NoviceTrio in #SMM4H-HeaRD 2026: Hybrid Clinical Transformer Ensembles for Insomnia Detection and Evidence Extraction from Clinical Notes

Abir Naskar and Mike Conway
School of Computing and Information Systems, University of Melbourne
Parkville, Melbourne, 3053, VIC, Australia
anaskar@student.unimelb.edu.au, mike.conway@unimelb.edu.au

Abstract

We present two systems for the #SMM4H-HeaRD 2026 Task 2 (Lopez-Garcia et al., 2026) shared task of automated insomnia detection from clinical notes. Our system addresses both subtasks: (1) binary insomnia classification and (2) multi-label rule prediction with evidence span extraction. For Subtask 1, we employ an ensemble architecture combining Qwen3-4B-Instruct[1] and Bio_ClinicalBERT[2] to capture both general semantic reasoning and domain-specific clinical representations. The framework utilizes chunk-based processing with overlapping token windows to handle long clinical notes efficiently. For Subtask 2, we develop a dual-head multi-task transformer model that jointly predicts insomnia labels and token-level evidence spans using BIO tagging[3]. To improve clinical relevance, we additionally incorporate sentence-level filtering using sentence-transformer embeddings and similarity-based retrieval of insomnia-related contexts. Experimental results suggest competitive performance relative to the shared task mean and median scores across both subtasks. Our best Subtask 1 system achieves a recall of 0.9474, surpassing the shared-task mean and median recall, while our Subtask 2 system exceeds the mean and median scores across label classification, exact match, and partial match metrics. The end-to-end implementation is publicly available on GitHub[4].

1 Dataset Description

The dataset comprises 156 training and 23 validation clinical discharge notes from MIMIC-III, with a test set, all provided by the shared task organisers. For Subtask 1, annotations specify a binary insomnia label (*yes*/*no*) per note. For Subtask 2, each note is annotated for four insomnia-related criteria, Definition 1, Definition 2, Rule B, and Rule C, where each rule carries a binary label and, when positive, a list of character-level evidence spans (formatted as `start` and `end` offsets). The annotations exhibit realistic clinical characteristics including correlated rule activations, partial rule triggering, evidence distributed across lengthy documents, and class imbalance favouring the negative class.

2 System 1: Heterogeneous Ensemble for Binary Classification

The architecture diagram of System 1 is provided in Figure 1.

2.1 Chunking and Input Representation

Each clinical note is tokenised and split into overlapping chunks of up to 512 tokens. The stride differs per model: 256 tokens for Qwen3-4B (Yang et al., 2025) and 128 tokens for Bio_ClinicalBERT (Devlin et al., 2019). This sliding-window approach ensures that no evidence region spanning a chunk boundary is lost through hard truncation. Every chunk inherits the document-level insomnia label.

2.2 Models

Two models are trained independently on the training split.

Qwen3-4B: We use Qwen3-4B-Instruct-2507, a 4B-parameter instruction-tuned decoder. Because the model has no native classification head, we attach a masked mean-pooling classifier. For a chunk with hidden states $\mathbf{H} \in \mathbb{R}^{L \times d}$ and attention mask $\mathbf{m} \in \{0,1\}^L$, the pooled representation is

$$\mathbf{p} = \frac{\sum_{t=1}^{L} m_t \, \mathbf{H}_t}{\sum_{t=1}^{L} m_t}, \tag{1}$$

[1] https://huggingface.co/unsloth/Qwen3-4B-Instruct-2507
[2] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT
[3] BIO: Beginning, Inside, Outside
[4] https://github.com/AbeerNaskar/SMM4H2026_SharedTask2_Insomnia

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 338–344

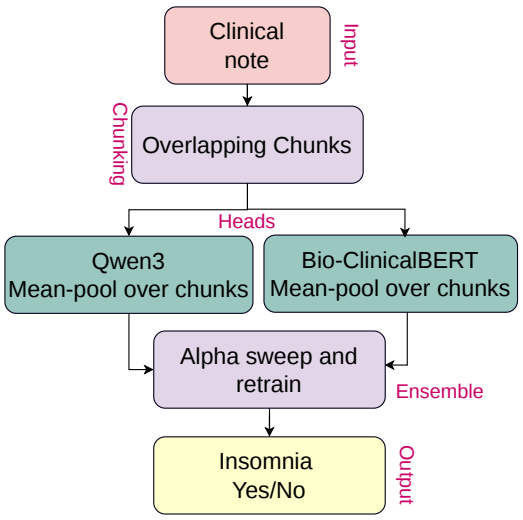

Figure 1: Overview of System 1, a heterogeneous ensemble combining Qwen3-4B and Bio_ClinicalBERT for document-level insomnia classification.

followed by a dropout layer ($p_{\text{drop}} = 0.1$) and a linear projection to two output logits. The learning rate is set to 1×10^{-5}.

Bio_ClinicalBERT: We use Bio_ClinicalBERT, a BERT-base model pre-trained on MIMIC-III clinical notes, making it naturally suited to this domain. The same mean-pooling head is attached. The learning rate is 2×10^{-5}.

Both are trained with AdamW (weight decay 0.01), a cosine learning rate schedule with 10% warmup, an effective batch size of 16 (batch size 4 with gradient accumulation over 4 steps), and mixed-precision training (bfloat16 where available, float16 otherwise). Training always runs for 6 epochs; the checkpoint achieving the highest validation F1 across all epochs is retained.

2.3 Class Imbalance

Insomnia-positive notes constitute a minority class. We address this by weighting the cross-entropy loss inversely by class frequency. Specifically, the weight assigned to the positive class is $w_{\text{pos}} = n_{\text{neg}}/n_{\text{pos}}$, where n_{pos} and n_{neg} are the number of positive and negative training notes, while the negative class weight is fixed at 1.

2.4 Chunk-to-Document Aggregation

At inference, each chunk produces a softmax probability for the positive class. The document-level positive-class probability is the maximum across all chunks:

$$p_{\text{doc}} = \max_{c \in \mathcal{C}} p_c^{(+)}. \tag{2}$$

This max-pooling strategy favours recall: if any chunk receives a high positive-class probability, the document is flagged, consistent with the clinical setting where missing a true positive is more costly than a false alarm.

2.5 Alpha Sweep and Final Training

After both models are trained, the blending weight α is selected by sweeping five candidate values $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ and evaluating the weighted combination

$$p_{\text{ens}} = \alpha \cdot p_{\text{Qwen}} + (1 - \alpha) \cdot p_{\text{BERT}} \tag{3}$$

on the validation set, keeping the α that maximises binary F1 on the positive class. No other hyperparameters are tuned at this stage; learning rates, batch size, and epoch count are all fixed constants. After α is fixed, both models are retrained from scratch on the union of training and validation data using the same fixed hyperparameters, and the final test predictions are produced with the selected α.

3 System 2: Multi-Task Joint Model for Classification and Span Extraction

The architecture diagram of System 2 is provided in Figure 2.

3.1 Sentence-Level Relevance Filtering

Clinical discharge summaries are long and contain large sections unrelated to insomnia (e.g., surgical history, laboratory values). Before chunking, we apply a sentence-level filtering step to concentrate the model on insomnia-relevant content. Sentences are extracted using a biomedical spaCy model (`en_core_sci_sm`[5]). If the note contains 90 or fewer sentences (i.e. $\leq 3K$ where $K = 30$), it is passed through unchanged. Otherwise, each sentence is embedded with sentence transformers (Reimers and Gurevych, 2019) (`all-MiniLM-L6-v2`[6]), cosine similarity is computed against a fixed prototype string covering insomnia symptoms, sleep complaints, and hypnotic medications (e.g., *zolpidem*, *trazodone*, *quetiapine*), and the top $K = 30$ most similar sentences are selected. A context window of ± 1 sentence

[5] https://allenai.github.io/scispacy/

[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

around each selected sentence is added to preserve local coherence, and the resulting sentences are concatenated in their original order.

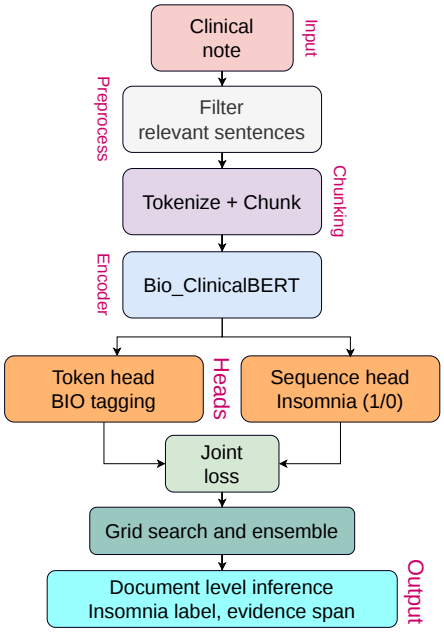

Figure 2: Overview of System 2, a multi-task Bio_ClinicalBERT framework jointly performing insomnia classification and evidence span extraction using BIO tagging.

3.2 Tokenisation and Chunking

The filtered text is tokenised without special tokens, and character-level offset mappings are preserved per token. Chunks of up to 500 tokens are created with an overlap of 100 tokens (stride of 400 tokens). Storing offset mappings per chunk allows exact character-span recovery during inference.

3.3 BIO Span Labelling

The four insomnia criteria are treated as named entity types under a BIO tagging scheme, following the IOB tagging framework introduced by Ramshaw et al. (Ramshaw and Marcus, 1995). For each rule, a token is marked `B-{rule}` if it is the first token of a gold span, `I-{rule}` if it continues a span, and `O` otherwise. Token membership in a span is determined by character-level overlap between the token's offset range and the annotated span. With four rules, the full label set has $4 \times 2 + 1 = 9$ classes. Padding tokens are assigned the ignore index -100 so the loss function disregards them.

3.4 Dual-Head Architecture

A single Bio_ClinicalBERT encoder is shared between two task-specific linear heads:

- **Token head:** a linear layer projecting each token hidden state $\mathbf{h}_t \in \mathbb{R}^d$ to 9 BIO logits. Drives span extraction.
- **Sequence head:** a linear layer projecting the `[CLS]` token representation $\mathbf{h}_0$ to 2 logits. Drives binary classification.

The two heads are trained jointly; the total loss is

$$\mathcal{L} = \mathcal{L}_{\text{token}} + \mathcal{L}_{\text{seq}}, \tag{4}$$

where $\mathcal{L}_{\text{token}}$ is a cross-entropy loss over BIO labels (padding ignored) and $\mathcal{L}_{\text{seq}}$ is a class-weighted cross-entropy over insomnia labels. Because notes containing evidence spans are also positive examples for insomnia, the two objectives provide mutually reinforcing gradient signal.

3.5 Training Details

Training uses AdamW (default weight decay) with a linear warmup-then-decay schedule (10% warmup), gradient accumulation over 2 steps, and gradient norm clipping at 1.0. Hyperparameters: learning rate $\in \{2 \times 10^{-5},\ 5 \times 10^{-5}\}$, batch size $\in \{2, 4\}$, and number of epochs $\in \{15, 20\}$, are selected by an exhaustive grid search over all $2 \times 2 \times 2 = 8$ combinations on the validation set, with early stopping (patience $= 3$ epochs) applied to each candidate, monitoring chunk-level insomnia F1 on the validation set.

3.6 Seed-Diversified Ensemble

Using the best hyperparameters found above, three model instances are trained on the combined train+validation set with different random seeds ($\{42, 123, 999\}$) for the full epoch count without early stopping. Using multiple seeds introduces diversity from random weight initialisation and mini-batch ordering, without requiring different architectures or data partitions.

3.7 Document-Level Inference and Span Aggregation

Insomnia prediction. Each model produces a softmax probability $p^{(+)}_{m,c}$ for each chunk c and model m. For each chunk, probabilities are first averaged across models, then averaged across chunks

to give the document score:

$$p_{\text{doc}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{M} \sum_{m=1}^{M} p_{m,c}^{(+)}, \tag{5}$$

and the document is labelled positive if $p_{\text{doc}} > 0.5$.

Span aggregation. For each chunk and each rule, each model's BIO predictions are decoded to character-level spans. A span is retained only if predicted by a strict majority of models ($\geq \lfloor M/2 \rfloor + 1 = 2$ out of $M = 3$). Surviving spans across all chunks are deduplicated, sorted by start offset, and merged: any two spans where the later one starts at or before the earlier one ends are joined into a single contiguous span.

3.8 Span Evaluation Metrics

Predicted spans are evaluated against gold spans under two criteria:

- **Exact match:** a predicted span (s, e) is a true positive only if an identical gold span exists.
- **Partial match:** a predicted span is a true positive if it has any character-level overlap with any gold span, i.e. $\max(s_{\text{pred}}, s_{\text{gold}}) < \min(e_{\text{pred}}, e_{\text{gold}})$.

Both metrics report micro-averaged precision, recall, and F1 across all four rules and all documents.

4 Results

Our best results for Subtask 1 and Subtask 2 are presented in Table 1 and Table 2, respectively. We compare our test-set performance against the mean and median Codabench scores across all participating teams, as provided by the shared task organizers via email. Our team name is "NoviceTrio".

4.1 Subtask 1

Our best Subtask 1 system achieved strong recall performance.

Our system surpasses both the mean and median recall scores, demonstrating strong sensitivity in detecting insomnia-positive clinical notes.

4.2 Subtask 2

Our best Subtask 2 submission outperformed both the mean and median scores across all evaluation metrics.

The results suggest that the proposed multi-task framework effectively captures both rule-level insomnia indicators and clinically relevant evidence spans.

System	Precision	Recall	F1
System 1	0.5455	**0.9474**	0.6923
System 2	0.4483	0.6842	0.5417
Shared Task Mean	0.7336	0.6935	0.6805
Shared Task Median	**0.8333**	0.6842	**0.7037**

Table 1: Subtask 1 test-set performance (F1 computed on the positive insomnia class). Both System 1 and System 2 were submitted to Codabench; System 1 achieved the better performance and is therefore considered our primary system for Subtask 1. The shared task mean and median, computed across all participating teams, are reported by the organisers.

System	Label Classification	Exact Match	Partial Match
System 2	**0.6444**	**0.4472**	**0.5093**
Shared Task Mean	0.5888	0.3129	0.4584
Shared Task Median	0.6000	0.3586	0.4524

Table 2: Subtask 2 test-set performance. Label Classification is micro-averaged F1 over the four rule labels. Exact Match and Partial Match are micro-averaged span F1 scores. The shared task mean and median are reported by the organisers, computed across all participating teams.

4.3 Ablation Study

We conduct a systematic ablation study across six design axes for System 2: relevance-based sentence filtering, encoder backbone, multi-task learning objective, ensemble size, chunk overlap, and class-weight balancing. Full results and per-axis analysis are provided in Appendix A.1.

5 Conclusion

We presented two systems for automated insomnia detection from clinical notes. System 1, a heterogeneous ensemble of Qwen3-4B and Bio_ClinicalBERT, achieves high recall (0.9474) through max-pooling aggregation and imbalance-aware training. System 2, a multi-task Bio_ClinicalBERT model with seed-diversified ensemble training, exceeds the shared task mean and median across all Subtask 2 metrics. Limitations include a single train–validation split without cross-validation, fixed sentence-filtering heuristics that may discard relevant evidence spans, and the absence of a systematic ablation for System 1. Future work should explore learnable filtering, cross-validation, and threshold tuning to better balance precision and recall.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Appendix

A.1 Ablation Study

We conduct a comprehensive ablation study to quantify the contribution of each design decision in our pipeline. Because the official Codabench evaluation server imposed a strict submission limit (maximum of three submissions per team), hyperparameter selection and ablation experiments were conducted exclusively on the validation split. After selecting the final configuration, models were retrained on the combined training and validation sets to maximize the amount of supervised data available for the final test submission. All variants are trained on the official training set and evaluated on the validation set. Three random seeds (42, 123, 999) are used and their predictions are ensembled by alpha-weighted blending (Subtask 1) and span-union (Subtask 2). Performance is reported on four metrics: Subtask 1 F1 (**S1**); Subtask 2 label micro-F1 (**S2-Lbl**); Subtask 2 exact-match span F1 (**S2-Exact**); and Subtask 2 partial-match span F1 (**S2-Part**). Table 3 summarises all results.

A.2 A – Relevance-Based Sentence Filtering

Rationale. MIMIC-III clinical notes are lengthy documents covering many clinical domains. To reduce noise and limit GPU memory pressure, our full system applies a semantic sentence filter: sentences are ranked by cosine similarity to an insomnia-specific prototype embedding (constructed from symptom terms and hypnotic medication names), and only the top-K sentences plus a one-sentence context window are retained.

Result. Removing the filter (*A – No Sentence Filter*) causes a sharp drop in Subtask 1 F1 ($0.73 \rightarrow 0.55$, -0.18), indicating that the filter effectively suppresses irrelevant clinical content and focuses the model on insomnia-relevant passages. However, the unfiltered variant achieves markedly higher span extraction scores (S2-Exact: 0.62 vs. 0.45; S2-Part: 0.72 vs. 0.59). This reveals a tension: aggressive filtering can discard sentences that contain gold evidence spans, hurting span recall even while it sharpens overall classification. Future work could explore softer filtering strategies (e.g., reranking rather than truncation) to retain classification gains without sacrificing span coverage.

A.3 B – Encoder Backbone

Rationale. We use Bio_ClinicalBERT as our default encoder because it is pre-trained on MIMIC-III discharge summaries, making it particularly suited to clinical language. This setting examines whether domain-specific pre-training provides a genuine advantage over general-purpose `bert-base-uncased`.

Result. Swapping to base BERT (*B*) slightly improves S1 F1 (0.76 vs. 0.73), but substantially degrades span extraction (S2-Exact: 0.39 vs. 0.45; S2-Part: 0.45 vs. 0.59). The S1 improvement is likely a statistical artifact of the small validation set rather than a genuine signal. The consistent span extraction degradation confirms that clinical-domain pre-training is critical for the fine-grained token-level reasoning required by Subtask 2, where the model must locate character-level evidence within highly specialised clinical prose.

Variant	S1 F1	S2-Lbl F1	S2-Exact F1	S2-Part F1
Full System	0.72	0.85	0.45	0.59
A – No Sentence Filter	0.54	0.9	0.62	0.72
B – Base BERT Backbone	0.76	0.79	0.39	0.45
C1 – Seq Head Only (no NER)	0.64	0.75	0	0
C2 – Token Head Only (no Seq Cls)	0	0.84	0.55	0.65
D1 – Single Model (no ensemble)	0.66	0.86	0.49	0.64
D2 – 2-Model Ensemble	0.64	0.83	0.51	0.58
E1 – Zero Chunk Overlap	0.7	0.83	0.46	0.57
E2 – 50-Token Overlap	0.73	0.87	0.44	0.63
F – No Class-Weight Balancing	0.7	0.85	0.44	0.57

Table 3: Ablation study results for the proposed system.

A.4 C – Multi-Task Learning Objective

Rationale. Our full system uses a dual-head architecture that jointly optimises a sequence classification head (for Subtask 1) and a token-level NER head using BIO tagging (for Subtask 2 span extraction). We ablate each head independently to measure the mutual benefit of joint training.

Result. When only the sequence classification head is retained (*C1 – Seq Head Only*), S2-Exact and S2-Part are identically 0.00 by construction: without a token-level NER head the model has no mechanism to produce character offset predictions, so zero is the only architecturally possible outcome rather than a sign of poor optimisation. S1 F1 also drops relative to the full system (0.64 vs. 0.73), providing evidence that the NER objective supplies useful auxiliary supervision that strengthens the shared encoder representations even for the classification task. Symmetrically, retaining only the token-level NER head (*C2 – Token Head Only*) produces an S1 F1 of exactly 0.00 by the same logic: without a sequence classification head the model emits no document-level insomnia prediction, making a non-zero S1 score impossible regardless of training quality. Within Subtask 2, this variant achieves higher span scores than the full system (S2-Exact: 0.55; S2-Part: 0.65), likely because the encoder can specialise entirely for token-level reasoning without sharing capacity with the classification objective. Taken together, these results demonstrate that the two heads are mutually irreplaceable: joint training is the only configuration that enables both subtasks simultaneously, and the auxiliary signal from each head benefits the other.

A.5 D – Ensemble Size

Rationale. We ensemble $N = 3$ models trained from different random seeds to reduce variance and improve calibration. This experiment evaluates whether the benefit saturates quickly ($1 \rightarrow 2$ models) or whether all three seeds are necessary.

Result. A single model (*D1*) achieves surprisingly competitive span scores (S2-Exact: 0.49; S2-Part: 0.64) but lower S1 F1 (0.67). Moving to two models (*D2*) does not consistently improve over the single model—S1 F1 remains at 0.64 and S2-Part drops to 0.58—suggesting that a poorly-aligned second seed can introduce noise. The three-model ensemble recovers S1 F1 to 0.73 and S2-Part to 0.59, consistent with the variance-reduction intuition that odd-sized ensembles with diverse seeds are preferable to even-sized ones where votes can tie. The monotonic improvement in S1 F1 from $N = 1$ to $N = 3$ validates the ensembling strategy for the classification subtask.

A.6 E – Chunk Overlap

Rationale. Long clinical notes are segmented into fixed-length chunks of 500 tokens. Overlapping adjacent chunks ensures that evidence spans crossing chunk boundaries are not missed. We test overlap values of 0, 50, and 100 tokens (the default).

Result. Zero overlap (*E1*) modestly reduces both S1 F1 (0.70) and S2 span scores relative to the full system, consistent with the expectation that non-overlapping chunking causes boundary spans to be split or missed. The 50-token overlap (*E2*) matches the full system on S1 F1 (0.73) and improves S2-Lbl (0.87) and S2-Part (0.63), though S2-Exact slightly lags (0.44). The 100-token default provides the best S2-Exact score (0.45), suggesting that wider context around boundaries helps precise span boundary recovery. The overall differences across overlap settings are modest, implying that the sentence filter already reduces note length enough to limit the boundary-truncation problem.

A.7 F – Class-Weight Balancing

Rationale. The insomnia *yes* class is a minority in the training set. We apply inverse-frequency class weights to the cross-entropy loss to mitigate this imbalance. This ablation investigates whether such weighting is necessary.

Result. Removing class weights (*F*) reduces S1 F1 by about 0.03 (0.69 vs. 0.73), confirming that without re-weighting the model is biased toward predicting the majority *no* class. Subtask 2 metrics are largely unchanged (S2-Lbl and S2-Part differ by < 0.02), which is expected because the NER head operates at the token level where the *O* tag dominates regardless of document-level class balance. Class-weight balancing is therefore a low-cost intervention with a meaningful impact specifically on the classification subtask.