regardless of architecture. Thus, clause-level methods are well-suited when spans are clean, intents are operationalized, and messages contain a dominant actionable intent.

Limits of clause-only modeling. Our span inconsistency analysis shows that many annotations either merge multiple clauses or misalign with syntactic boundaries, with mean offsets of 2.5-4.8 tokens from clause boundaries. Such irregularity reduces the discriminability of labels and obscures the local semantic cues that clause-level models rely on. In these cases, classifiers trained solely on isolated spans may confuse overlapping categories or fail to resolve context-dependent intents.

Why graph methods succeed in their original settings. Graph-based methods were originally developed for tasks that rely heavily on inter-clause or discourse-level relations. For example, dependency-anchor graphs enhance clause representations for predicting connectives like *because* or *although*, where syntactic anchors and clause-to-clause dependencies are central signals [16]. Similarly, discourse-level pooling architectures improve relation extraction by aggregating features across clauses and sentences, capturing long-range dependencies and rhetorical functions such as *cause-effect* or *elaboration* [19]. In such settings, graphs provide explicit structural cues that sequence-only models struggle to infer.

When graph/discourse methods help in intent classification. By contrast, clause-level intent classification in patient-provider communication often involves short, locally interpretable units (e.g., requests, acknowledgments, symptom reports)

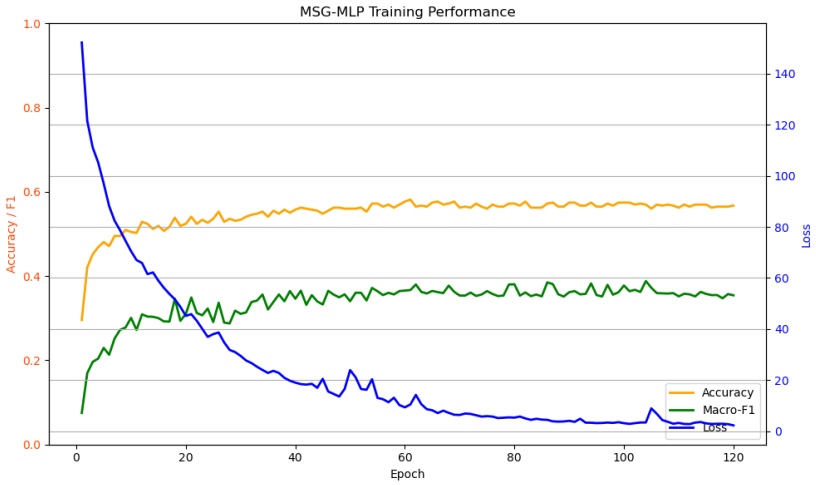


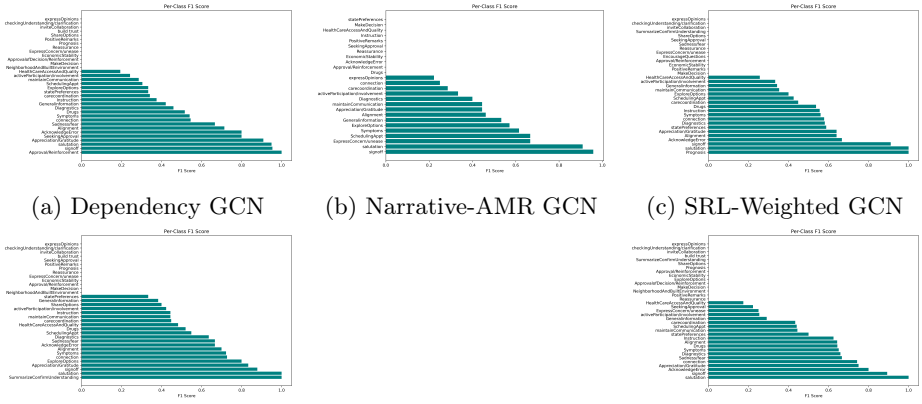
Fig. 21: Training curve for MSG-MLP.

where intents can be recovered from a single clause. Here, graph modeling is not strictly necessary unless annotation spans are inconsistent or discourse signals are critical. Nevertheless, graphs can add value in the following scenarios:

- **Multi-intent messages.** When a span bundles several communicative acts (e.g., reporting symptoms and requesting medication), graph structures can expose clause boundaries and support multi-label prediction.
- **Discourse-sensitive intents.** Distinguishing a follow-up question from a new request may hinge on rhetorical relations (*contrast*, *cause*, *elaboration*) that can be encoded as graph edges.
- **Cross-sentence dependencies.** Some provider responses reference prior patient clauses; discourse graphs help capture anaphora, temporal flow, or coherence across turns.
- **Syntactic enrichment.** Dependency anchors (e.g., subject-verb pairs) provide more stable units of meaning than noisy annotated spans, improving robustness when span boundaries are inconsistent [16].

In these contexts, discourse-aware graphs expand the receptive field of models, making it possible to recover coherence signals that clause-only methods overlook.

Practical implications. These considerations suggest a staged workflow. First, stabilize annotation guidelines and establish strong clause-level baselines with single-intent spans. Second, perform targeted error analysis to identify subsets of data where failures arise from discourse dependencies. Only in these cases should graph/discourse features be introduced, for example by enriching clause semantics with dependency anchors [16], adding local context windows, or modeling



(a) Dependency GCN (b) Narrative-AMR GCN (c) SRL-Weighted GCN

(d) Sentence MLP (e) Clause-level MLP

Fig. 22: Per-class F1 scores across five models. Each plot shows sorted F1 across 35 sub-CODE labels.

Table 3: Representative nearest-neighbor conflicts. Despite high semantic similarity, clauses receive divergent labels, illustrating annotation ambiguity.

Anchor Text	**Anchor Label**	**Neighbor Text**	**Neighbor Label**	**Sim.**
In person or telemed?	Invite Collaboration	We can either schedule for in person or via telemed ...	Share Options	0.74
Please send to	Instruction	I will also send it via email	Maintain Communication	0.71
Take care	Connection	Take care	Signoff	0.92
We can look into this together	Invite Collaboration	Let's review this together tomorrow	Active Participation/Involvement	0.68
I just sent in the refill	Drugs	The prescription refill has been sent	Maintain Communication	0.79

explicit rhetorical relations with pooling strategies [19]. This selective integration ensures that graph-based methods are applied where they provide measurable gains, while clause-level classification remains the foundation for reliable intent modeling.

A.2 Feasibility and Limitations of Graph-Based Methods

Graph-based representations were introduced to capture dependencies beyond surface lexical cues, offering interpretable structures for clause-level intent modeling. In principle, symbolic graphs provide advantages such as explicit relation modeling, localized context propagation (e.g., `next`, `prev`, `elaboration`), and the ability to query or manipulate reasoning paths. However, under the current task design and dataset conditions, their benefits were limited.

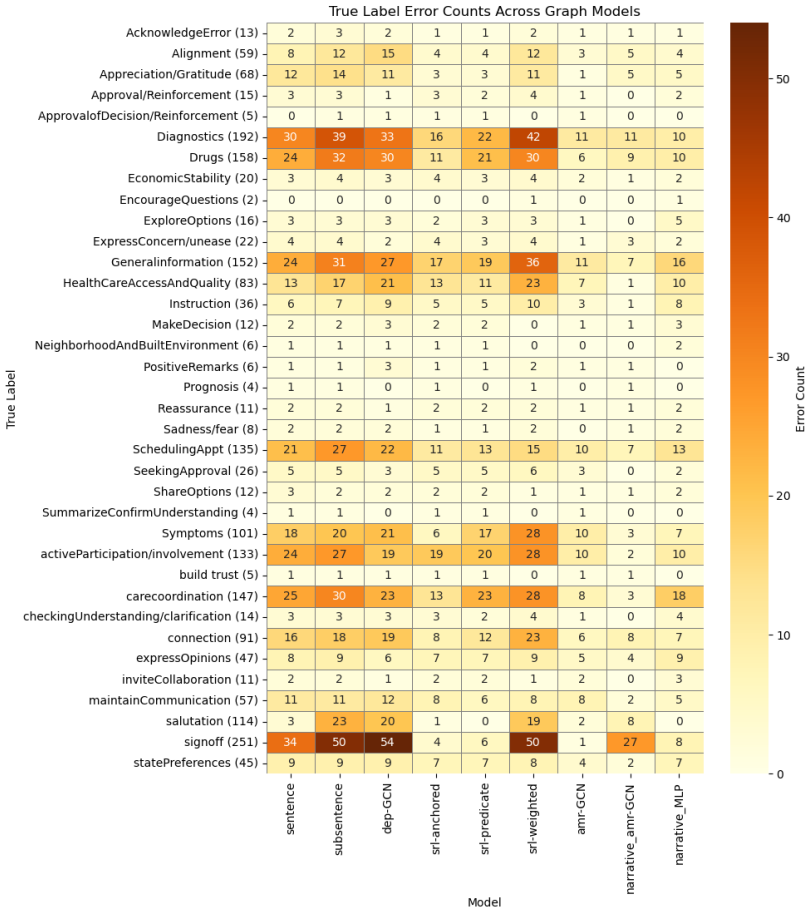


Fig. 23: Error heatmap by true label: darker shades indicate higher misclassification frequency.

Feasibility Under Current Task Settings Experimental results showed that graph-based models achieved performance comparable to simple baselines but did not yield substantial improvements. This outcome reflects the influence of noisy and inconsistent labels, which impose a ceiling on classification accuracy independent of the representational form. When label variability dominates, the additional structure encoded in graphs cannot manifest as performance gains.

To better ground the discussion of graph feasibility, we considered established measures of graph complexity and expressivity. Clause-level graphs in our dataset are structurally simple (few nodes/edges, low relational variety), which constrains the range of features that graph neural networks can exploit. Table 4 summarizes diagnostic metrics, from basic topological properties (node/edge counts, density, diameter) to algebraic measures (spectral diversity, graph en-

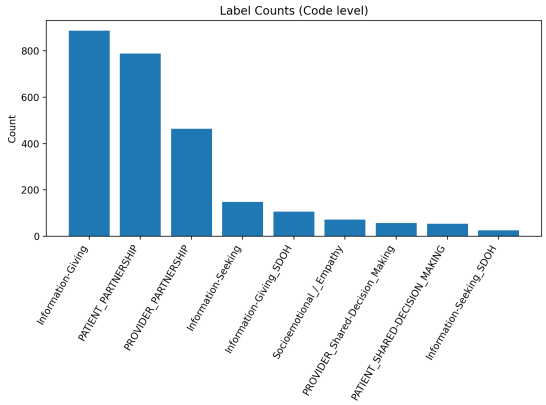


Fig. 1: Distribution of annotation counts at the code level.

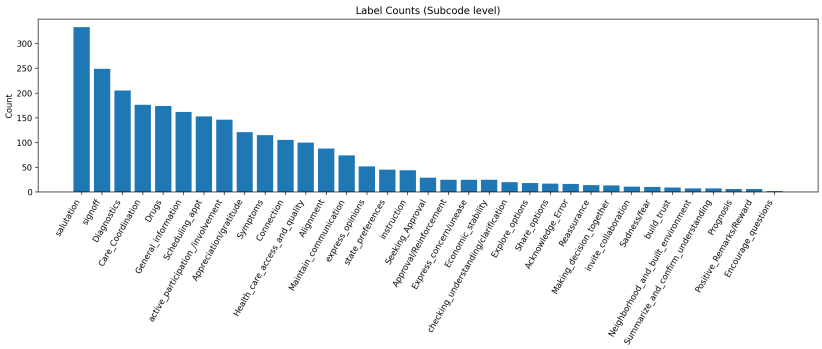


Fig. 2: Distribution of annotation counts at the subcode level.

ergy) and theoretical expressivity tests (e.g., Weisfeiler-Lehman refinement [41]). These metrics provide a principled way to quantify whether generated graphs are sufficiently rich to support meaningful learning. Preliminary statistics (average nodes $\sim$4, diameter $\sim$2, near-zero clustering coefficient) suggest that our clause-level graphs occupy the low-complexity regime, which helps explain the observed plateau in performance.

Collectively, these diagnostics suggest that clause-level graphs occupy the low-complexity regime, explaining why symbolic structure alone did not yield substantial performance improvements under the current task design.

Metric definitions and interpretation. To contextualize these results, we summarize the meaning of each metric and the implications of high vs. low values. Where applicable, we indicate typical thresholds observed in the graph learning literature.

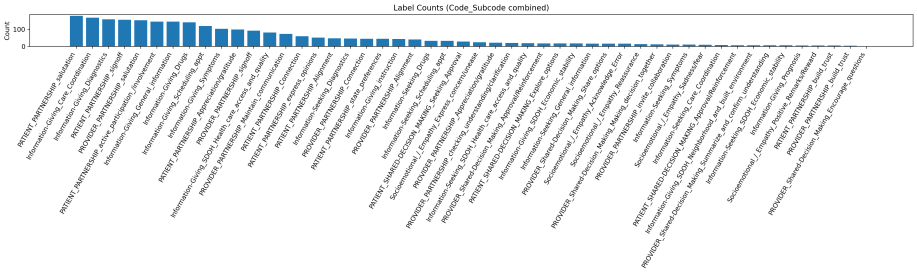


Fig. 3: Distribution of annotation counts at the combined code–subcode level.

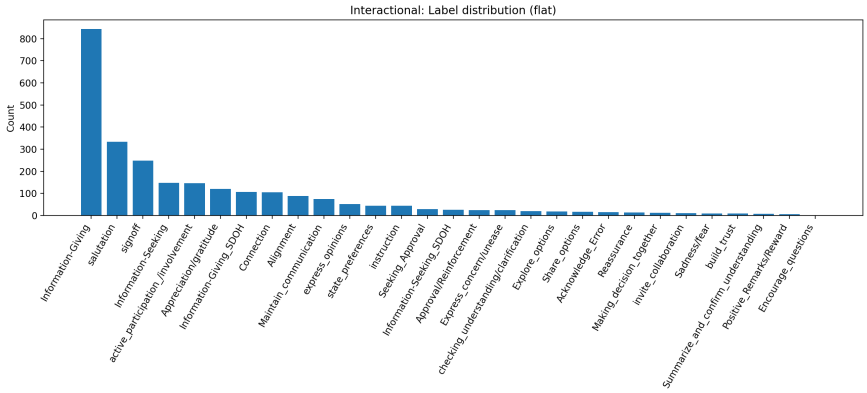


Fig. 4: Label distribution for the **Interactional** subset (flat). The distribution is dominated by a few formulaic classes, such as *salutation* and *signoff*, reflecting high skew.

- **Nodes / Edges:** Average graph size. Small graphs (<10 nodes) limit relational variety; larger graphs (>50 nodes) generally provide richer contexts.
- **Density / Average Degree:** Ratio of actual to possible edges, and mean degree per node. Low values (≤ 2 neighbors) indicate shallow neighborhoods where message passing collapses quickly. Moderate densities (0.2–0.5) with higher degrees (3–5) support richer propagation.
- **Diameter / Average Shortest Path Length (ASPL):** Longest and average shortest path lengths. Small diameters (1–3) mean limited scope for long-range information flow. Moderate diameters (∼5–10) allow multi-hop reasoning.
- **Clustering Coefficient:** Probability that neighbors of a node are connected. Near-zero clustering indicates absence of higher-order motifs. Values >0.1 typically indicate richer local structure.
- **Degree Entropy:** Shannon entropy of degree distribution. Low entropy (∼1) means most nodes have similar degree, reflecting structural uniformity.

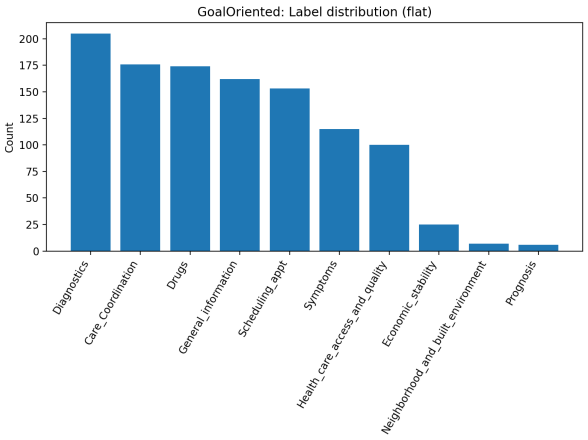


Fig. 5: Label distribution for the **Goal-Oriented** subset (flat). While absolute counts are smaller, the distribution is relatively more balanced across classes.

Higher entropy (>2) implies the presence of hubs or more diverse connectivity.

- **Node Label Entropy:** Diversity of node labels (semantic roles, clause types). Higher values reflect richer semantic scaffolds. Very low values (<1) suggest limited differentiation.
- **Edge Label Entropy:** Diversity of edge types. Zero entropy means all edges encode the same relation, offering no relational variety. Higher entropy (>1) is desirable to represent multiple discourse or syntactic relations.
- **Unique Laplacian Eigenvalues:** Number of distinct eigenvalues of the Laplacian spectrum. Larger numbers indicate richer structural diversity. Values close to the number of nodes suggest high variability; very low numbers imply trivial graphs.
- **Spectral Entropy:** Entropy of normalized Laplacian eigenvalue distribution. Higher values (>3 for medium graphs) correspond to greater structural complexity. Low values indicate near-regular or tree-like graphs.
- **Graph Energy:** Sum of absolute adjacency eigenvalues. Scales with graph size and complexity. Higher energy reflects richer connectivity; values below 10 often correspond to very small/sparse graphs.
- **WL Distinguishability:** Fraction of graphs uniquely identified under Weisfeiler–Lehman (WL) refinement [**?**]. High distinguishability (>80%) indicates structural or label diversity across graphs. Low values (<50%) suggest many graphs collapse to the same WL hash and are indistinguishable to standard GNNs.

Structural Simplicity of Clause-Level Graphs Clause-level graphs are inherently sparse: they contain relatively few nodes and edges, shallow connectivity, and limited relation types. This structural simplicity restricts the diversity of

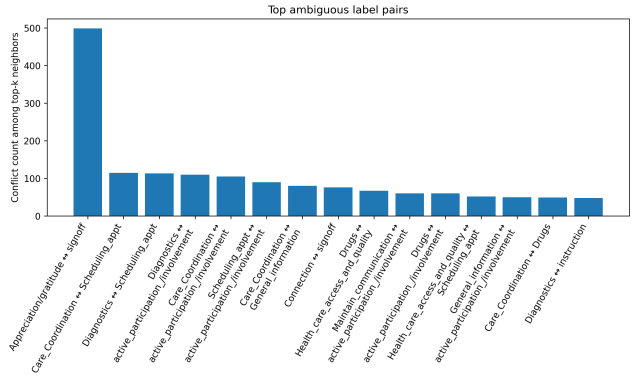


Fig. 6: Distribution of the most frequent ambiguous label pairs. Height indicates the number of nearest-neighbor conflicts between categories.

paths and features that graph convolutional layers can exploit. Compared with richer document- or discourse-level graphs, clause-level graphs have:

- Low node and edge counts, resulting in shallow neighborhoods.
- Small graph diameter, limiting long-range information flow.
- Restricted relational variety, with most edges representing only local adjacency.

Such constraints reduce expressive capacity and limit the potential benefit of graph-based reasoning at the clause level. *[Insert table/figure here: average node/edge counts, average degree, and graph diameter for generated clause-level graphs].*

Task Constraints and Implications The current classification tasks rely heavily on lexical signals, making them less sensitive to the structural cues that graphs provide. Graph-based methods are better aligned with reasoning-oriented tasks—such as discourse relation prediction, implicit intent inference, or message-level aggregation—where structural dependencies play a central role. This explains why clause-level classification shows limited gains, whereas more complex reasoning tasks may benefit from symbolic scaffolds.

Limitations Observed Two key limitations emerged: (1) structural simplicity restricted representational power at the clause level, and (2) the computational overhead of graph construction (dependency parsing, SRL, AMR) did not yield proportional accuracy improvements under noisy conditions. These findings highlight a mismatch between the promise of symbolic graphs and the constraints of the present task design.

Implications and Next Steps Future work should evaluate graph expressiveness more formally, using structural metrics such as degree distribution, clustering coefficient, or spectral gap. Incorporating richer relations (e.g., discourse

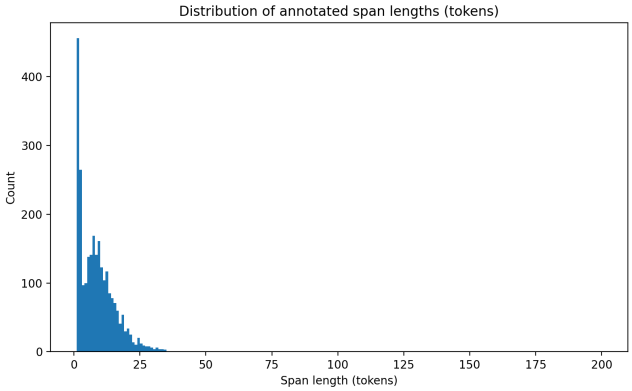


Fig. 7: Distribution of annotated span lengths. While most spans are short, a long tail of extended spans reflects annotation inconsistency.

markers, narrative roles) or multi-level graph abstractions (clause → sentence → document) may increase representational power. Moreover, shifting from flat classification to reasoning-oriented tasks could better showcase the advantages of symbolic graphs.

D Discussion

Our experiments highlight several important insights for modeling communicative intent in patient-provider messaging. First, the dominant barrier to performance is not model choice but data quality. The current corpus exhibits multiple challenges, including label imbalance, semantic overlap between categories, and inconsistent span segmentation. For example, pleasantries and interactional phrases are sometimes conflated with goal-oriented intents, while multi-intent clauses are forced into single labels. These factors depress both inter-annotator agreement and model discriminability, resulting in limited headroom for even strong neural architectures. In this context, graph-based methods cannot compensate for noise in the annotation schema; structural modeling adds complexity but does not resolve ambiguous supervision.

Second, while graph-based methods are theoretically attractive, our results show limited benefit when applied directly to clause-level intent classification. This is consistent with the design of many graph architectures, which were developed for tasks where explicit structural relations drive prediction—for instance, connective prediction [16] or discourse-level relation extraction [19]. By contrast, most patient-provider intents are locally expressed and can be classified from short spans without requiring discourse coherence. Graphs, therefore, offer little advantage unless the task explicitly requires reasoning across clauses or integrating multiple discourse cues.

Table 4: Graph complexity and expressivity metrics relevant for clause- and sentence-level intent graphs.

Metric	Definition	Interpretation for Expressivity
Nodes / Edges	Average number of nodes and edges per graph.	Very small graphs (e.g., 3-6 nodes) limit relational variety and reduce representational power.
Density / Degree	Ratio of edges to possible edges; average node degree.	Low density and low degree → shallow neighborhoods; message passing collapses quickly.
Diameter / Path Length	Longest and average shortest distance between nodes.	Small diameters (1-2) imply limited context propagation beyond local neighbors.
Clustering Coefficient	Probability that neighbors of a node are connected.	Near-zero clustering indicates lack of higher-order motifs; richer graphs show more relational closure.
Structural Entropy	Entropy of degree or connectivity distribution.	Low entropy = structurally impoverished graphs with few distinguishable patterns.
WL Test Distinguishability	Fraction of graphs distinguishable under Weisfeiler-Lehman refinement.	If many graphs collapse to the same WL hash, expressivity is low regardless of GNN architecture.
Spectral Diversity	Diversity of eigenvalues of Laplacian or adjacency matrix.	Collapsed eigenvalues suggest trivial structures; richer spectra → more structural signals.
Graph Energy	Sum of absolute adjacency eigenvalues.	Higher energy reflects more structural richness; low values indicate near-trivial graphs.
Label Entropy	Entropy of node/edge label distributions.	Low label entropy → limited semantic differentiation; high entropy → richer symbolic scaffolds.

Third, graphs appear more promising when reframed toward reasoning-oriented objectives rather than direct intent labeling. For example, predicting rhetorical relations between adjacent clauses, inferring missing or implicit intents, or aggregating clause-level predictions into message-level workflows all naturally benefit from structural representations. In these contexts, graph modeling can leverage syntactic anchors, discourse relations, or message-level coherence to enrich intent understanding beyond what span-level classifiers achieve.

E Potential Directions

Building on these observations, we identify three concrete directions for future work:

1. **Refine label design.** A key priority is to redesign the label schema to reduce ambiguity and better reflect clinical actionability. This may involve support-

Table 5: Graph complexity and expressivity metrics across different graph types. Placeholders "–" indicate values to be filled in.

Graph	Nodes	Edges	Density	WL (%)	Spectral Ent. / Energy
Dependency	7.75 ± 4.69	6.72 ± 4.66	0.37 ± 0.26	83.23	1.95 ± 1.01 / 7.76 ± 4.90
SRL-weighted	6.00 ± 0.00	5.00 ± 0.00	0.33 ± 0.00	83.42	1.49 ± 0.22 / 1.12 ± 0.17
SRL-predicate	2.95 ± 1.15	1.95 ± 1.15	0.68 ± 0.28	85.26	0.65 ± 0.58 / 2.62 ± 0.96
SRL-anchored	5.88 ± 3.94	5.03 ± 4.16	0.24 ± 0.19	83.42	1.55 ± 1.06 / 5.31 ± 4.22
AMR	4.37 ± 2.62	3.70 ± 3.00	0.45 ± 0.28	71.54	1.24 ± 0.89 / 4.58 ± 3.34

ing multi-label annotation for spans containing multiple communicative acts, or explicitly distinguishing between *interactional intents* and *goal-oriented intents*. Such refinements can improve annotation reliability and provide a clearer foundation for modeling.

2. **Explore reasoning-oriented tasks.** Rather than applying graphs solely for direct classification, future work should target tasks that require structured reasoning. Examples include predicting discourse relations between clauses, modeling the progression of intents across a conversation, or detecting when an implicit request is embedded within pleasantries. These tasks align more naturally with the strengths of graph/discourse methods and may yield stronger performance gains.
3. **Benchmark on public datasets.** To disentangle modeling capacity from dataset-specific noise, it is valuable to evaluate graph-based approaches on public corpora with cleaner labels and well-defined spans (e.g., DailyDialog, CLINC150). This not only enables comparison with prior work but also provides a controlled setting to validate whether graph architectures deliver measurable improvements when label definitions and span boundaries are stable.

Taken together, these directions suggest a staged trajectory: first, stabilize labels and span policies to strengthen clause-level baselines; second, incorporate graphs in settings where reasoning across discourse units is essential; and third, validate these methods on both private and public data to clarify their generalizability.

References

1. CLINC150. UCI Machine Learning Repository (2020), DOI: https://doi.org/10.24432/C5MP58
2. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
3. Attardi, G., Sartiano, D., Simi, M.: Biaffine dependency and semantic graph parsing for enhanceduniversal dependencies. In: Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021). pp. 184–188 (2021)

4. Bunt, H., Alexandersson, J., Choe, J.W., Fang, A.C., Hasida, K., Petukhova, V., Popescu-Belis, A., Traum, D.: ISO 24617-2: A semantically-based standard for dialogue annotation. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 430–437. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), https://aclanthology.org/L12-1296/
5. Chen, Q., Du, J., Kim, S., Wilbur, W.J., Lu, Z.: Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. BMC Medical Informatics and Decision Making **20**, 1–10 (2020)
6. Chen, Y.C.: Structured graph representations for visual narrative reasoning: A hierarchical framework for comics. arXiv preprint arXiv:2506.10008 (2025)
7. Chen, Y.C., Jhala, A.: Cpst: Comprehension-preserving style transfer for multimodal narratives. arXiv preprint arXiv:2312.08695 (2023)
8. Chen, Y.C., Jhala, A.: Collaborative comic generation: Integrating visual narrative theories with ai models for enhanced creativity. arXiv preprint arXiv:2409.17263 (2024)
9. Cohen, K.B., Palmer, M., Hunter, L.: Nominalization and alternations in biomedical language. PloS one **3**(9), e3158 (2008)
10. Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., et al.: Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. arXiv preprint arXiv:1805.10190 (2018)
11. Dong, H., Mao, J., Lin, T., Wang, C., Li, L., Zhou, D.: Neural logic machines. arXiv preprint arXiv:1904.11694 (2019)
12. Fodeh, S., Li, T., Menczynski, K., Burgette, T., Harris, A., Ilita, G., Rao, S., Gemmell, J., Raicu, D.: Using machine learning algorithms to detect suicide risk factors on twitter. In: 2019 international conference on data mining workshops (ICDMW). pp. 941–948. IEEE (2019)
13. Fodeh, S.J., Brandt, C., Luong, T.B., Haddad, A., Schultz, M., Murphy, T., Krauthammer, M.: Complementary ensemble clustering of biomedical data. Journal of biomedical informatics **46**(3), 436–443 (2013)
14. Fodeh, S.J., Finch, D., Bouayad, L., Luther, S.L., Ling, H., Kerns, R.D., Brandt, C.: Classifying clinical notes with pain assessment using machine learning. Medical & biological engineering & computing **56**, 1285–1292 (2018)
15. Gao, Y., Huang, T.H., Passonneau, R.J.: Abcd: A graph framework to convert complex sentences to a covering set of simple sentences. arXiv preprint arXiv:2106.12027 (2021)
16. Gao, Y., Huang, T.H., Passonneau, R.J.: Learning clause representation from dependency-anchor graph for connective prediction. In: Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15). pp. 54–66 (2021)
17. Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E.T., Wu, J.T., Welt, J., Foote Jr, J., Moseley, E.T., Grant, D.W., Tyler, P.D., et al.: Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PloS one **13**(2), e0192360 (2018)
18. Haider Rizvi, S.M., Imran, R., Mahmood, A.: Text classification using graph convolutional networks: A comprehensive survey. ACM Computing Surveys (2025)
19. Hsu, I., Guo, X., Natarajan, P., Peng, N., et al.: Discourse-level relation extraction via graph pooling. arXiv preprint arXiv:2101.00124 (2021)

20. Huang, J., Wang, Y., Wang, Y., Dong, Y., Xiao, Y.: Relation aware semi-autoregressive semantic parsing for nl2sql. arXiv preprint arXiv:2108.00804 (2021)
21. Hui, B., Geng, R., Wang, L., Qin, B., Li, B., Sun, J., Li, Y.: S^2 sql: Injecting syntax to question-schema interaction graph encoder for text-to-sql parsers. arXiv preprint arXiv:2203.06958 (2022)
22. Jagannatha, A.N., Yu, H.: Structured prediction models for rnn based sequence labeling in clinical text. In: Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing. vol. 2016, p. 856 (2016)
23. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)
24. Kacupaj, E., Plepi, J., Singh, K., Thakkar, H., Lehmann, J., Maleshkova, M.: Conversational question answering over knowledge graphs with transformer and graph attention networks. arXiv preprint arXiv:2104.01569 (2021)
25. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: DailyDialog: A manually labelled multi-turn dialogue dataset. In: Kondrak, G., Watanabe, T. (eds.) Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 986–995. Asian Federation of Natural Language Processing, Taipei, Taiwan (Nov 2017), https://aclanthology.org/I17-1099/
26. Liang, C., Berant, J., Le, Q., Forbus, K.D., Lao, N.: Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. arXiv preprint arXiv:1611.00020 (2016)
27. Liu, J., Zhang, Z., Razavian, N.: Deep ehr: Chronic disease prediction using medical notes. In: Machine Learning for Healthcare Conference. pp. 440–464. PMLR (2018)
28. Lopez, K., Fodeh, S.J., Allam, A., Brandt, C.A., Krauthammer, M.: Reducing annotation burden through multimodal learning. Frontiers in big Data **3**, 19 (2020)
29. Lorenzo, A.C.M., Maru, M., Navigli, R.: Fully-semantic parsing and generation: The babelnet meaning representation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1727–1741 (2022)
30. Luo, X., Gandhi, P., Zhang, Z., Shao, W., Han, Z., Chandrasekaran, V., Turzhitsky, V., Bali, V., Roberts, A.R., Metzger, M., et al.: Applying interpretable deep learning models to identify chronic cough patients using ehr data. Computer Methods and Programs in Biomedicine **210**, 106395 (2021)
31. Naseem, T., Ravishankar, S., Mihindukulasooriya, N., Abdelaziz, I., Lee, Y.S., Kapanipathi, P., Roukos, S., Gliozzo, A., Gray, A.: A semantics-aware transformer model of relation linking for knowledge base question answering. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 256–262 (2021)
32. Naseem, U., Musial, K., Eklund, P., Prasad, M.: Biomedical named-entity recognition by hierarchically fusing biobert representations and deep contextual-level word-embedding. In: 2020 International joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2020)
33. Nguyen, P., Tran, T., Wickramasinghe, N., Venkatesh, S.: Deepr: A convolutional net for medical records. arxiv. org (2016)
34. Nie, L., Cao, S., Shi, J., Sun, J., Tian, Q., Hou, L., Li, J., Zhai, J.: Graphq ir: Unifying the semantic parsing of graph query languages with one intermediate representation. arXiv preprint arXiv:2205.12078 (2022)

35. Procopio, L., Tripodi, R., Navigli, R.: Sgl: Speaking the graph languages of semantic parsing via multilingual translation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 325–337 (2021)
36. Rousseau, F., Kiagias, E., Vazirgiannis, M.: Text categorization as a graph classification problem. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1702–1712 (2015)
37. Sakai, H., Lam, S.S.: Large language models for healthcare text classification: A systematic review. arXiv preprint arXiv:2503.01159 (2025)
38. Samuel, D., Barnes, J., Kurtz, R., Oepen, S., Øvrelid, L., Velldal, E.: Direct parsing to sentiment graphs. arXiv preprint arXiv:2203.13209 (2022)
39. Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., Gonzalez, G.: Utilizing social media data for pharmacovigilance: a review. Journal of biomedical informatics **54**, 202–212 (2015)
40. Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., Okruszek, L.: Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research **304**, 114135 (2021)
41. Shervashidze, N., Schweitzer, P., Van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-lehman graph kernels. Journal of Machine Learning Research **12**(9) (2011)
42. Weißenhorn, P., Donatelli, L., Koller, A.: Compositional generalization with a broad-coverage semantic parser. In: Proceedings of the 11th Joint Conference on Lexical and Computational Semantics. pp. 44–54 (2022)
43. Wu, S., Chen, B., Xin, C., Han, X., Sun, L., Zhang, W., Chen, J., Yang, F., Cai, X.: From paraphrasing to semantic parsing: Unsupervised semantic parsing via synchronous semantic decoding. arXiv preprint arXiv:2106.06228 (2021)
44. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.: Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. Advances in neural information processing systems **31** (2018)
45. Yuan, R., Wang, Z., Li, W.: Event graph based sentence fusion. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 4075–4084 (2021)
46. Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., Chen, J.: Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In: Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020. pp. 109–117 (2020)
47. Zeng, Q.T., Redd, D., Divita, G., Jarad, S., Brandt, C., Nebeker, J.R.: Characterizing clinical text and sublanguage: A case study of the va clinical notes. J Health Med Informat S **3**(2) (2011)
48. Zheng, C., Chen, X., Xu, R., Chang, B.: A double-graph based framework for frame semantic parsing. arXiv preprint arXiv:2206.09158 (2022)
49. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI open **1**, 57–81 (2020)