Analysis and Report for Bethesda Data

Yi-Chun (Rimi) Chen

November 24, 2025

1 Report Summary

- **Alignment between NLP structure and annotation spans:** Clause-level segmentation (subsentences) achieves 74.98% alignment with gold spans, compared to only 25.02% at the sentence level. *Conclusion:* clause segmentation provides a much closer fit to annotation practice.
- **Label distribution:** Both codes and subcodes show heavy skew, with a few frequent categories dominating and many rare ones forming a long tail. *Conclusion:* the dataset exhibits uneven coverage, especially for socially routine versus goal-oriented categories.
- **Dataset balance:** Imbalance metrics (IR_{max}, Gini, effective class count) confirm severe long-tail distributions. Codes effectively collapse to ∼2-3 classes, subcodes to ∼7, and combined labels to ∼8 of 35. *Conclusion:* models trained on this dataset will be biased toward majority classes unless corrected.
- **Ambiguity analysis (subcode level):** Several subcodes (e.g., *Gratitude*, *Maintain Communication*) show disagreement rates above 40%. *Conclusion:* subcodes suffer from overlapping definitions and inconsistent application.
- **Ambiguity analysis (code level):** Ambiguity is somewhat lower than subcodes but still exceeds 10–15% for categories like *PartnershipProvider* and *SocioEmotionalBehaviour*. *Conclusion:* broader codes are not immune to overlap and require clearer boundaries.
- **Ambiguity analysis (combined code-subcode level):** Highest ambiguity overall, with some categories exceeding 70% disagreement. *Conclusion:* the fine-grained schema is too noisy for stable modeling without consolidation or hierarchical treatment.
- **Cross-level comparison:** Ambiguity increases with granularity (code < subcode < combined), with partnership and socio-emotional categories consistently problematic. *Conclusion:* downstream modeling should prioritize schema refinement, label grouping, or uncertainty-aware methods.

1.1 Alignment Between NLP Structure and Annotation Spans

Goal. We evaluate how well our segmentation procedure (at both the sentence and clause level) aligns with human-provided annotation spans in patient-provider messages. Each message includes annotated text spans, and we compare these against automatically derived sentences and clause-like subsentences.

Segmentation method. Messages are first split into sentences using spaCy. Clause-like subsentences are then derived using a rule-based procedure guided by syntactic dependency relations, including *advcl*, *ccomp*, *xcomp*, *relcl*, and *parataxis*, along with discourse connectives and punctuation. Very short fragments are merged and very long ones are divided with token-length thresholds. Each sentence and subsentence is compared with gold spans using fuzzy matching at a threshold of 0.6.

Results. Table 1 reports corpus-level statistics. Subsentence segmentation yields substantially higher coverage of gold spans (74.98%) compared to sentence-level segmentation (25.02%).

Table 1: Alignment between segmentation units and gold spans.

Total messages	149
Total sentences (N_S)	908
Total subsentences (N_C)	2,297
Total gold annotations	1,547
Matched sentences (M_S)	387
Sentence alignment rate	25.02%
Matched subsentences (M_C)	1,160
Subsentence alignment rate	74.98%

1.2 Label Distribution

We computed message-level counts for each code and subcode. Figures 1, 2, and 3 illustrate the distributions.

At the **code level**, a few categories dominate (Figure 1). At the **subcode level**, the skew is more severe, with socially routine acts (e.g., greetings, acknowledgments) heavily overrepresented (Figure 2). The **joint distribution** (Figure 3) shows uneven coverage of code-subcode pairs, with many rare or absent.

2 Dataset Balance Analysis

We quantified skew at the *code*, *subcode*, and *combined* levels using imbalance ratio ($\mathrm{IR}_{\max}$), coefficient of variation (CV), normalized entropy (H_{norm}), Gini index (G), and effective number of classes (K_{eff}). Results are shown in Table 2.

Table 2: Imbalance metrics for the Bethesda dataset.

Level	Classes	$\mathrm{IR}_{\max}$	CV	H_{norm}	G	K_{eff}
Code	8	128.0	1.47	0.62	0.61	2.54
Subcode	26	402.0	1.65	0.73	0.86	6.96
Combined	35	402.0	1.78	0.74	0.88	8.40

Findings. A balanced dataset would typically have $\mathrm{IR}_{\max}<10$, $G < 0.5$, and K_{eff} close to the nominal label count. Bethesda falls far outside these ranges: codes collapse to only 2-3 effective classes, subcodes to $\sim$7 of 26, and combined labels to $\sim$8 of 35. Models trained on this data will therefore overfit to majority classes, neglecting rare but clinically important intents.

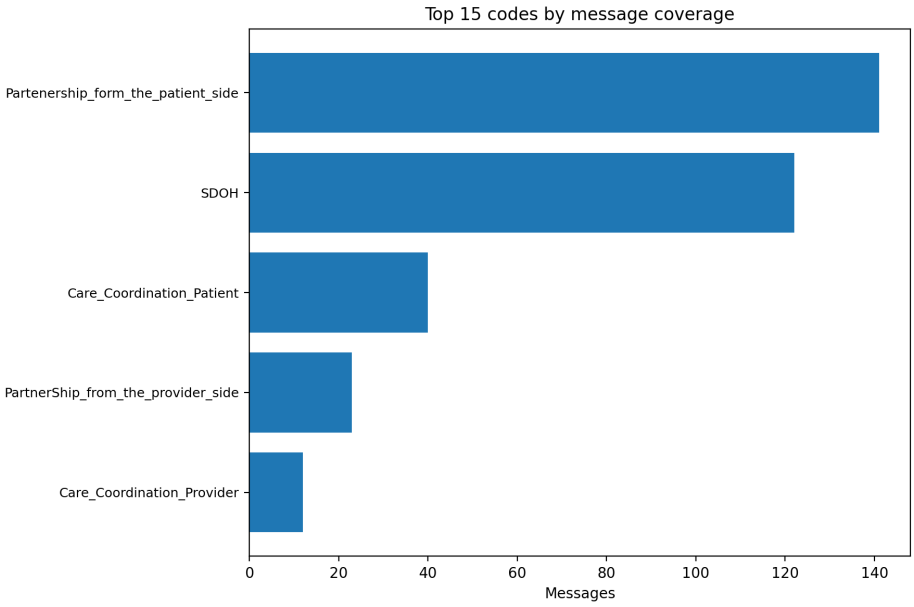

Figure 1: Top codes by message coverage.

3 Ambiguity Analysis

To examine annotation ambiguity, we evaluate how often semantically similar text spans receive different labels. We operationalize this by representing each annotated span with four alternative similarity measures: a dataset-independent embedding model (Sentence-BERT), dataset-dependent embeddings (TF-IDF and LSA), and a lexical overlap baseline (Jaccard similarity). For each method, we compute nearest-neighbor pairs of spans in the representation space and flag a conflict when the pair is annotated with different labels. We then aggregate (i) the number of conflict pairs, (ii) the number of distinct label pairs involved, and (iii) per-label disagreement rates. This procedure is applied at three levels of granularity: subcode, code, and combined code-subcode.

3.1 Subcode Level

Sentence-BERT found 1,365 conflicts (10.9%), TF-IDF 703 (7.3%), LSA 2,461 (23.3%), and Jaccard 317 (2.6%) (Table 3). Ambiguity hotspots (Figure 4) included *Approval of Decision/Reinforcement*, *Appreciation/Gratitude*, and *Maintain Communication*, exceeding 40%. Frequent conflicts (Figure 5) included *Appreciation/Gratitude* vs. *Signoff* and *Clinical Care* vs. *Active Participation/Involvement*. Supervised datasets typically aim for below 10-15% disagreement; several subcodes exceed this.

Table 3: Subcode-level ambiguity statistics across representation methods.

Method	Conflict Pairs	Distinct Label Pairs	Avg. Disagreement	Median Disagreement
Sentence-BERT	1,365	55	0.109	0.064
TF-IDF	703	62	0.073	0.043
LSA	2,461	110	0.233	0.204
Jaccard	317	16	0.026	0.001

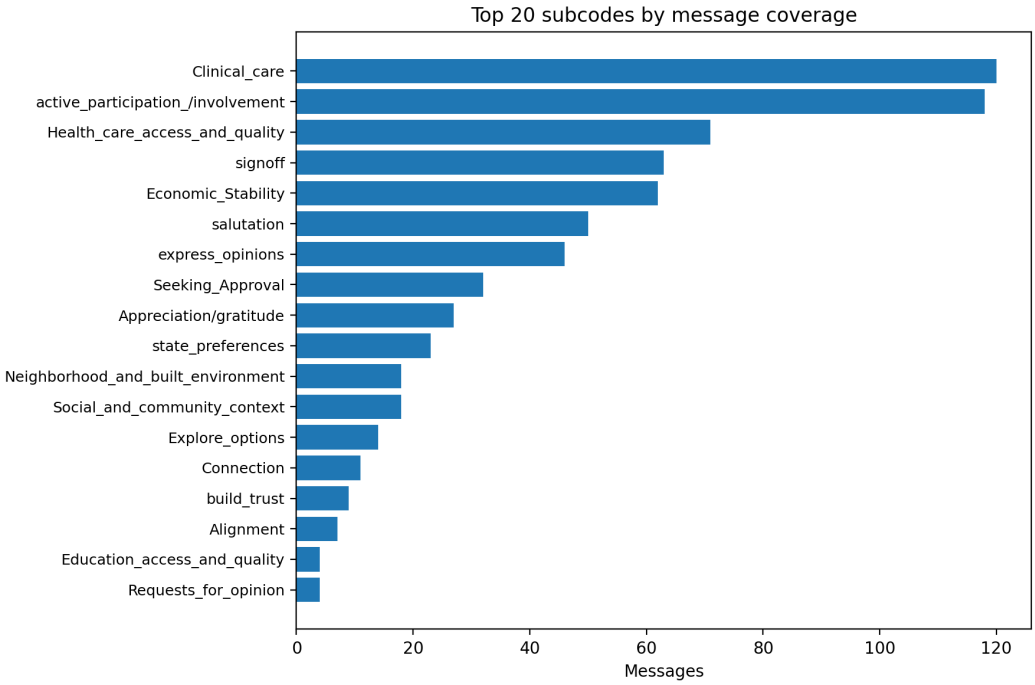

Figure 2: Top subcodes by message coverage.

3.2 Code Level

At the code level, results (Table 4) show 1,203 conflicts with SBERT (14.9%), 678 with TF-IDF (5.7%), 2,610 with LSA (23.7%), and 269 with Jaccard (2.0%). Problematic codes (Figure 6) include *PartnershipProvider* (36%), *SocioEmotionalBehaviour* (30%), and *SharedDecisionProvider* (27%). Frequent conflicts (Figure 7) include *PartnershipPatient* vs. *SDOH* and *PartnershipPatient* vs. *PartnershipProvider*. While somewhat lower than subcodes, many still exceed the 10-15% threshold.

Table 4: Code-level ambiguity statistics across representation methods.

Method	Conflict Pairs	Distinct Label Pairs	Avg. Disagreement	Median Disagreement
Sentence-BERT	1,203	22	0.149	0.105
TF-IDF	678	17	0.057	0.042
LSA	2,610	25	0.237	0.221
Jaccard	269	5	0.020	0.004

3.3 Combined Code-Subcode Level

At the combined level, ambiguity was highest: 1,863 conflicts with SBERT (18.7%), 1,018 with TF-IDF (13.1%), 3,566 with LSA (32.7%), and 499 with Jaccard (7.5%) (Table 5). High-disagreement categories (Figure 8) included *PartnershipProvider* → *Appreciation/Gratitude* and *PartnershipProvider* → *Salutation*, both exceeding 70%. Frequent conflicts (Figure 9) included *Appreciation/Gratitude* vs. *Signoff* and *Clinical Care* vs. *Active Participation/Involvement*. These results far exceed the 10-15% guideline, suggesting the schema is too fine-grained without adjudication or

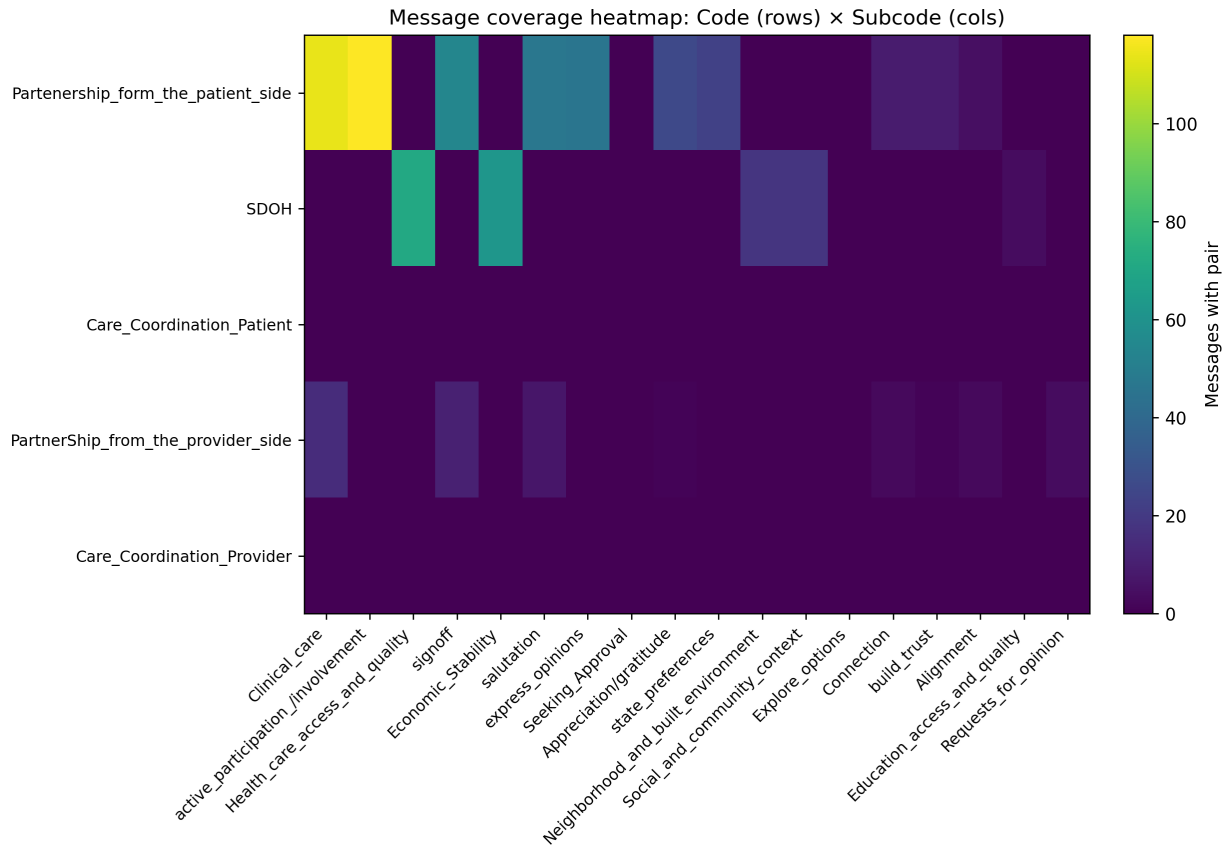

Figure 3: Heatmap of message coverage across code (rows) and subcode (columns).

label consolidation.

Table 5: Combined code-subcode ambiguity statistics across representation methods.

Method	Conflict Pairs	Distinct Label Pairs	Avg. Disagreement	Median Disagreement
Sentence-BERT	1,863	93	0.187	0.110
TF-IDF	1,018	82	0.131	0.050
LSA	3,566	181	0.327	0.300
Jaccard	499	27	0.075	0.005

3.4 Cross-Level Comparison

Ambiguity increases with granularity: lowest at the code level, higher at the subcode level, and highest at the combined code-subcode level. Partnership, socio-emotional, and communication-related categories consistently appear as ambiguity hotspots. These patterns suggest downstream models will overfit to majority classes while underperforming on overlapping or fine-grained ones, motivating schema refinement, hierarchical grouping, or uncertainty-aware modeling.

References

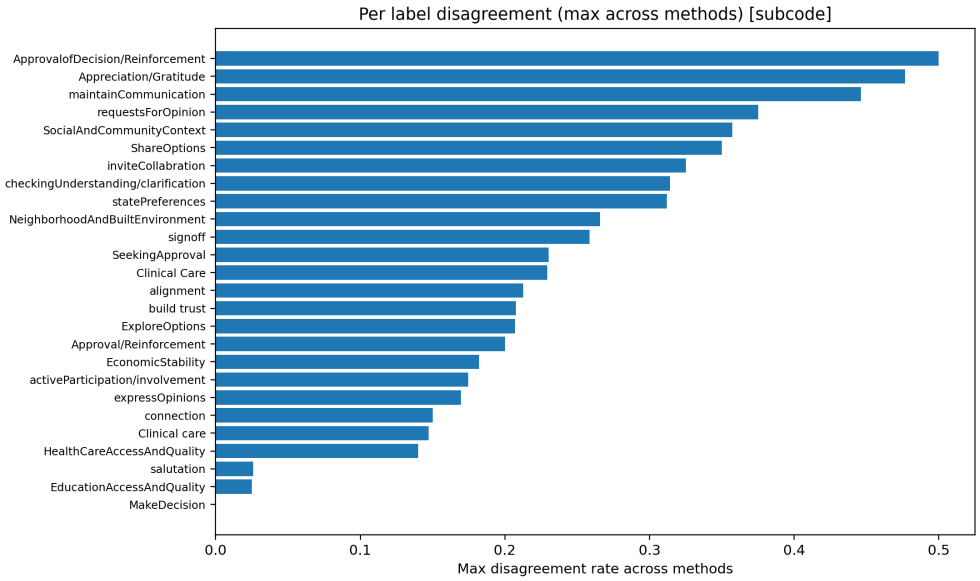

Figure 4: Maximum per-label disagreement across methods (subcode level).

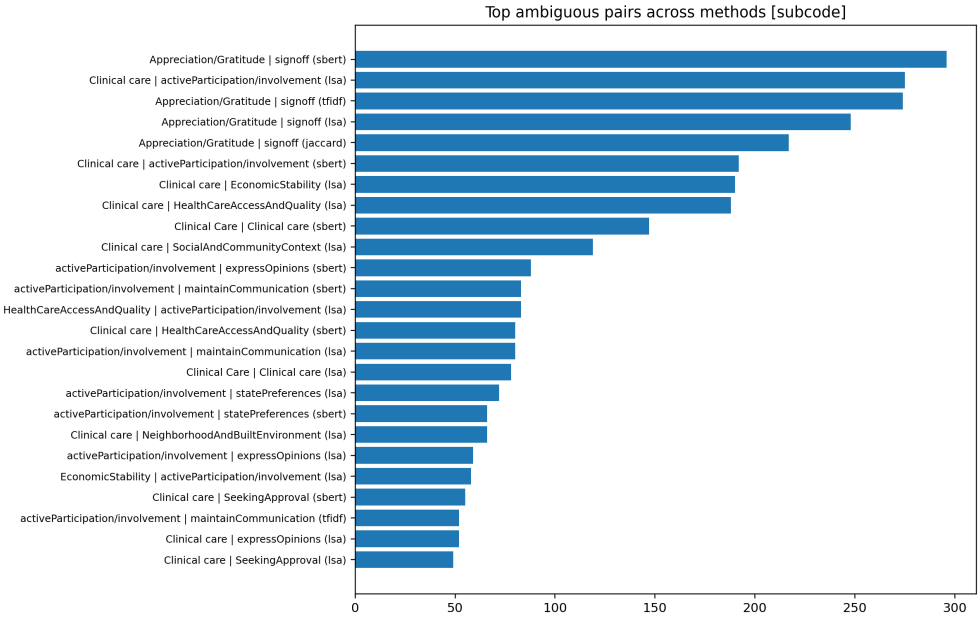

Figure 5: Top ambiguous subcode pairs across methods.

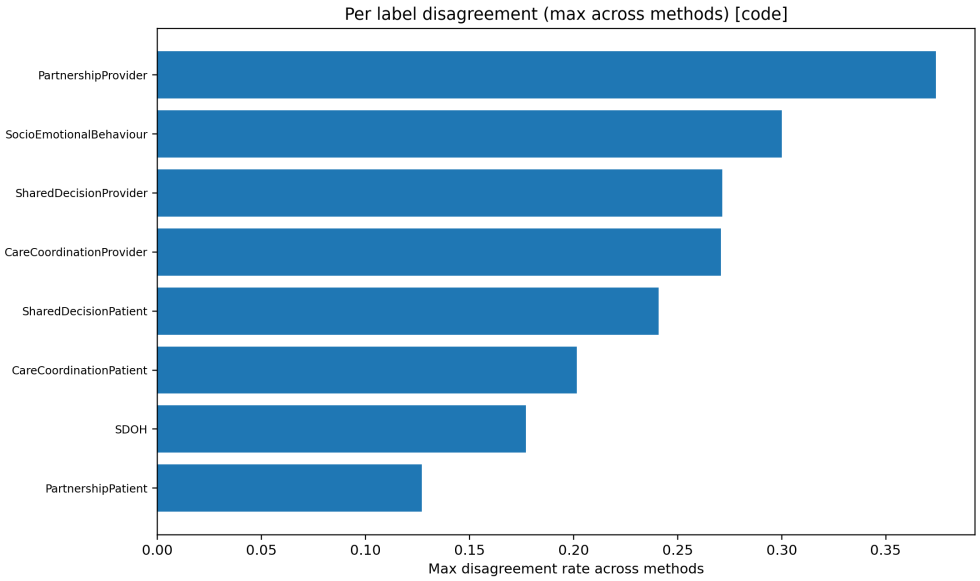

Figure 6: Maximum per-label disagreement across methods (code level).

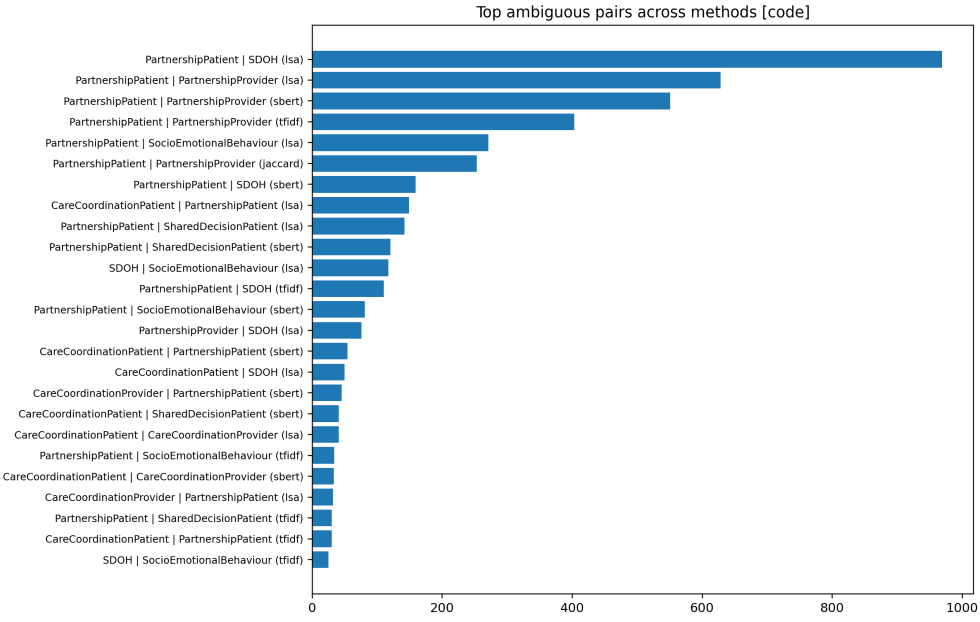

Figure 7: Top ambiguous code pairs across methods.

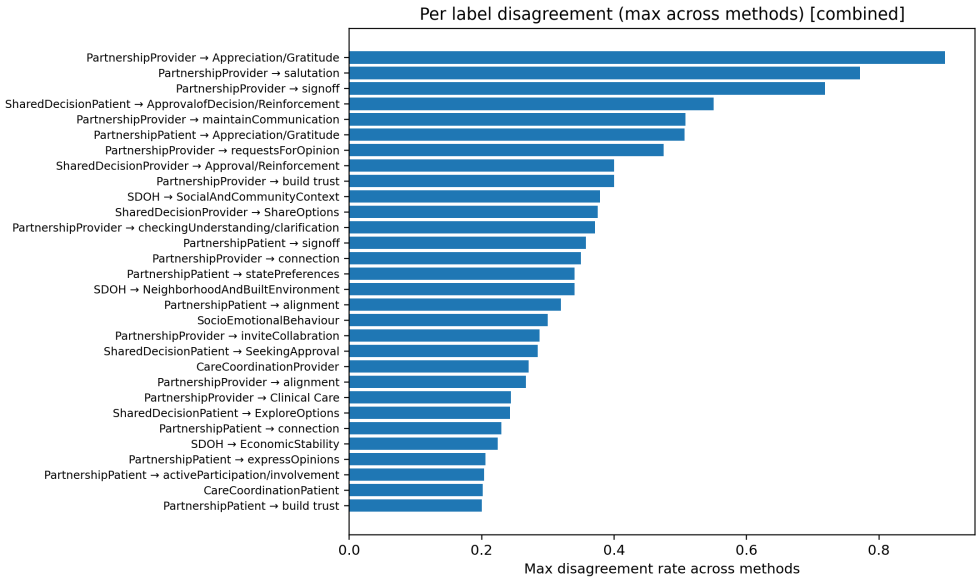

Figure 8: Maximum per-label disagreement across methods (combined level).

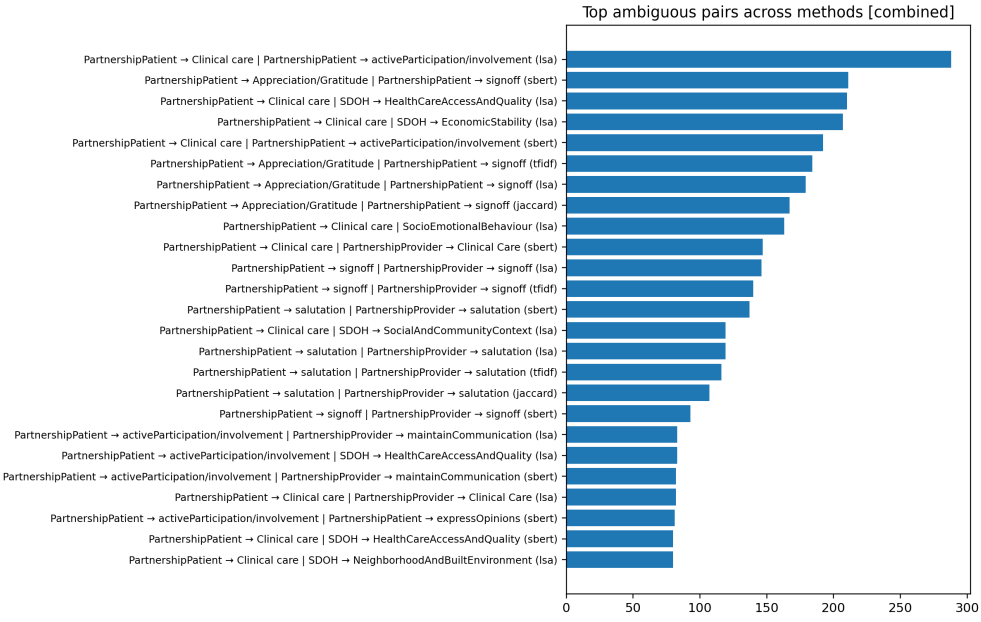

Figure 9: Top ambiguous combined pairs across methods.