Detecting Inter-Annotator Inconsistency in Non-Overlapping Dialogue Act Annotations via Semantic Alignment and Sparse Latent Features

Anonymous ACL submission

Abstract

Dialogue-act annotation is often distributed across disjoint corpus partitions, where standard inter-annotator agreement measures are inapplicable. We propose a diagnostic framework that aligns semantically similar turns, extracts residual-stream SAE features, and compares semantically filtered feature inventories across annotators. On AMI and DialogBank, annotator-level feature drift is lower than tagset-, corpus-, language-, and layer-induced drift, while marker-level validation reveals systematic differences in annotators' reliance on surface vs. pragmatic evidence. The framework enables interpretable inspection of annotation behaviour without overlapping annotations.

1 Introduction

DA annotation is often influenced by subtle pragmatic and contextual cues, which makes annotation consistency a major challenge (Geertzen et al., 2008; Qian et al., 2021; Duran et al., 2022; Lee et al., 2026). Annotation consistency has traditionally been assessed with inter-annotator agreement measures such as Cohen's κ and Krippendorff's α (Artsein and Poesio, 2008). Recent work treats inconsistency as a meaningful signal, arising from linguistic ambiguity, underspecified schemes, overlapping label semantics, or systematic differences in annotator interpretation (Uma et al., 2022; Jiang and de Marneffe, 2022; Ignatev et al., 2025; Mostafazadeh Davani et al., 2022; Rodríguez-Barroso et al., 2024; Xu and Jurgens, 2026; Ivey et al., 2025). However, this work still mainly assumes overlapping annotation partitions, where multiple annotators label the same instances. Yet in practice, annotators are often assigned disjoint partitions for scalability and cost reduction, making agreement metrics and inconsistency-aware models inapplicable and leaving systematic annotator differences undetected. We address this gap with a diagnostic framework that flags semantically aligned turns as candidate inconsistencies when they receive different DA labels and show divergent inferred decision-relevant features. To this end, we align semantically equivalent turns across disjoint annotator partitions, extract interpretable sparse features with residual-stream SAEs, and lexicalize these features into human-readable cues. Experiments on AMI and English/Dutch DialogBank show that the framework surfaces inconsistency patterns not captured by embedding similarity alone. **Our contributions** are a diagnostic framework for non-overlapping annotations, a feature-conditioned drift measure based on interpretable SAE latents, and a lexicalization and marker-level validation pipeline for analyzing annotator-specific labeling behavior.

2 Related Work

Our work relates to three research lines. First, the use of SAEs to decompose language-model activations into more interpretable latent features, often viewed as approximately monosemantic dimensions of polysemantic representations (Cunningham et al., 2023; Bricken et al., 2023). Lieberum et al. (2024) released *Gemma Scope*, an open suite of SAEs trained on Gemma 2 activations (Gemma Team, 2024). Closest to our work, Kuznetsov et al. (2025) combine Gemma Scope features with XGBoost for artificial text detection and identify a small subset of predictive, interpretable latents. We adapt this layer-sweep and tree-based feature-importance setup to DA classification. Second, our task builds on DA annotation work, in particular ISO 24617-2 (Bunt et al., 2012), the ISO-annotated AMI corpus (Carletta et al., 2006), and DialogBank (Bunt et al., 2019). DA labeling often yields only moderate inter-annotator agreement (Geertzen et al., 2008; Duran et al., 2022), motivating recent "disagreement-as-signal" approaches (Mostafazadeh Davani et al., 2022; Rodríguez-

Barroso et al., 2024; Ivey et al., 2025; Xu and Jurgens, 2026). Unlike these approaches, which typically assume overlapping annotations of the same items, we infer annotator-specific criteria from non-overlapping annotation partitions. Third, interpretable feature attribution. Tree ensembles provide internal importance scores such as gain (Chen and Guestrin, 2016); SHAP and TreeSHAP provide additive explanations for model predictions (Lundberg and Lee, 2017; Lundberg et al., 2020); and permutation importance/model reliance estimate relevance via feature perturbation (Breiman, 2001; Fisher et al., 2019). We combine these model-based scores with class-conditional selectivity measures, Cohen's d (Cohen, 1988) and single-feature AUC, to link DA-relevant SAE latents to lexical and pragmatic markers of annotator-specific labeling behavior.

3 Methodology

Let the dialogue corpus $\mathcal{D}$ consist of a set of dialogue turns $\{t_1, t_2, \ldots, t_n\}$ and $\mathcal{D}$ be partitioned into m disjoint subsets $\mathcal{P} = P_1, P_2, \ldots, P_m$, where each partition $P_k \subset \mathcal{D}$ contains the dialogue turns annotated by annotator a_k, such that $P_i \cap P_i = \emptyset, \forall i \neq j$ and $\bigcup_{k=1}^{m} P_k = \mathcal{D}$. Each dialogue turn t_i receives a DA label y_i^l, where $y^l \in \mathcal{Y}$ (with $\mathcal{Y}$ being the predefined DA label inventory).

Since annotators label disjoint corpus partitions, no shared instances are available for standard inter-annotator agreement metrics. We therefore aim to detect potentially inconsistent annotations without requiring annotation overlap.

3.1 Aligned sparse features

To reconstruct a proxy shared annotation space, we identify semantically similar dialogue turns across non-overlapping annotation partitions. Given a turn t_i, we compute a dense semantic embedding $\mathbf{e}(t_i) = f_{\text{emb}}(t_i)$ with a pretrained sentence encoder. For each $t_i \in P_a$, we retrieve nearest turns $t_j \in P_b$, $a \neq b$, using cosine similarity:

$$\text{sim}(t_i, t_j) = \frac{\mathbf{e}(t_i) \cdot \mathbf{e}(t_j)}{\|\mathbf{e}(t_i)\| \|\mathbf{e}(t_j)\|}. \quad (1)$$

We retain cross-partition pairs above a threshold τ:

$$\mathcal{S} = \{(t_i, t_j) \mid t_i \in P_a,\ t_j \in P_b,\ a \neq b,\ \text{sim}(t_i, t_j) \geq \tau\}. \quad (2)$$

$\mathcal{S}$ is computed once with a permissive threshold $\tau = 0.80$, and stricter cutoffs are applied downstream: 0.90 for marker-level validation and 0.95 for activation-ranked visualisation.

Embedding similarity captures global semantic proximity, but not necessarily the cues salient for annotation decisions. We therefore represent each turn with residual-stream SAE features. Following Kuznetsov et al. (2025), we treat the extraction layer as a hyperparameter: for each layer ℓ, per-token activations $\mathbf{z}^{(\ell)}(t_i)$ are pooled into an utterance-level vector. For each aligned pair $(t_i, t_j) \in \mathcal{S}$, we compare the most active latents, $F(t_i) = \text{TopK}(\mathbf{z}^{(\ell)}(t_i))$ and $F(t_j) = \text{TopK}(\mathbf{z}^{(\ell)}(t_j))$. Because only a small fraction of SAE latents is expected to be DA-informative (Kuznetsov et al., 2025), we filter the inventory using model-based importance scores (gain, mean-|SHAP| (Lundberg and Lee, 2017), permutation importance) and class-conditional selectivity measures (Cohen's d (Cohen, 1988), single-feature AUC). The retained features are then lexicalized and compared for inconsistency detection.

3.2 Lexicalization and semantic filtering

Selected SAE features are statistically informative, but their semantics are not directly observable from feature identifiers. We therefore lexicalize each retained feature with an auto-interpretation procedure (Bricken et al., 2023). We then apply a label-contrastive semantic filter (*Sem-Contr*). For each feature–label pair (f, y), a second prompt judges whether the description is related to y, given the feature's top-5 alternative labels in the same partition. It returns IS_RELATED $\in \{\text{T}, \text{F}\}$ and CONF $\in [0, 100]$. Sem-Contr removes features judged unrelated with high confidence, i.e., IS_RELATED $=$ F and CONF $\geq \theta$. We set $\theta = 80$; a sweep over $\theta \in \{50, \ldots, 85\}$ shows stable F1 for $\theta \in [50, 80]$ and a drop at 85 (Appendix C).

3.3 Feature drift and validation

Given the aligned cross-partition turns $\mathcal{S}$ and, for each annotator k and dialogue-act label y, the semantically filtered inventory of lexicalized SAE features $F_k(y)$, we measure whether annotators rely on similar features for the same DA label. We define cross-annotator feature drift as mean pairwise Jaccard disagreement:

$$\text{Drift}(y) = 1 - \frac{1}{\binom{m}{2}} \sum_{k<l} J\big(F_k(y), F_l(y)\big), \quad (3)$$

where m is the number of annotators and J the Jaccard index. A drift of 0 indicates full feature

consensus; values close to 1 indicate largely disjoint feature inventories. We use $\mathcal{S}$ to distinguish variation from potential inconsistency: pairs with different DA labels but overlapping filtered features are treated as *legitimate variation*, whereas pairs with different labels and non-overlapping feature sets are treated as *candidate-inconsistent*.

4 Experiments

4.1 Data

We evaluate our methodology on three DA-annotated datasets: AMI with its original tagset (Carletta et al., 2006), an ISO 24617-2 re-annotation of AMI (Bunt et al., 2012), and the English and Dutch subsets of DialogBank (Bunt et al., 2019). ISO 24617-2 provides a general-purpose, domain-independent DA inventory (Bunt et al., 2012; Mezza et al., 2018), while DialogBank serves as its multilingual reference resource (Bunt et al., 2019; Ribeiro et al., 2022). Including both AMI variants allows us to separate effects of annotator behaviour from effects of annotation scheme.

The two AMI variants are annotated by three annotators (DS, SN, VK) on disjoint corpus partitions, matching the non-overlapping setting targeted by our method. For AMI-ISO, we use the deterministic AMI ISO mapping specified by Bunt et al. (2012), with context-sensitive disambiguation conventions for *inform*; the full mapping is given in Appendix A.1. After harmonization, six ISO dimensions/communicative functions have sufficient support (see Figure 2 in Appendix A.2) and are retained; AMI-original retains its full tagset. DialogBank is used to assess cross-language and cross-domain generalization.

Since drift measures feature-set divergence rather than pragmatic inconsistency directly, we further validate candidate disagreements with a two-stage marker-level analysis (Annex C).

We compute $\mathcal{S}$ once with a permissive $\tau = 0.80$ and apply tighter cutoffs downstream (0.90 for marker-level validation, 0.95 for the activation-ranked visualization); Then we aligned $|\mathcal{S}|$ per annotator pair using $\tau = 0.80$ $\tau = 0.90$ and $\tau = 0.95$ cosine-similarity thresholds (Table 5). Cross-lingual DialogBank admits only a few near-identical pairs ($|\mathcal{S}| = 4$ at $\tau \geq 0.95$), motivating $\tau = 0.90$ as the operating point for the marker-level analysis.

4.2 SAE configuration and feature filtering

We use Gemma-2-2B and its Gemma Scope residual-stream SAEs (16,384 latents per layer; Lieberum et al., 2024). For each layer $\ell \in \{0, \ldots, 25\}$, we extract SAE encoder activations directly, without decoder reconstruction, and aggregate them with max or mean pooling. For each layer–pooling pair, we train an XGBoost classifier (Chen and Guestrin, 2016) and select the configuration with the highest macro-F1 averaged over three seeds. Dense residual-stream hidden states from the same layer are evaluated analogously as an embedding-only baseline, allowing us to test whether the interpretable SAE space preserves predictive performance (Appendix B, Figure 4).

To filter features, we combine model-based importance (XGBoost gain, mean-$|\mathrm{SHAP}|$ (Lundberg and Lee, 2017), permutation importance) with class-conditional selectivity (Cohen's d (Cohen, 1988), single-feature AUC). A latent is retained if $|d| \geq 1.0$ and $\mathrm{AUC} \notin (0.2, 0.8)$, or if it lies in the intersection of the importance and selectivity rankings at percentile rank ≥ 0.5; the top 30 latents per label are kept. We use layer 18 with max pooling for English, AMI, and cross-dataset comparisons. Dutch peaks at layer 12, but remains within the layer-18 performance plateau; using L18 therefore keeps feature indices comparable across datasets.

Implementation details, LLM prompt summaries, the Sem-Contr threshold sweep, and the layer sweep are provided in Appendix C. Additional aligned-set sizes, filtering trajectories, per-label drift values, and qualitative cases are reported in Appendix D.

5 Results

Feature filtering reduces the SAE inventory by roughly two orders of magnitude with limited performance loss. On English, the trajectory is 16,384 $\rightarrow$ 950 $\rightarrow$ 134 features (All$\rightarrow$Selected$\rightarrow$Sem-Contr), with macro-F1 $0.278 \rightarrow 0.316 \rightarrow 0.296$ (Appendix D, Figure 3). Thus, statistical selection improves predictive performance, while Sem-Contr removes roughly half of the selected features at a cost of 0.02 macro-F1. High-confidence Sem-Contr rejections are dominated by non-pragmatic surface detectors, such as file paths, BOS tokens, and console-log strings.

The resulting SAE classifiers remain usable across all four regimes used in the drift analysis: AMI-ISO, AMI-original, DialogBank EN, and Di-

Comparison	Pairs	$\overline{\text{drift}}$	Consensus
AMI-ISO 3 ann. (Selected)	3	0.357	51.7%
AMI-ISO 3 ann. (Sem-Contr)	3	0.503	38.0%
AMI-orig 3 ann. (Sem-Contr)	3	0.584	27.6%
AMI-ISO DS vs. AMI-orig DS	1	0.695	30.5%
AMI-ISO DS vs. DB EN	1	0.804	19.6%
DB EN vs. DB NL L18	1	0.783	21.7%
AMI-ISO DS vs. DB NL L18	1	0.939	6.1%

Table 1: Feature drift across comparison regimes. Drift is computed on the Sem-Contr inventory, except for the Selected row. DB = DialogBank; EN = English; NL = Dutch.

alogBank NL. Per-dataset macro-F1 is 0.30–0.39 on DialogBank and 0.40–0.54 on the AMI partitions (Appendix D, Table 4). Moreover, in the layer sweep, SAE-max activations match or exceed dense residual-stream hidden states at the same layer almost everywhere on English and remain competitive on Dutch (Appendix D, Figure 4). This suggests that the interpretability gain of the SAE space does not come with a measurable F1 penalty relative to a dense embedding-only baseline.

Feature drift forms a graded signal across annotator, tagset, corpus, language, and layer differences. On Sem-Contr, AMI-ISO three-annotator drift is 0.503, while AMI-original reaches 0.584 (Table 1). Drift increases for cross-tagset comparison of the same annotator (0.695), cross-corpus comparison with DialogBank English (0.804), cross-language comparison between DialogBank English and Dutch (0.783), and stacked cross-corpus/cross-language comparison (0.939).

Overall, annotator divergence is the smallest non-trivial regime, below tagset, corpus, language, and layer effects. The drift score itself depends on the per-label Sem-Contr feature inventory rather than on the alignment threshold τ, which controls only the size of $\mathcal{S}$ (Table 5) and not the inventories $F_k(y)$ that enter Eq. 3; the ordering of comparison regimes in Table 1 is therefore stable across the three downstream thresholds ($\tau \in \{0.80, 0.90, 0.95\}$). Per-label AMI-ISO drift ranges from 0.212 for *setQuestion* to 0.738 for *answer* (Appendix D, Table 6).

Marker-level validation shows that Sem-Contr does not remove all surface artefacts. On AMI-ISO disagreements at $\tau \geq 0.90$, the marker-conditioned pipeline maps 799 fired Sem-Contr features to taxonomy nodes; 62–88% per annotator are classified as residual surface confounds (Appendix D, Table 7), with annotators differing not only in which features they recruit but also in the balance between pragmatic and surface-based evidence. Figure 1 illustrates this on the identical marker *"Maybe two?"*: DS's *inform* is supported by surface-format detectors, whereas SN's *suggest* is supported by modal-possibility and hedging features.

Marker. *"Maybe two ?"* cos sim = 1.00, activation rank #1
DS *"Maybe two ?"* **inform**
SN *"Maybe two ?"* **suggest**

Fired Sem-Contr features (annotator-unique):
DS / *inform*:
f_{660} "list-introducing phrases" [SURFACE]
f_{4449} "colon introducing code/lists/explanations" [SURFACE]
SN / *suggest*:
f_{3981} "modal possibility (*may/might be*)" [PRAGMATIC]
f_{2089} "sentence-initial hedging (*Maybe*)" [PRAGMATIC]

Diagnosis. DS's *inform* inventory recruits format detectors that misfire on the trailing punctuation; SN's *suggest* inventory latches onto the modal-possibility cue.

Figure 1: Top-activation AMI-ISO DS–SN disagreement traced through marker-level validation: identical marker, annotator-unique Sem-Contr features, and the surface-vs.-pragmatic split.

6 Conclusions and Future Work

We presented a framework for detecting inter-annotator inconsistency in DA-annotated corpora with non-overlapping annotation partitions. It aligns semantically similar turns, extracts label-contrastive SAE features, and summarizes annotator divergence with a cross-annotator drift score. On AMI-ISO, annotator drift is lower than cross-dataset drift against DialogBank (0.503 vs. 0.708–0.978), showing that the measure captures graded variation across annotators, corpora, languages, and representation settings. Marker-level validation further shows that annotators differ not only in which features they recruit, but also in their reliance on pragmatic vs. surface-based evidence. Thus, the framework supports targeted inspection of annotator-specific labeling behavior without requiring overlapping annotations. AI was used as support for coding and writing, but not for research.

Future work will apply the framework to additional non-overlapping DA corpora and other LMs to assess the robustness of layer selection and feature inventories. We also plan a human-adjudication study comparing high- and low-drift aligned pairs to test whether the proposed diagnostics match perceived annotation inconsistency.

Limitations

The presented work has still a number of limitations, , which we outline here together with how we plan to address them. First, the framework surfaces *candidate* inconsistencies for human inspection; it does not directly confirm them. Drift values in Table 1 should be read as *relative* comparisons across regimes, not as absolute inconsistency rates. Also, we do not validate flagged pairs against human judgment; a small expert-adjudication study is planned (cf. Section 6) to calibrate these relative scores against human-confirmed inconsistencies.

Second, both the Sem-Contr filter and the marker-level validation rely on `gpt-5-mini` judgment. We report neither inter-run agreement nor prompt-perturbation stability; the $\theta \in [50, 80]$ macro-F1 plateau (Appendix C) provides indirect evidence of robustness but does not replace a direct reliability study. Sem-Contr and marker-pipeline labels should therefore be treated as model-derived diagnostic categories, not as gold pragmatic annotations.

Third, classifier performance varies by partition (Table 4): macro-F1 is 0.30–0.39 on DialogBank and 0.40–0.54 on AMI. DialogBank is small and label-imbalanced, limiting interpretability at the low-frequency tail, expanding and re-balancing this data is a natural next step.

Finally, cross-language coverage is restricted to one typologically close pair (English/Dutch) and one backbone (Gemma-2-2B with Gemma Scope residual-stream SAEs at layer 18). The cross-layer comparison is not possible since SAE features are not transferable across layers; generalization to other SAE families, distant languages, and non-meeting genres is left to future work.

References

R. Artsein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang, and Kars Wijnhoven. 2019. The DialogBank: dialogues with interoperable annotations. *Language Resources and Evaluation*, 53(2):213–249.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maelle Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction (MLMI 2005)*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39, Berlin, Heidelberg. Springer.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, San Francisco, California, USA. ACM.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

N. Duran, S. Battle, and J. Smith. 2022. Inter-annotator agreement using the conversation analysis modelling schema, for dialogue. *Communication Methods and Measures*, 16(3):182–214.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Jeroen Geertzen, Volha Petukhova, and Harry Bunt. 2008. Evaluating dialogue act tagging with naive and expert annotators. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

D. Ignatev, D. Paperno, and M. Poesio. 2025. Annotator disagreement in RST annotation schemes. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 18–32, Eugene, Oregon.

Jonathan Ivey, Susan Gauch, and David Jurgens. 2025. NUTMEG: Separating signal from noise in annotator disagreement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2874–2887, Suzhou, China. Association for Computational Linguistics.

N.-J. Jiang and M.-C. de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*.

Kristian Kuznetsov, Laida Kushnareva, Anton Razzhigaev, Polina Druzhinina, Anastasia Voznyuk, Irina Piontkovskaya, Evgeny Burnaev, and Serguei Barannikov. 2025. Feature-level insights into artificial text detection with sparse autoencoders. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25727–25748, Vienna, Austria. Association for Computational Linguistics.

Jinsook Lee, Kirk Vanacore, Zhuqian Zhou, Bakhtawar Ahtisham, Jeanine Grutter, and Rene F. Kizilcec. 2026. Codebook-injected dialogue segmentation for multi-utterance constructs annotation: Llm-assisted and gold-label-free evaluation. *Preprint*, arXiv:2601.12061.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2:56–67.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 4765–4774. Curran Associates, Inc.

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. ISO-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Annotation inconsistency and entity bias in MultiWOZ. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 326–337, Singapore and Online. Association for Computational Linguistics.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2022. Automatic recognition of the general-purpose communicative functions defined by the ISO 24617-2 standard for dialog act annotation. *Journal of Artificial Intelligence Research*, 73.

N. Rodríguez-Barroso, E. Martínez Cámara, J. Camacho Collados, M.V. Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

A Uma, D Almanea, and M. Poesio. 2022. Scaling and disagreements: Bias, noise, and ambiguity. *Front Artif Intell.*

Yinuo Xu and David Jurgens. 2026. Beyond consensus: Perspectivist modeling and evaluation of annotator disagreement in nlp. *Preprint*, arXiv:2601.09065.

A AMI dataset re-annotation

A.1 AMI original → ISO 24617-2

The AMI Meeting Corpus is distributed with an in-house tagset (*inf*, *sug*, *bck*, . . .). For the AMI-ISO variant we re-annotate every segment to the ISO 24617-2 inventory (Bunt et al., 2012) using a deterministic rule-based map followed by context-sensitive disambiguation for *inform*. Both versions are then analysed independently (the AMI *original* variant retains the AMI tagset; the AMI-ISO variant uses the mapping below). Table 2 lists the static mapping; the disambiguation rules for *inform* are:

- *inf* after *el.inf* → `task_answer`
- *inf* after *inf* with a positive relation tag → `task_agreement`
- *inf* after *inf* with a negative relation tag → `task_disagreement`
- otherwise → `task_inform`

AMI code	ISO dimension	ISO function
inf	task	inform
sug	task	suggest
off	task	offer
ass	task	inform
el.inf	task	setQuestion
el.und	task	checkQuestion
el.sug	task	NO_LABEL
el.ass	alloFeedback	NO_LABEL
bck	alloFeedback	alloPositive
und	autoFeedback	autoPositive
stl	timeManagement	stalling
be.pos	socialObligations-Management	NO_LABEL
be.neg	socialObligations-Management	NO_LABEL
fra	ownCommunication-Management	NO_LABEL
oth	ownCommunication-Management	NO_LABEL

Table 2: Static AMI short-code → ISO 24617-2 mapping used to build AMI-ISO. *inf* is further disambiguated by the previous turn's AMI code and the AMI relation tag.

Short codes that do not correspond to a canonical ISO function in our inventory are mapped to `NO_LABEL` on their dimension and are excluded from the per-label analyses; the same convention is used for unannotated AMI segments.

A.2 AMI DA label redistribution

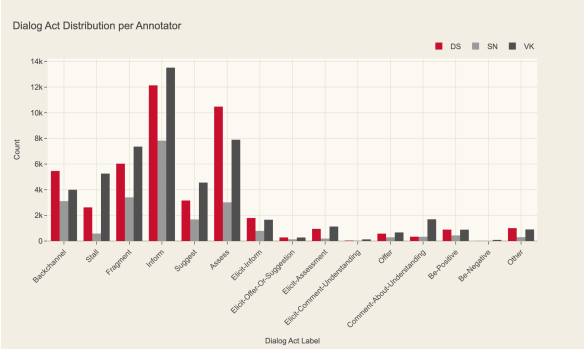

Figure 2: DA label distribution on AMI after re-annotation to the ISO 24617-2 inventory. Highly skewed, with *inform* dominating.

B Illustration of the Gemma-2-2B layer sweep

Figure 4 illustrates the Gemma-2-2B's layers 0–25 sweep for our datasets. Performance plateaus in layers 10–20, SAE features peak at L18 (English) and L12 (Dutch), and SAE-max dominates hidden-state-max almost everywhere on English while remaining competitive on Dutch. We use L18 for all cross-dataset analyses so that feature indices share a common SAE basis.

C Implementation Details

LLM prompts. All LLM stages use `gpt-5-mini` (Singh et al., 2026) with low reasoning effort and structured JSON output. Full prompt text is included in the supplementary material and will be released with the code. We use LLMs in four stages. First, for *feature lexicalization*, the model receives the top-10 max-activating snippets for an SAE feature from Neuronpedia's public broad-domain activation database and returns a one-sentence `gpt-5-mini` generated description plus a categorical tag, following the auto-interpretation format of Bricken et al. (2023). Second, for *label-contrastive classification*, the model receives the cached lexicalization, the ISO definition of a target DA label y, and the feature's importance-ranked top-5 alternative labels in the same partition, and returns $\{\text{IS_RELATED} \in \{\text{T}, \text{F}\}, \text{CONF} \in [0, 100], \text{RELATION_TYPE}, \text{SEMANTIC_ASPECT}\}$. Sem-Contr removes features with $\text{IS_RELATED} = \text{F}$ and $\text{CONF} \geq \theta$.

For marker-level validation, Stage M_1 induces a marker-specific taxonomy. Given a marker, its observed label-pair frequencies, and up to three ± 4-turn dialogue excerpts, the model produces a hierarchy of at most seven top-level pragmatic categories, with definitions and example markers per node. Stage M_2 maps fired features to this taxonomy. Given the marker, the Stage M_1 taxonomy, and per-annotator lists of fired Sem-Contr features with cached lexicalizations, contrastive-classifier metadata, and max activation values, the model assigns each feature to a taxonomy node and flags $\text{IS_SURFACE_CONFOUND} \in \{\text{T}, \text{F}\}$.

Sem-Contr threshold. We sweep the Sem-Contr confidence threshold $\theta \in \{50, 60, 70, 80, 85\}$ on English DialogBank. Macro-F1 is stable for $\theta =$ 50–80 (0.302–0.305) and drops at $\theta = 85$ (0.293). We therefore use $\theta = 80$, the strictest cutoff that yields the smallest interpretable feature inventory without material performance loss.

Layer sweep. For each layer $\ell \in \{0, \ldots, 25\}$ of Gemma-2-2B, we compare SAE activations with dense residual-stream hidden states under max and mean pooling. Dense hidden states serve as a within-pipeline embedding-only baseline. Figure 4

reports the full English/Dutch DialogBank sweep. Performance is highest or near-highest in the middle layers; English peaks at layer 18 with max pooling, while Dutch peaks at layer 12 but remains within the layer-18 plateau. We therefore use layer 18 for English, AMI, and cross-dataset comparisons.

D Additional Results

Aligned-set sizes. Table 5 reports the number of aligned cross-partition pairs $|\mathcal{S}|$ at three cosine-similarity thresholds. We compute $\mathcal{S}$ once with a permissive $\tau = 0.80$ and apply tighter cutoffs downstream (0.90 for marker-level validation, 0.95 for the activation-ranked visualization). Aligned-set sizes per annotator pair are given in Appendix D, Table 5; cross-lingual DialogBank admits only a few near-identical pairs ($|\mathcal{S}| = 4$ at $\tau \geq 0.95$), motivating $\tau = 0.90$ as the operating point for the marker-level analysis.

Table 5 displays the size of the aligned set $|\mathcal{S}|$ per annotator pair at three cosine-similarity thresholds.

Pair	$\tau \geq 0.80$	≥ 0.90	≥ 0.95
AMI-ISO DS–SN	14,008	2,452	617
AMI-ISO DS–VK	26,026	3,715	965
AMI-ISO SN–VK	12,076	1,699	500
AMI-orig DS–SN	300,355	59,924	13,725
AMI-orig DS–VK	570,312	109,471	24,252
AMI-orig SN–VK	405,229	82,177	17,886
DialogBank EN–NL	68	12	4

Table 3: Size of the aligned set $|\mathcal{S}|$ per annotator/language pair at three cosine-similarity thresholds.

Figure 3 shows the performance trajectory for Gemma-2-2B layers sweep for the English and Dutch DialogBanks.

Per-dataset classifier performance. Table 4 reports the per-partition feature inventory size and classification performance at the Sem-Contr stage (seed 42, $\theta = 80$) on every dataset used in the drift analysis. The SAE classifier remains usable across all four regimes: macro-F1 lies in 0.30–0.39 for the cross-language DialogBank partitions and in 0.40–0.54 for the AMI partitions, with accuracy 0.55–0.87 correlating with label-distribution skew. AMI partitions train on larger inventories and yield higher macro-F1 and accuracy than DialogBank; AMI-ISO is slightly above AMI-original at comparable inventory sizes. Drift estimates therefore rest on classifiers that are non-degenerate, though absolute F1 should be read with the small DialogBank label-support caveats discussed in the Limitations section.

Dataset	#feat	Macro-F1	Acc.
DialogBank EN	112	0.296	0.662
DialogBank NL L12	62	0.391	0.548
DialogBank NL L18	66	0.359	0.575
AMI-ISO DS / SN / VK	55–76	0.487–0.541	0.833–0.865
AMI-orig DS / SN / VK	85–101	0.396–0.480	0.620–0.686

Table 4: Per-dataset Sem-Contr feature-inventory size and classification performance at $\theta = 80$ (seed 42), the production setting used for the drift analysis. AMI rows give the min–max across the three annotators.

Per-label drift. Table 6 reports per-label AMI-ISO drift after Sem-Contr filtering. Drift is highest for *answer* and *inform*, and lowest for *setQuestion*. This supports the main-text finding that Sem-Contr exposes label-specific annotator signatures that are partly masked by statistical selection alone.

Marker-level validation details. Table 7 reports the surface-confound rate per annotator from Stage 2 of the marker pipeline. VK shows the highest rate in every pair in which it appears.

Qualitative case studies. In addition to the main-text marker-validation example, Table 8 reports activation-ranked qualitative cases, one per AMI annotator pair and one for the cross-lingual DialogBank comparison. In each row, *Side A* corresponds to the first annotator (or language) named in the pair heading (e.g. *DS* in *AMI-ISO DS-vs-SN*) and *Side B* to the second; listed features are the Sem-Contr features unique to that side's label inventory that fire above the activation floor on the aligned turn. The table illustrates how feature drift can be inspected at the level of individual aligned turns: e.g. in AMI-ISO DS–SN, DS's *offer* label on *"Well, I shall go to the next slide."* is supported by a first-person-intention feature, whereas SN's *inform* label on the near-paraphrase *"I'm going to go to that next slide,"* is supported by a topic-focus feature. These pairs are returned as *diagnostic candidates*, not as confirmed inconsistencies (cf. the Limitations section).

Pair	$\tau \geq 0.80$	≥ 0.90	≥ 0.95
AMI-ISO DS–SN	14,008	2,452	617
AMI-ISO DS–VK	26,026	3,715	965
AMI-ISO SN–VK	12,076	1,699	500
AMI-orig DS–SN	300,355	59,924	13,725
AMI-orig DS–VK	570,312	109,471	24,252
AMI-orig SN–VK	405,229	82,177	17,886
DialogBank en–nl	68	12	4

Table 5: Size of the aligned set $|\mathcal{S}|$ per annotator pair at three cosine-similarity thresholds.

Label	N	Consensus	Drift
answer	16	3 (19%)	0.738
inform	12	3 (25%)	0.622
checkQuestion	31	9 (29%)	0.564
offer	19	7 (37%)	0.500
suggest	26	13 (50%)	0.380
setQuestion	19	13 (68%)	0.212
mean			**0.503**

Table 6: Per-label cross-annotator drift on AMI-ISO after Sem-Contr filtering ($\theta = 80$). N: union of Sem-Contr features over the three annotators for the label. *Consensus*: features selected by all three annotators.

Pair	Ann.	N	Confound	Rate
DS–SN	DS	95	72	0.76
DS–SN	SN	240	192	0.80
DS–VK	DS	85	53	0.62
DS–VK	VK	188	151	0.80
SN–VK	SN	88	57	0.65
SN–VK	VK	229	201	0.88

Table 7: Surface-confound rates from marker-level validation. N: fired Sem-Contr features mapped to taxonomy nodes across the 200 most frequent markers at $\tau \geq 0.90$.

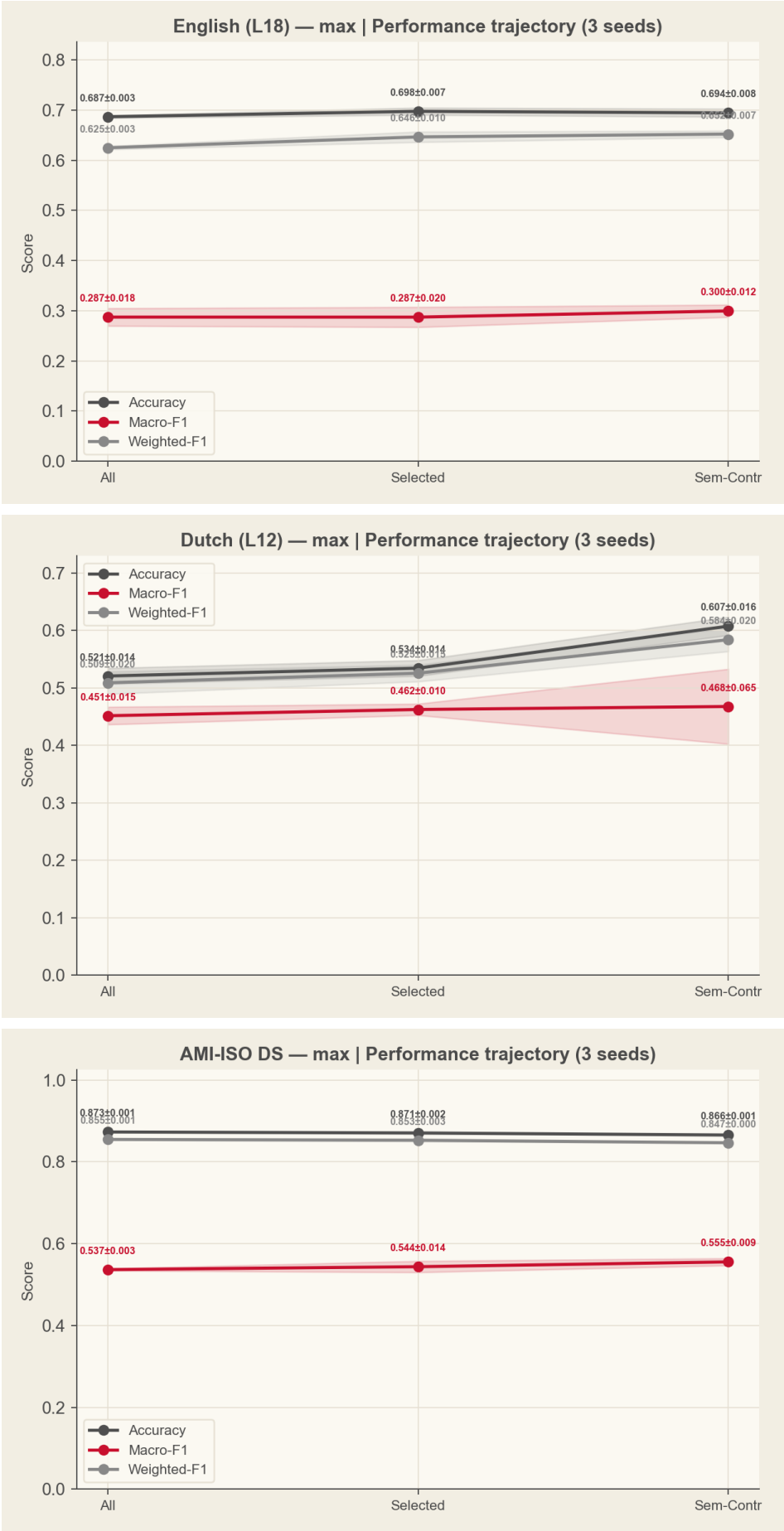

Figure 3: Performance trajectory *All*→*Selected*→*Sem-Contr* for English (top), Dutch (middle), and AMI-ISO DS (bottom); accuracy / macro-F1 / weighted-F1, ±1 s.d. over three seeds. SN, VK and AMI-original trajectories.

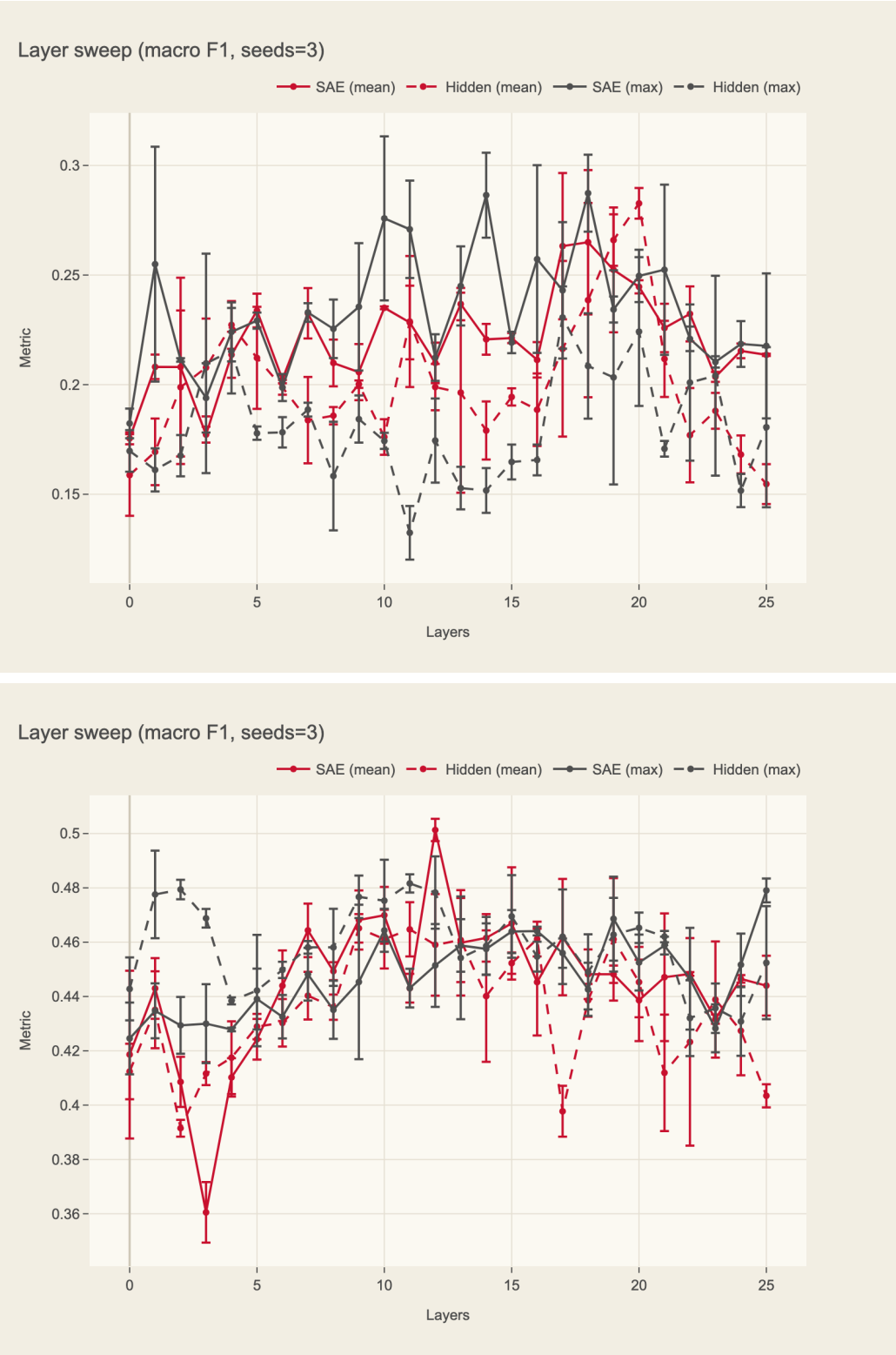

Figure 4: Layer sweep across Gemma-2-2B layers 0–25 for ISO DialogBank English (top) and Dutch (bottom), macro-F1 averaged over three seeds (bands: ± 1 s.d.). SAE (max/mean) traces are shown in red/grey, hidden-state baselines as dashed lines.

Marker pair (cos, rank)	Side A: features (unique to A's label inventory)	Side B: features (unique to B's label inventory)
AMI-ISO DS-vs-SN		
"Maybe two ?" DS/*inform* | SN/*suggest* (cos=1.00, #1)	f_{660}: list-introducing phrases f_{4449}: colon introducing code/lists/explanations	f_{3981}: modal possibility (*may/might be*) f_{2089}: sentence-initial *Maybe* hedge
"Are they ?" DS/*inform* | SN/*setQuestion* (cos=1.00, #3)	f_{660}: list-introducing phrases f_{4449}: colon punctuation	f_{13853}: sentence-initial interrogative (*Is/How*)
"Well, I shall go to the next slide." vs. *"I'm going to go to that next slide,"* DS/*offer* | SN/*inform* (cos=0.95, #6)	f_{10333}: first-person intention (*so I thought I'd...*) f_{5191}: first-person plural subject	f_{15275}: topic-focus / key content words f_{15612}: mentions of specific topics/featured items
AMI-ISO DS-vs-VK		
"Yes they are ." vs. *"Uh, yes they are ."* DS/*inform* | VK/*answer* (cos=0.98, #1)	f_{4449}: colon punctuation f_{5807}: copular verbs (*be/is*, state)	f_{12542}: first-person statements / personal experience f_{6072}: sentence-initial *Yes* (often followed by comma)
"I think um something comfortable to hold," vs. *"I think something that's comfortable to hold"* DS/*inform* | VK/*suggest* (cos=0.96, #7)	f_{4449}: colon punctuation f_{5807}: copular verbs	f_{5878}: modal/epistemic belief (*we must, believes, ought*)
AMI-ISO SN-vs-VK		
"Maybe that's uh better ." vs. *"That's maybe that's better ."* SN/*suggest* | VK/*inform* (cos=0.97, #4)	f_{3981}: modal possibility f_{5080}: modal/ability (*you can* / contracted *can't*)	f_{7343}: mentions of customisable equipment/features f_{10731}: data/samples collected or discarded
"Are there any more questions ," vs. *"Uh are there any more questions ?"* SN/*setQuestion* | VK/*checkQuestion* (cos=0.98, #5)	f_{13853}: sentence-initial interrogative f_{14676}: question phrasing with auxiliary *does*	f_{5169}: sentence-start / emphasised tokens f_{9706}: opening quotation marks / dialog starts
AMI-orig DS-vs-SN		
"Oh Alright ." vs. *"Ah . Okay ."* DS/*oth* | SN/*und* (cos=0.97, #1)	f_{8506}: interjections of surprise (*Wow, Geez*) f_{972}: interjection	f_{4930}: casual affirmations *OK/ok* f_{8224}: speaker stating perception (*I see, I hear*)
"So, I think we're ready ." vs. *"And then we're ready ."* DS/*ass* | SN/*inf* (cos=0.95, #9)	f_{14365}: short opinion phrases (*to me, in my opinion*) f_{16326}: expressive punctuation (commas, exclamations) f_{408}: mild positive review phrases	f_{7173}: assessments, judgments, evaluations f_{3175}: illness/sickness affecting test results
AMI-orig DS-vs-VK		
"Yeah, okay, mm-hmm ." vs. *"Mm yeah, okay ,"* DS/*bck* | VK/*und* (cos=0.97, #1)	f_{10514}: utterance-initial filled pauses (*hmm, hmpf*)	f_{4930}: informal acknowledgments (*ok*, *OK*)
"Oh shit ." vs. *"Shit ."* DS/*oth* | VK/*be.neg* (cos=0.96, #5)	f_{972}: interjection f_{7708}: short informal phrase starters	f_{12350}: negative evaluations / problems (defects, failures) f_{551}: threatening/violent imperatives
AMI-orig SN-vs-VK		
"Or something like that . Yeah ." vs. *"Yeah, yeah, something like that ."* SN/*sug* | VK/*inf* (cos=0.95, #7)	f_{6445}: implementation/action verbs (*carry out, achieve*) f_{10962}: examine / investigate phrases f_{9294}: sentence-opening discourse cues (*Take, Consider*)	f_{15275}: topic-focus / subject emphasis f_{9555}: participation/being involved in something
DialogBank en-vs-nl		
"no no no" vs. *"nee"* en/*answer* | nl/*disagreement* (cos=0.96, #1)	f_{7481}: first-person experience statements f_{6837}: sentence-starting quoted replies f_{6850}: start of answer/response block (*A:* label)	f_{8023}: contraction *don't* (informal negation) f_{11226}: denial of having forced or coerced

Table 8: Twelve activation-ranked AMI and DialogBank disagreements, one to three per pair. Features listed are the Sem-Contr features unique to each annotator's labeled inventory that fire above the activation floor on the disagreement text. Lexicalizations are the Neuronpedia-grounded autointerp descriptions cached in Stage 2 of the filter pipeline.