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

Yi-Chun (Rimi) Chen

November 23, 2025

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.

1 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:

1.1 Task Definition

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

1.2 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?

1.3 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?

1.4 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?

2 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.

2.1 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 $K_{\mathrm{eff}} = 4.1$, 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 $K_{\mathrm{eff}} = 16.1$, 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 $K_{\text{eff}} = 24.4$, 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.

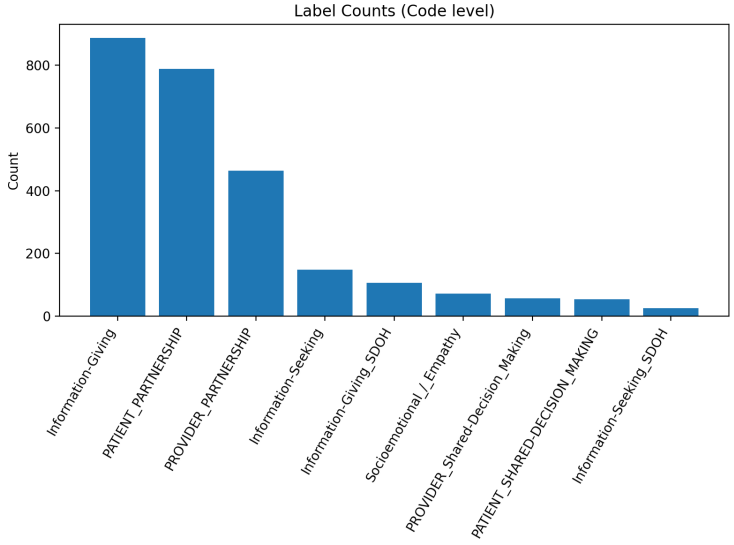

Figure 1: Distribution of annotation counts at the code level.

2.1.1 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 ($\mathbf{IR_{max}}$). The maximum imbalance ratio is defined as:

$$\mathrm{IR}_{\max} = \frac{\max_i n_i}{\min_i n_i}.$$

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:

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

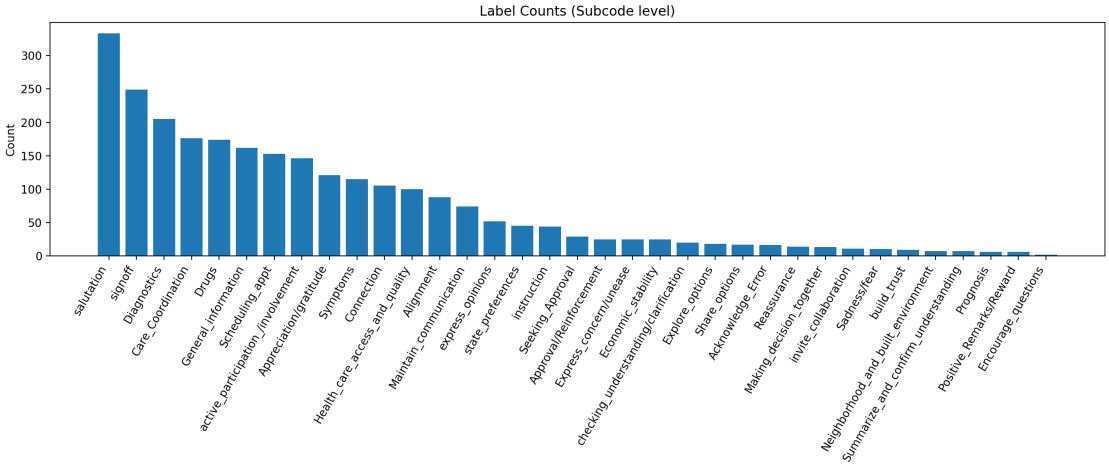

Figure 2: Distribution of annotation counts at the subcode level.

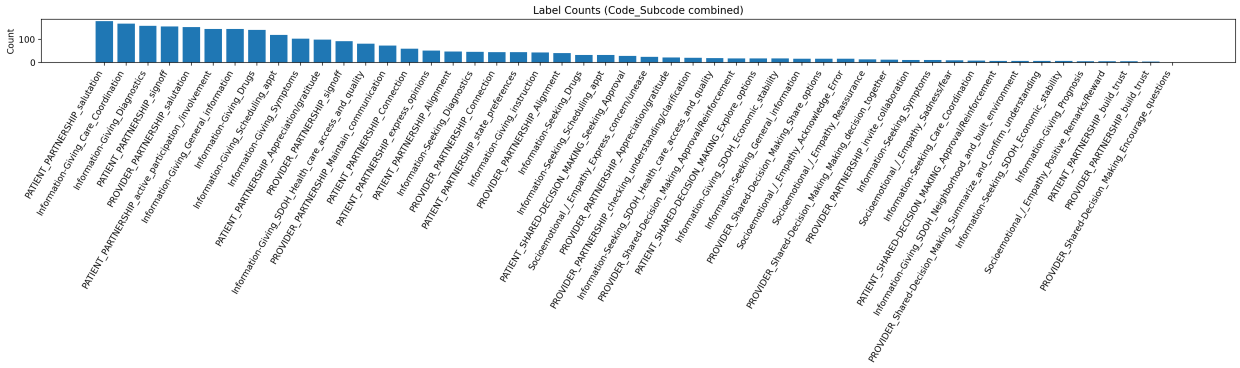

Figure 3: Distribution of annotation counts at the combined code–subcode level.

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 $H_{\text{norm}} = H/\log K$, 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:

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

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: $\text{IR}_{\max}$ highlights extremes, CV captures overall spread, entropy and Gini measure diversity and inequality, and K_{eff} summarizes the effective class space.

2.2 Mixed Semantics

2.2.1 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.

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

Subset	**Classes**	$\mathbf{IR_{max}}$	**CV**	$H_{\mathbf{norm}}$	$K_{\mathbf{eff}}$
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

Findings.

- **Interactional labels** are frequent but shallow. The distribution is dominated by a few formulaic categories such as *salutation* and *signoff*. This yields an extreme imbalance ($\text{IR}_{\max} = 421.5$) and a very low effective number of classes ($K_{\text{eff}} = 6.9$ 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 ($\text{IR}_{\max} = 34.2$), and the effective number of classes ($K_{\text{eff}} = 7.1$ 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.

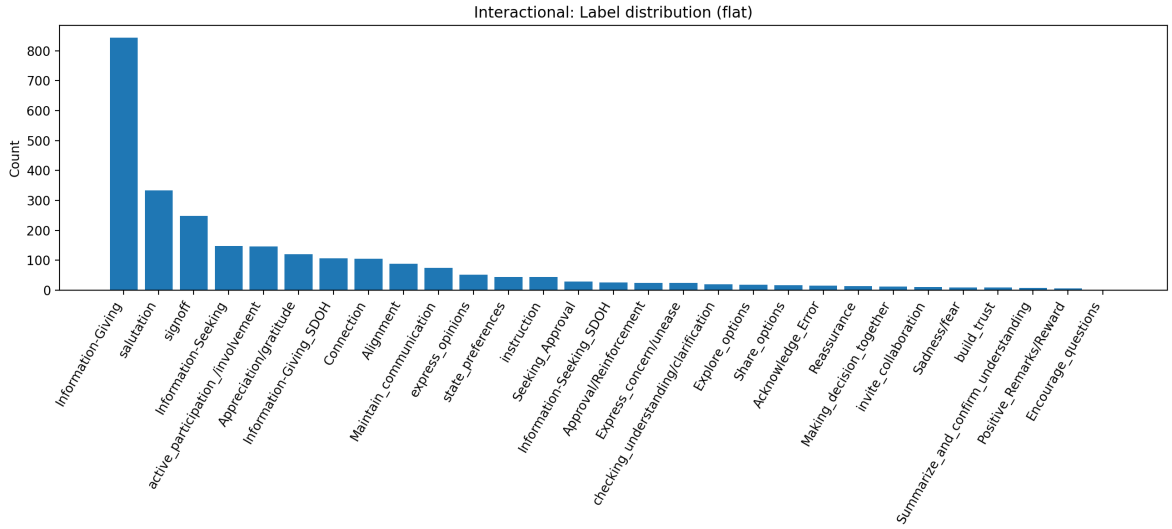

Figure 4: Label distribution for the **Interactional** subset (flat). The distribution is dominated by a few formulaic classes, such as *salutation* and *signoff*, reflecting high skew.

2.3 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.

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

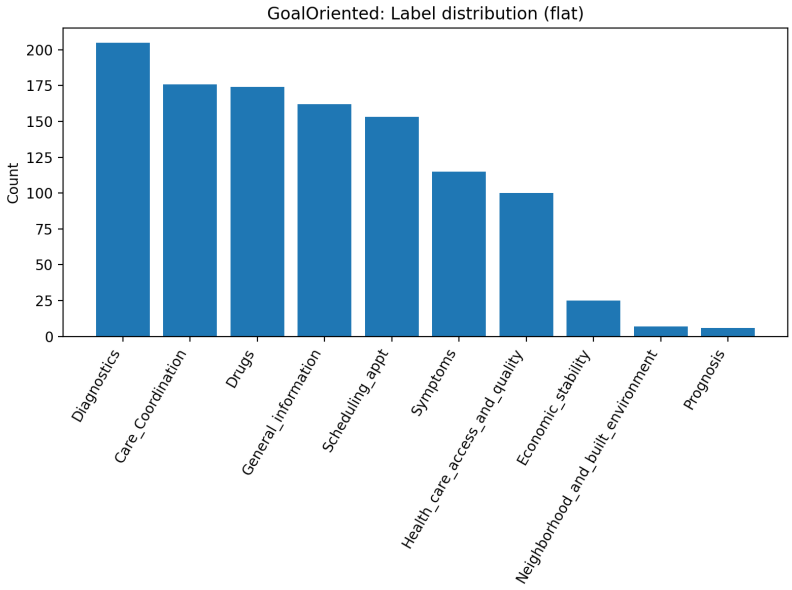

Figure 5: Label distribution for the **Goal-Oriented** subset (flat). While absolute counts are smaller, the distribution is relatively more balanced across classes.

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

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.

2.4 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 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.

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

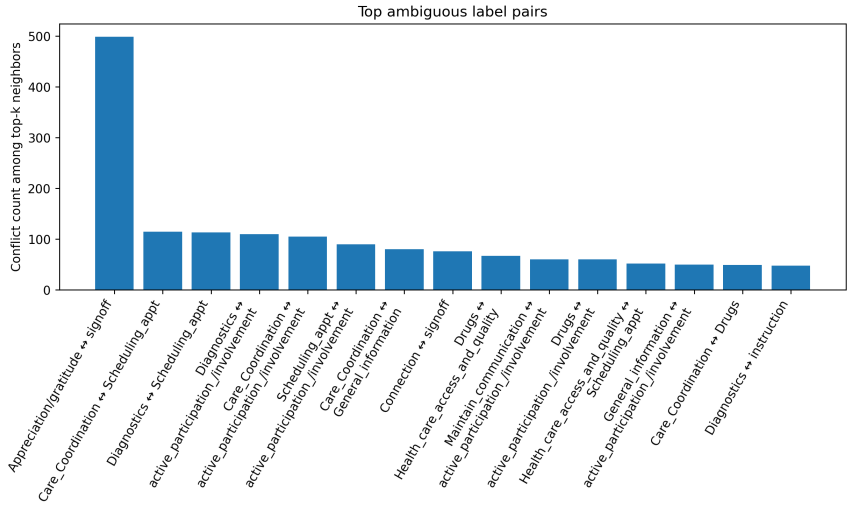

Figure 6: Distribution of the most frequent ambiguous label pairs. Height indicates the number of nearest-neighbor conflicts between categories.

- **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

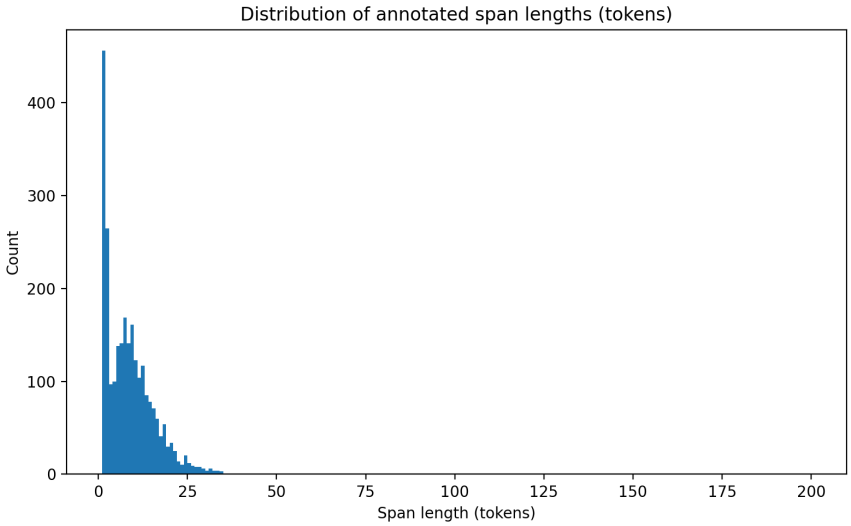

Figure 7: Distribution of annotated span lengths. While most spans are short, a long tail of extended spans reflects annotation inconsistency.

models can memorize with little generalization value, while very long spans conflate multiple intents into a single label, obscuring the decision boundary. Misaligned 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.

3 Method Feasibility

3.1 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 [3]. 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 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 [1]. 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* [2]. 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) 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 [1].

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 [1], adding local context windows, or modeling explicit rhetorical relations with pooling strategies [2]. 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.

3.2 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.

3.2.1 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 energy) and theoretical expressivity tests (e.g., Weisfeiler-Lehman refinement [4]). 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.

- **Nodes / Edges:** Average graph size. Small graphs ($<$10 nodes) limit relational variety; larger graphs ($>$50 nodes) generally provide richer contexts.
- **Density / Average Degree:** Ratio of actual to possible edges, and mean degree per node. Low values ($\leq$ 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 ($\sim$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 $>$0.1 typically indicate richer local structure.
- **Degree Entropy:** Shannon entropy of degree distribution. Low entropy ($\sim$1) means most nodes have similar degree, reflecting structural uniformity. Higher entropy ($>$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.

3.2.2 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 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].*

3.2.3 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.

3.2.4 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.

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

4 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 [1] or discourse-level relation extraction [2]. 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.

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.

5 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 supporting 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] 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)

[2] Hsu, I., Guo, X., Natarajan, P., Peng, N., et al.: Discourse-level relation extraction via graph pooling. arXiv preprint arXiv:2101.00124 (2021)

[3] Sakai, H., Lam, S.S.: Large language models for healthcare text classification: A systematic review. arXiv preprint arXiv:2503.01159 (2025)

[4] 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)

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.

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