MetaMiners at SMM4H-HeaRD 2026: A Semantic-Structural Knowledge-Enriched Ensemble for SARS-CoV-2 Metadata Identification

Claudia-Alexandra Ursu
University of Bucharest
ursu.claudia99@yahoo.com

Alecsandru Florin Soare
University of Bucharest
alecs.twch@gmail.com

Abstract

This paper presents a hybrid solution for a binary classification of medical PubMed articles created for identifying reports that associate clinical metadata with SARS-CoV-2 genomic sequences. The system is designed to catch the subtle distinction between reports of sequence-associated patient metadata and sentences where such metadata is either unrelated, irrelevant, or linked to previous studies. The biggest challenge is the fact that the medical dataset is highly imbalanced, consisting of only 13.3 % of medical reports labeled positive. Our system proposes a hybrid solution combining four approaches that include dual-evidence tagging, negation-aware suppression, semantic frame extraction, and adversarial training. All these approaches were tested on multiple models: BiomedBERT-base-abstract, BioLinkBERT-large, PubMedBERT-base-fulltext, followed by a best-subset ensemble search to obtain an F1 score of 0.786, setting a new benchmark and positioning the solution in 1st place in the competition.

1 Introduction

One of the core concepts behind genomic epidemiology is that patient's metadata (age, location, symptoms, vaccinated status, history of travel) plays an essential role in understanding how a virus behaves. The Covid-19 pandemic highlighted a double-sided bottleneck. Firstly, in public repositories like GISAID (Grubaugh et al., 2019) or GenBank (Sayers et al., 2019), which store the raw viral sequences, the metadata fields often remain empty. The second side of the issue is related to speed and scale, because during a pandemic, there are thousands of articles published monthly, a volume that would require way too much time for a manual processing in a crisis situation. As a response to this challenge, we introduce a hybrid approach: an ensemble of biomedical transformers fine-tuned through a preprocessing pipeline that includes both semantic and structural analysis. Previous work described by Klein et al. (2025) showed best results on fine-tuned BiomedBERT-Large, achieving an F1 score of 0.776.

2 Methodology

2.1 DataSet

The dedicated corpus for SMM4H-2026 HeaRD Lopez-Garcia et al. (2026) task 5 consists of medical articles that report SARS-CoV-2 sequences, totalling 22,147 sentences, partitioned into 15,504 for training, 2,214 used for validation and 4,429 for testing. Due to the sparse representation of metadata (13.3% only positive labeled sentences), we implemented a rigorous preprocessing pipeline to prevent the model from developing a bias toward the majority class.

2.2 Feature-Enriched Preprocessing

2.2.1 Value-Enriched Attribute Tags

This is the primary feature engineering layer, filtering raw text into a categorical density map before the model processes the narrative. This step finds patient metadata (age, sex, geography) via regular expression (regex) based functions and adds them into the dataset, extracting the corresponding values. New structured metadata signals are added: specifically [AGE:XX], [SEX:Code], and [GEO:Location]. For example, to maximize the model's ability to identify metadata, a sentence describing a 54-year-old female patient in Philadelphia would be enriched with tags like [AGE:54], [SEX:F], and [GEO:Philadelphia].

2.2.2 Metadata Count Tags

The `[META:N]` tag serves as high-level density counter, aggregating the count of unique, non-negated clinical metadata categories detected via regex functions. Each category contributes a single unit to the overall count, even when

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 272–279

they are value-enriched (e.g., [AGE:77], [SEX:M], [GEO:Philadelphia]). The tags are grouped into six thematic domains: (1) demographics and identity ([AGE], [SEX], [RACE]); (2) clinical status and severity ([SYMPTOM], [SEVERITY], [VITALS], [COMORBID], [RISK]); (3) interventions ([TREATMENT], [HOSPITAL], [VACCINE], [LAB]); (4) observation metrics ([VIRAL_LOAD], [DURATION], [OUTCOME]); and (5) geographic context ([LOCATION], [TRAVEL]). Negated metadata is explicitly excluded from the [META:N] density count to minimize false positives.

2.2.3 Anti-Metadata Categories

The [ANTI:M] tags act as a "negative evidence" filter designed to help the model identify sentences that belong to Label 0 even though medical data is present. So these patterns search for technical, laboratory, or bioinformatic terms that usually indicates the report is about data processing or viral research rather than about a specific human case. The regex functions identify three distinct domains of technical noise: bioinformatics - [BIOINF], laboratory experiments - [IN_VITRO] and molecular & genomic analysis -[MOLECULAR].

2.2.4 Verb Detection

Reporting verbs associated with patient outcomes and clinical status (e.g., present, diagnose, hospitalize) are tagged with [REPORTS], while procedural verbs indicative of sequencing and bioinformatics protocols (for example amplify, align, deposit) are marked as [PROCEDURE]. In instances where sentences contain evidence of both clinical reporting and laboratory processes, a composite [REPORTS+PROC] tag is applied. This pattern-matching transformation helps the model process the semantic overlap between patient descriptions and technical documentation.

2.2.5 Biomedical Entity Highlighting

This preprocessing step identifies specific clinical and medical terms using regular expressions and wraps them in XML-style tags, in order to highlight the high-importance tokens that may be strongly correlated with patient metadata. When the match is found, it replaces it with its value included between categorical tags around it. Tags included: DISEASE, DEMOGRAPHIC, MEASUREMENT, DRUG, VACCINE_ENT, BODYPART.

2.2.6 Focus Patterns

This regex based preprocessing step is used to reflect the topic of the sentence, based on the grammatical subject. There are 5 different categories of topics: human-subjects (PATIENT_FOCUS), laboratory protocols or sequencing steps (METHOD_FOCUS), variants or mutations of the virus (VIRUS_FOCUS), epidemiological trends (EPI_FOCUS) and diverse studies (STUDY_FOCUS). This technique is particularly useful in cases where a sentence is likely a technical description even if it contains medical terms. So, for example, a tag like [METHOD_FOCUS] would make the difference between a laboratory protocol description and a patient's report, maximizing the chances of the model to label it as 0.

2.2.7 Negation Detection

The system also utilizes a regex-based detector to identify negation cues like "no" or "without", triggering a specialized filter that examines a 60-character window following the cue. Any metadata keywords found within this scope are transformed with a [NEG] prefix to signify that the attribute is being denied rather than confirmed. Excluding these negated tokens from the [META:N] score prevents laboratory procedures (e.g., "no contamination", "no symptoms in mice") from being mistaken as patient reports, filtering out thousands of potential false positives.

2.2.8 Semantic Frame Tags

This preprocessing step utilizes scispaCy library, which is specialized in parsing biomedical data (Neumann et al., 2019) to analyze the grammatical relationships between subjects, verbs, and objects within a sentence (version scispaCy 0.6.2). Its purpose is to verify that clinical metadata is explicitly linked to a human patient rather than appearing in a technical or procedural context, which is essential for filtering out false positives related to medical procedures. The [FRAME:*] tag categorizes the overall relationship between the sentence's subject and its main verb. For example, [FRAME:patient_report] indicates a human subject performing a communicative action (e.g., "The man described..."), while [FRAME:sample_method] indicates a lab object undergoing a procedure (e.g., "The RNA was extracted..."). The [PATIENT→*] tag uses dependency parsing to confirm that a specific metadata attribute is grammatically linked to a patient subject. It moves beyond key-

word presence to prove "ownership," such as [PATIENT→SYMPTOM] to confirm the patient has a cough, rather than the sentence simply mentioning "cough" in a general research context.

2.3 Weighted Ensemble Classifier

The classification architecture utilizes a strategic weighted ensemble of four high-performing transformer models. This approach leverages the diversity of different pre-training objectives and specialized training refinements to robustly distinguish patient metadata in SARS-CoV-2 genomic studies. Each of the 4 components of the ensemble has its own specific processing steps, described below.

2.3.1 BioLinkBERT-Large With Focal Loss

To handle the dataset's significant imbalance, the model BioLinkBERT-Large[1] was fine-tuned using Focal Loss. This specialized loss function improves performance by reducing the importance of "easy" negative samples (sentences clearly lacking metadata) and forcing the model to concentrate on the "hard" positive samples that are difficult to distinguish. The preprocessing part includes the metadata category prefix injector that scans the text for keywords related to 17 specific categories. It also performs biomedical entity highlighting to identify and wrap key clinical terms including diseases, drugs and measurements. The best F1-score for this model on the validation set was 0.822, with a positive label threshold of 0.36.

2.3.2 BiomedBERT-Base With Semantic Frame Extraction

This component of the ensemble integrates structural linguistic insights by including dependency-parsed semantic frames. It moves beyond simple keyword matching by analyzing the relationship between subjects (for example "patient" vs. "sample") and reporting verbs. The model used is biomedbert-base[2] and is trained using a weighted cross-entropy loss to compensate for the imbalance. By including tags for patient-dominant polarity and predicate-argument compositions, this model excels at identifying the "who" and "what" of a sentence, ensuring that metadata is correctly attributed to human patients rather than laboratory procedures. The best F1-score for this model alone on the validation set was 0.815, with a positive label threshold of 0.40.

2.3.3 Adversarially Robust BioLinkBERT-Large

Constructed upon the BioLinkBERT-Large model, this third ensemble member uses a training strategy called the Fast Gradient Method (FGM). During the fine-tuning stage, the model's word embeddings are intentionally shifted by tiny, calculated amounts. This process forces the model to ignore minor differences in how scientists write and instead focus on broader, more important language patterns. To make the model even more reliable, label smoothing is applied to prevent it from becoming too certain of its own guesses, which helps it perform better on new, unseen data. As a result, this component provides a strong foundation for the ensemble, remaining accurate even when the writing style changes. The best F1-score obtained for this model on the validation set was 0.800, with a positive label threshold of 0.55.

2.3.4 PubMedBERT-Fulltext With Layer Freezing

The final ensemble component is a specialized version of PubMedBERT base[3] that was pre-trained using both research abstracts and full-text articles. This broad exposure allows the model to better understand the specific language and technical details typically found in the methods and results sections of genomic studies. To protect these valuable pre-trained features while training for this specific task, we used a layer freezing strategy. In this setup, the bottom 8 encoder layers remained untouched, while only the top 4 layers and the final classification head were updated. This technique ensures the model keeps its deep biomedical knowledge intact while focusing its remaining learning power on identifying the specific patterns used to report patient metadata. The best F1-score achieved with this model alone on the validation set was 0.819, with a positive label threshold of 0.41.

2.3.5 Ensemble Architecture

The complete classification methodology employs a weighted ensemble consisting of the above four specialized transformer models. By combining these diverse models through a weighted average of their predicted logits, assigning weights of 0.182 to the focal-loss BioLinkBERT-Large, 0.273 for the BiomedBERT-B (Semantic) model and the BioLinkBERT-L (FGM) and 0.272 for the

[1] https://huggingface.co/michiyasunaga/BioLinkBERT-large

[2] https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract

[3] https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext

PubMedBERT-FT (Layer freez.), the system captures a broad range of linguistic signals while minimizing the impact of dataset imbalance.

2.3.6 Development History and Configuration Plans

While a formal, isolated ablation study was not performed due to competition time constraints, we trace the incremental impact of our pipeline components via our chronological development history. This progress is captured by Figure 2, showing a sequence of independently trained runs, each adding a new step. For this, we define some notations, to reflect our experiments.

We note plan A for metadata category injection by regex-based detection of 17 metadata categories. If any category keyword fires, the corresponding tag (for example[AGE], [SEX], [SYMPTOM]) is prepended. Multiple tags may fire simultaneously. More details about these metadata tags are found in Appendix F (Table 6).

We note another plan A-v2 for value-enriched tags, geolocation expansion and [META:N] tags. It also injects concrete values that include: [AGE:46], [SEX:M/F], [GEO:Brazil]. It prepends a density header '[META:N] [ANTI:M]' counting positive-metadata vs. anti-metadata count tags. Serves as the base for all optional add-on operations listed below.

Finally we note plan C the biomedical entity highlighting. It includes inline XML-style tagging of six entity types within the sentence: <DISEASE>, <DEMOGRAPHIC>, <MEASUREMENT>, <DRUG>, <VACCINE_ENT>, <BODYPART>. This step is applied on top of plan A or plan A-v2 in every configuration.

While our used configurations are defined here, a conceptualized but unimplemented approach (Plan B) is detailed in Appendix D as future work.

3 Results

Final predictions are generated by applying a positive threshold of 0.55, which was optimized to achieve an F1-score of 0.851 on the validation dataset and an F1-score of 0.786 on the test dataset, resulting in the first place in the shared task 5 of SMM4H-HeaRD 2026. We provide granular validation metrics (Precision, Recall, and F1) across individual components and ensemble in Table 1. The corresponding threshold sensitivity curves, precision-recall curves, and exhaustive confusion matrices are detailed in Appendix A and

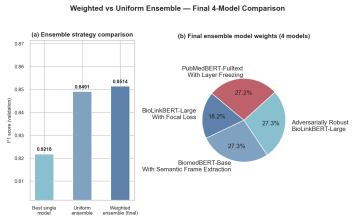

Figure 1: Ensemble analysis: (a) validation F1 for the best single model, uniform ensemble, and final weighted ensemble; (b) ensemble weight distribution;

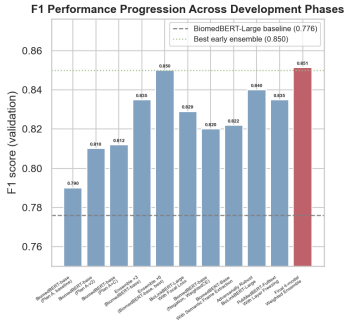

Figure 2: F1 performance progression across development phases

Appendix C. Error analysis indicates that aggregate density tokens ([META] / [ANTI]) remain a primary source of false positives, motivating type-specific or directional replacements in future iterations.

Table 1: Per-model and ensemble validation-set results

Model	Threshold	P	R	F1
BioLinkBERT-L (Focal)	0.36	0.798	0.847	0.822
BiomedBERT-B (Semantic)	0.40	0.786	0.847	0.815
PubMedBERT-FT (Layer freez.)	0.41	0.806	0.833	0.819
BioLinkBERT-L (FGM)	0.55	0.804	0.796	0.800
Weighted ensemble	**0.55**	0.836	0.867	**0.851**

4 Conclusion

By achieving an F1 score of 0.786, our hybrid transformer ensemble sets a new benchmark and secures first place in the shared task. The system successfully overcomes the severe 13.3% positive class imbalance by pairing feature-enriched preprocessing (metadata tags, negation suppression, and anti-metadata signals) with diverse training objectives, such as focal loss, adversarial perturbation, and layer freezing across four transformer models. Error analysis shows density tokens ([META] and [ANTI]) cause false positives via premature aggregation. Future work will replace these raw counts with granular, type-specific markers to improve precision.

References

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34(5):301–310.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Nathan D. Grubaugh, Jason T. Ladner, Philippe Lemey, Oliver G. Pybus, Andrew Rambaut, Edward C. Holmes, and Kristian G. Andersen. 2019. Tracking virus outbreaks in the twenty-first century. *Nat Microbiol*, 4:10–19.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Huang, Kevin Gimpel, Wei-Chen Chen, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Ari Z. Klein, Davy Weissenbacher, Karen O'Connor, Amir Elyaderani, Ivan Flores Amaro, Takeshi Onishi, Su Golder, Kaelen Spiegel, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2025. Detection of patient metadata in published articles for genomic epidemiology using machine learning and large language models.

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.

Lluís Màrquez, Xavier Carreras, Kenneth C Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling via framenet, verbnet and propbank. *Natural Language Engineering*, 14(1):59–109.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations (ICLR)*.

A. Mohammed and R. Kora. 2023. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35:757–774.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Eric W. Sayers, Mark Cavanaugh, Karen Clark, James Ostell, Kim D. Pruitt, and Ilene Karsch-Mizrachi. 2019. Genbank. *Nucleic Acids Research*, 47(D1):D94–D99.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

A Threshold Sensitivity and Precision-Recall Analysis

Figure 3 plots validation F1 vs. classification threshold for each model (a) and for the ensemble (b). The ensemble curve is notably flatter near its peak, reducing threshold sensitivity. Figure 4 shows precision-recall curves.

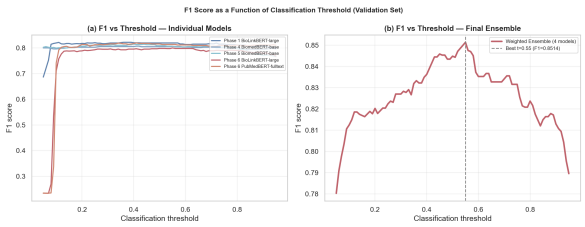

Figure 3: F1 score vs. classification threshold for each model and the final weighted ensemble (validation set).

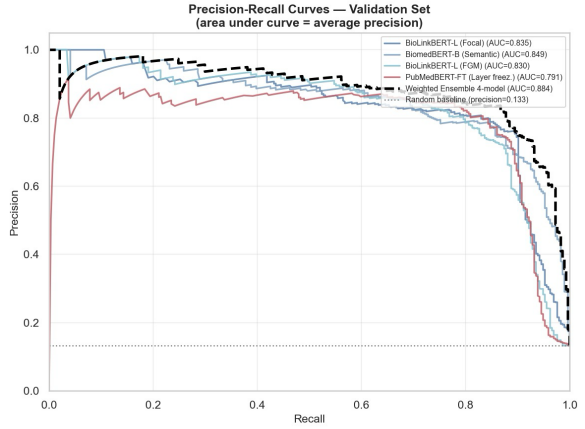

Figure 4: Precision-recall curves for each model and the final weighted ensemble (validation set).

B Confusion Matrices

Figure 5 shows per-model confusion matrices on the validation set (294 positive / 1,920 negative). Figure 6 shows the weighted ensemble confusion matrix. Table 2 summarizes the counts.

Table 2: Confusion matrix counts (validation set).

Model	TN	FP	FN	TP
BioLinkBERT-L (Focal)	1857	63	45	249
BiomedBERT-B (Semantic)	1852	68	45	249
BioLinkBERT-L (FGM)	1863	57	60	234
PubMedBERT-FT (Layer freez.)	1861	59	49	245
Weighted Ensemble	**1870**	**50**	**39**	**255**

Figure 5: Per-model confusion matrices (validation set)

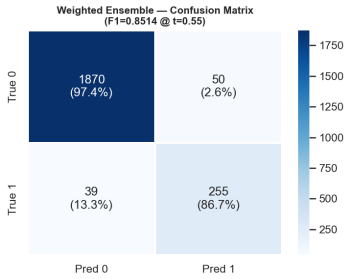

Figure 6: Weighted ensemble confusion matrix (validation set, $t = 0.55$).

C Error Analysis

The `[META:N]/[ANTI:M]` density-header tokens are a systematic driver of false positives: they encode category count but discard which categories fired and their syntactic role (*premature aggregation*). Replacing raw counts with type-specific signals (e.g., `[META_DEMO]`, `[META_CLIN]`) is a priority for future work. This limitation is clearly reflected in the training set distribution shown in Figure 7, where `META` and `ANTI` tags vastly dominate the preprocessing landscape with 15,504 occurrences each. Because these tokens injected into every sentence of the dataset, the models develop a high baseline bias toward predicting the positive class whenever these high-frequency markers are encountered.

D Unimplemented Approaches and Future Work

Beyond our implemented configurations, we also conceptualized **Plan B** (Contextual Window Modeling), which aimed to incorporate context by appending neighboring sentences from the article in a

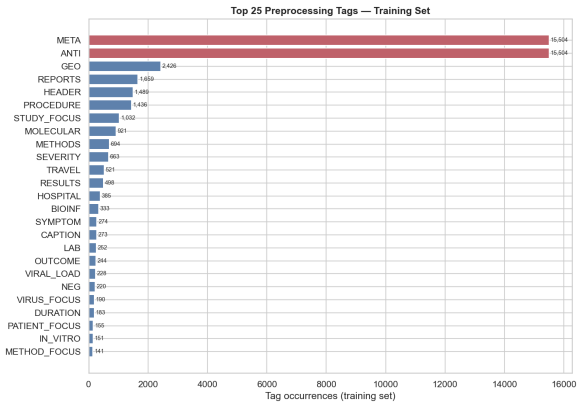

Figure 7: Tag frequency across the training set

structured way. However, this approach was deprioritized and left unimplemented, but still remains a target for future work related to this topic. The risks associated with this approach would be the injection of noise data from neighboring sentences while the data set is already imbalanced and reaching GPU memory limits.

E Reproducibility Details

Random seed. No explicit seed was passed to `TrainingArguments`; HuggingFace Trainer defaults to `seed=42`.

Hardware. NVIDIA GeForce RTX 4070 Laptop GPU (8.6 GB VRAM), AMD Ryzen AI 9 HX 370 CPU, 64 GB RAM.

Shared hyperparameters. Table 4 outlines the shared hyperparameter configurations applied uniformly across all models during training.

F Regex Pattern Reference

Table 6 lists all evaluated metadata tag categories. The exact regular expression definitions used for all of them, as well as the contextual negation rules, and anti-metadata extraction patterns are publicly available within the `code-utils` directory of our GitHub repository.

Table 3: Per-model training configuration. BS = batch size, GA = gradient accumulation steps, MaxL = max sequence length, WCE = weighted cross-entropy, LS = label smoothing.

Model	Ep.	BS	GA	MaxL	Loss	Extra
BioLinkBERT-L	16	4	4	256	Focal (γ=2)	–
BiomedBERT-B	10	8	2	512	WCE	–
PubMedBERT-FT	16	16	1	256	WCE	Freeze 8/12 layers
BioLinkBERT-L	10	4	4	256	WCE	FGM ε=1.0, LS=0.05

Table 4: Shared hyperparameters (all models).

Parameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Warmup ratio	0.1
Weight decay	0.01
Max. gradient norm	1.0
Precision	fp16
Eval. strategy	Per epoch
Early stopping	Patience 3 (val. F1)
Pos. class weight	4.0 (Weighted CE models)
Classifier head	Single linear layer

Table 5: Measured training wall-clock times (RTX 4070 Laptop GPU).

Model	Epochs	Time
BioLinkBERT-L (Focal)	16	5h 02m
BiomedBERT-B (Semantic)	9†	1h 07m
PubMedBERT-FT (Layer freez.)	15†	39m
BioLinkBERT-L (FGM)	10	5h 58m
Total		**13h 36m**

†Early stopping triggered.

Table 6: Positive-metadata tag categories (`METADATA_PATTERNS`, 17 total).

Tag	Matches
`[AGE]`	Age values (e.g., “54-year-old”, “months old”)
`[SEX]`	Gender terms (male, female, man, woman)
`[RACE]`	Ethnicity (Hispanic, Caucasian, Asian, etc.)
`[SYMPTOM]`	Clinical symptoms (fever, cough, dyspnea, etc.)
`[SEVERITY]`	Disease severity (mild, severe, ICU, intubated)
`[VIRAL_LOAD]`	Viral quantification (CT value, copies/mL)
`[DURATION]`	Temporal course (duration, days positive)
`[LAB]`	Serology/testing (antibody, IgG, PCR result)
`[VITALS]`	Physiological signs (blood pressure, SpO2)
`[TREATMENT]`	Drug/therapy (remdesivir, dexamethasone)
`[HOSPITAL]`	Hospitalization (admitted, ICU, length of stay)
`[OUTCOME]`	Patient outcomes (death, survived, recovered)
`[COMORBID]`	Comorbidities (diabetes, hypertension, obesity)
`[RISK]`	Risk factors (smoking, pregnant, elderly)
`[VACCINE]`	Vaccine-related (vaccinated, BNT162b2, booster)
`[LOCATION]`	Geographic residence (city, province, rural)
`[TRAVEL]`	Travel history (travel, flight, quarantine)