TermHunter: Neural Biomedical Term Extraction and Ontology-Enhanced Definition Generation with Structured Prompting

Nina Hosseini-Kivanani[1,2,*,†], Rossella Resi[3,†]

[1]University of Luxembourg, Belval, Luxembourg
[2]Radio Télévision Luxembourg, Luxembourg
[3]University of Innsbruck, Innsbruck, Austria

Abstract

We present TermHunter, a cascaded system for the DETECH 2026 shared task on biomedical term extraction and definition generation in the gut–brain interplay domain. For Task A, we combine BiomedBERT-large and BioLinkBERT-large sequence taggers through a confidence-calibrated token ensemble and compare this neural baseline with several linguistic post-processing modules for boundary correction, confidence filtering, lexical filtering, and abbreviation recovery. For Task B, we apply the strongest Task A extractor to the official definition-generation abstracts and generate definitions for the extracted in-corpus concepts using BioBART-v2-large, concept-type-aware prompting, retrieval augmentation, self-consistency decoding, and optional LLM-based re-ranking. On the held-out validation split, the best Task A configuration is the raw neural ensemble, which reaches 0.9545 Micro-F1 and 0.9624 Type-F1. Additional post-processing mainly increases recall or linguistic coverage but does not improve exact-match F1. For Task B, the best configuration is a retrieval-augmented best-fold BioBART model, achieving 0.0714 BLEU-4 and 0.5828 BERTScore-F1. Overall, the results show that controlled neural extraction and compact retrieval-grounded generation are more effective than heavier post-hoc correction in this shared-task setting.

Keywords

Biomedical term extraction, definition generation, terminology extraction, linguistic post-processing, retrieval-augmented generation, structured prompting, BioBART

1. Introduction

The automatic extraction of domain-specific terminology remains a central challenge in natural language processing, especially in specialized biomedical domains where precision, consistency, and conceptual clarity are essential. In such settings, even small boundary errors or lexical mismatches can reduce the usefulness of extracted terminology for downstream applications such as knowledge organization and information retrieval [1, 2]. The DETECH shared task, organized within the HEREDITARY project[1], addresses this problem in the domain of gut–brain interplay [3]. The shared task comprises two complementary subtasks. Task A focuses on automatic term extraction, with the goal of identifying every instance of single-word and multi-word domain-relevant terms in the corpus. Task B focuses on definition generation, with the goal of producing well-formed natural language definitions for the concepts designated by the extracted terms [3]. Together, these tasks capture two closely related aspects of terminology processing: identifying domain terms in text and defining the concept they designate in a standardized way.

Task A is challenging because term identification in biomedical text requires more than sequence labeling alone. Systems must determine accurate term boundaries, distinguish valid terms from general language expressions, and handle a range of complex phenomena, including multi-word expressions,

International Workshop on Definition and Term Extraction Challenge (DETECH) 2026, June 24, 2026, Zadar, Croatia
*Corresponding author.
†Both authors contributed equally to this work.
nina.hosseinikivanani@ext.uni.lu;nina.kivanani@rtl.com (N. Hosseini-Kivanani); rossella.resi@uibk.ac.at (R. Resi)
0000-0002-0821-9125 (N. Hosseini-Kivanani); 0000-0002-3261-5662 (R. Resi)

[1]https://hereditary-project.eu/

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-06-17

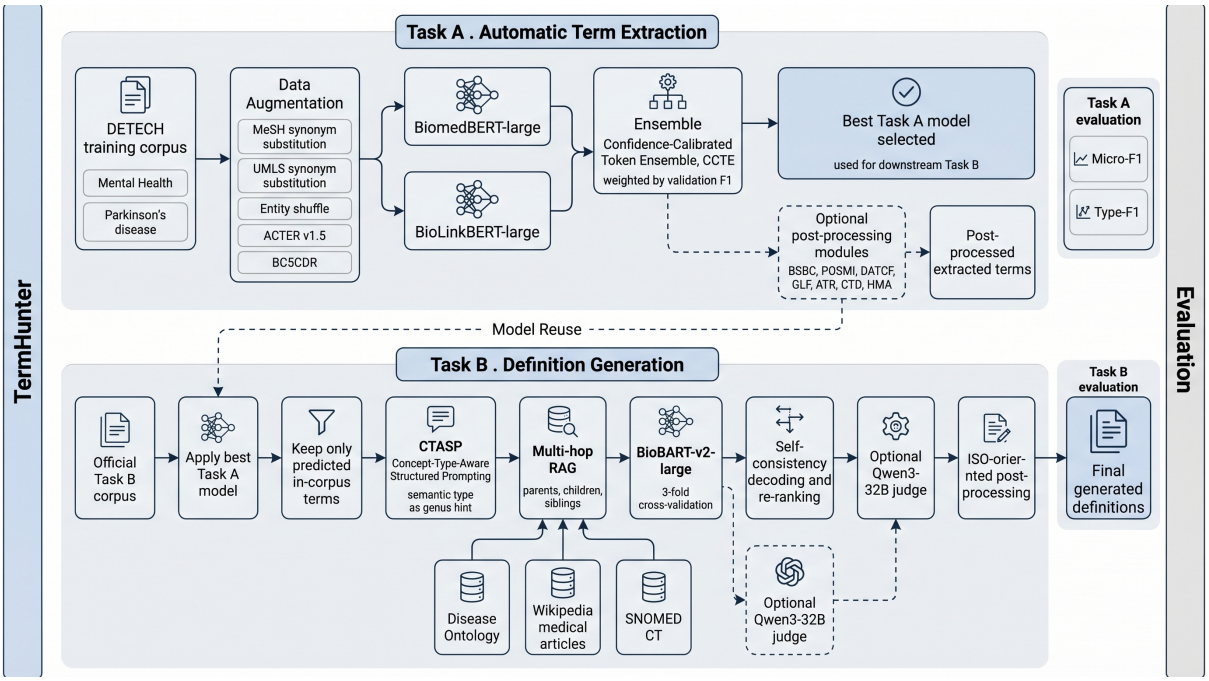

Figure 1: Overview of the TermHunter pipeline. For Task A, two biomedical sequence taggers are fine-tuned and combined through a confidence-calibrated token ensemble, followed by optional linguistic post-processing modules. For Task B, concepts are extracted from the Task B input abstracts using the best Task A configuration, then passed to a BioBART-based definition generator with concept-type-aware prompting, multi-hop retrieval, self-consistency decoding, and optional LLM-based re-ranking.

nested concepts, coordination patterns, and abbreviated forms. These difficulties are amplified in the domain of gut–brain interplay, where the distinction between general and specialized usage may be subtle [1].

Task B introduces an additional level of complexity. Definition generation depends not only on the quality of the extracted term inventory, but also on the system's ability to recover concept-relevant evidence and transform it into lexically precise, structurally well-formed definitions. In practice, this often benefits from combining generation with structured biomedical knowledge and retrieval-based grounding, rather than relying on unconstrained text generation alone [4, 5].

In this paper, we present TermHunter, a cascaded system tailored to the DETECH shared task. For Task A, we investigate transformer-based biomedical term extraction combined with confidence-calibrated ensembling and targeted linguistic post-processing. For Task B, we adopt a pipeline in which the best Task A model is first applied to the official Task B corpus, after which definitions are generated using BioBART, concept-type-aware prompting, and ontology-guided retrieval [6]. Our goal is to evaluate how far a compact but carefully designed pipeline can go in handling both terminology extraction and definition generation under the constraints of the gut–brain interplay domain. This setup allows us to examine whether a strong extraction backbone combined with controlled generation can provide a competitive strategy for both subtasks in a specialized biomedical domain.

2. Methodology

2.1. Overall pipeline

TermHunter addresses the two DETECH 2026 subtasks in a unified pipeline for the gut–brain interplay domain. Task A performs automatic term extraction from PubMed abstracts, while Task B generates natural-language definitions for the extracted concepts. Our primary system is neural but intentionally modular: biomedical encoders form the core of Task A, and ontology-enhanced sequence-to-sequence

Table 1
Task A training resources used in the final pipeline.

Resource	Docs	B-TERM	Role
DETECH official training set	580	in-domain	supervised target corpus
MeSH synonym substitution	1,160	25,371	terminology-preserving augmentation
UMLS synonym substitution	1,160	25,977	terminology-preserving augmentation
Entity shuffle	536	9,856	structure-preserving perturbation
ACTER v1.5	2,432	8,908	external term extraction data
BC5CDR	1,500	28,534	curriculum pre-training corpus
Total	**7,369**		

generation forms the core of Task B. Post-processing and LLM-based re-ranking are introduced as controlled additions rather than as the main performance drivers. Although we also explored a pure LLM setting during development, this paper focuses on the neural and neural-plus-judge pipelines that produced the selected runs. Figure 1 summarizes the overall architecture.

2.2. Task A. Automatic term extraction

The official DETECH 2026 Task A training set contains 580 PubMed abstracts, comprising 356 abstracts in the mental health subdomain and 224 abstracts in the Parkinson's disease subdomain. To reduce the effect of limited in-domain supervision, we augment the official corpus automatically with synonym-based replacements, structure-preserving entity perturbations, and external biomedical term extraction data. Synonym substitution is performed programmatically by querying the MeSH and UMLS thesauri: for each annotated term in the training abstracts, up to two synonymous surface forms are retrieved and substituted at the token level while preserving the BIO annotation; no manual synonym selection is involved. Entity perturbation shuffles annotated spans across abstracts while keeping sentence structure intact, also without manual intervention. We also use BC5CDR for curriculum pre-training, before fine-tuning on DETECH and the augmented in-domain corpus. Table 1 summarizes the resources used in the final setup.

For sequence labeling, we fine-tune two biomedical transformer encoders, BiomedBERT-large[2] [7] and BioLinkBERT-large[3] [8], on a 70% training, 15% validation, and 15% test split. Training is performed for 15 epochs with learning rates of 2e-5 and 1.5e-5, respectively, batch sizes of 8 for training and 16 for evaluation, cosine_with_restarts scheduling, a warmup ratio of 0.15, label smoothing of 0.05, gradient accumulation over 4 steps, and a maximum gradient norm of 1.0. All Task A models are trained on a single NVIDIA A100 40 GB GPU. We apply early stopping with patience 4 on the validation entity F1 and use fp16 precision throughout training. The final ensemble combines token-level probabilities using validation-F1-weighted averaging and excludes models whose validation F1 falls below 0.85. To strengthen biomedical adaptation, we adopt a curriculum setting in which BC5CDR is used before fine-tuning on DETECH and the augmented in-domain corpus.

After CCTE decoding, we apply a modular post-processing pipeline. The full system supports ten stages: Biomedical Span Boundary Correction (BSBC), POS-Based Modifier Inclusion (POSMI), Domain-Aware Term Confidence Filtering (DATCF), General Lexicon Filter (GLF), Abbreviation Term Rescuer (ATR), Gazetteer Boosting (GAZ), Coordinated Term Decomposition (CTD), Head-Modifier Expansion (HMA), KeyBERT Keyphrase Boosting (KKB), and a final analysis stage used only for linguistic logging. In practice, the shared-task comparison concentrates on representative configurations built from these modules rather than on the full combinatorial space.

BSBC repairs common BIO decoding errors, including illegal I-without-B transitions, detached hyphenated spans such as *gut–brain*, and incomplete biomedical suffix attachment. POSMI expands spans leftward when an adjectival modifier is likely part of the biomedical term. DATCF filters low-

[2] https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-large-uncased-abstract
[3] https://huggingface.co/michiyasunaga/BioLinkBERT-large

confidence predictions using average token confidence, with a default threshold of 0.3. GLF targets spuriously extracted single-word general-language items using two constraints, general-language frequency and corpus document frequency. Multi-word terms always pass this filter. ATR recovers abbreviations when explicit abbreviation definitions or abbreviation-as-modifier patterns are detected. GAZ, which was available but not used in the selected Task A runs, scans the text for ontology-derived or gold-derived terms that may have been missed by the tagger.

Structural components, HMA and CTD, are retained below because they provide the key linguistic motivation for decomposition and modifier redistribution: The HMA module supports the structural implicitness in English MWEs. It is implemented using SciSpaCy part-of-speech tagging and dependency parsing and examines the internal syntactic structure of multi-word expressions by identifying the head and its modifiers. Specifically, it profiles recurrent term patterns, such as ADJ+NOUN and NOUN+NOUN, and assigns each token a structural role, distinguishing the semantic core, head, from its descriptive elements, modifiers. For example, in the phrase *intestinal epithelial barrier*, the module identifies *barrier* as the syntactic head, while *intestinal* and *epithelial* are classified as modifiers. This supports downstream tasks such as controlled decomposition.

The CTD module, implemented with SciSpaCy dependency parsing, leverages syntactic dependency relations, specifically *conj* (coordination), *cc* (coordinating conjunction), *amod* (adjectival modifier), and *compound* (nominal modifier), to identify and decompose coordinated structures within multi-word terms. The approach uses a heuristic strategy, defaulting to decomposition in cases of shared heads. For example, given the coordinated MWE *diagnostic and molecular biomarkers*, the model preserves the complete entity while also generating the two specific terms *diagnostic biomarkers* and *molecular biomarkers*. It also defaults to a wide-scope interpretation in cases of shared modifiers. For example, given the coordinated MWE *αS aggregation and propagation*, the model preserves the full expression and decomposes it into *αS aggregation* and *αS propagation*, since the modifier may independently scope over each conjunct. Both the full entities and the decomposed terms are then subjected to downstream validation. By integrating CTD into the extraction pipeline, the model moves beyond surface-level entity recognition, as in NER, toward structurally informed decomposition that captures explicit and implicit terminological units embedded in coordinated expressions.

2.3. Task B. Definition generation

For Task B, the system generates natural-language definitions only for concepts that actually occur in the official Task B input abstracts. This design decision is important. Rather than assuming an externally fixed concept inventory, we first apply the best-performing Task A configuration, namely `run_2_ccte_only`, to the Task B corpus and then define only the extracted concepts. At test time, this step is performed on the 12 Task B-specific abstracts in `task_Definition/test/abstracts/`[4], comprising 6 Parkinson-related and 6 mental-health-related documents. The definition generator is based on `biobart-v2-large`[5], trained with 3-fold cross-validation. In addition to the DETECH [9] gold definitions, we augment the Task B training material with definitions from UMLS, Disease Ontology, Wikipedia medical articles, BioPortal/OLS, and a small number of LLM-generated examples. This yields a total of 630 definitional training instances, summarized in Table 2.

Our prompting strategy uses Concept-Type-Aware Structured Prompting, or CTASP. Each concept is mapped to a coarse semantic type, such as *disorder, chemical, gene, biological process, anatomy, organism, therapy, or diagnostic* concept. This type label is injected into the prompt as a genus hint, guiding the generator toward an intensional genus–differentia formulation instead of a looser definition.

To ground generation, we use multi-hop retrieval-augmented generation over the Disease Ontology graph. For each concept, the system retrieves graph-neighbor evidence from parents, children, and siblings, then compares and supplements it with additional evidence from Disease Ontology definitions, Wikipedia medical articles, and SNOMED CT. This step is intended to improve the factual specificity while keeping the evidence closely tied to the target concept. Decoding is performed with

[4]https://github.com/gmdn/DETECH2026/tree/main/task_Definition/test/abstracts
[5]https://huggingface.co/GanjinZero/biobart-v2-large

Table 2
Task B resources and generation configuration.

Section	Item	Specification
Definition resources		
Resources	DETECH definitions	315 supervised in-domain definitions
Resources	UMLS definitions	159 external biomedical definitions
Resources	Disease Ontology definitions	23 ontology-grounded definitions
Resources	Wikipedia medical articles	81 supplementary domain definitions
Resources	BioPortal/OLS	36 terminology-resource definitions
Resources	LLM-generated definitions	16 synthetic augmentation examples
Resources	Total	630 training definitions
Generation configuration		
Configuration	Base model	`biobart-v2-large`
Configuration	Cross-validation	3-fold, shuffled, fixed random seed
Configuration	Optimization	1e-5 learning rate, 18 epochs, batch size 6
Configuration	Training controls	cosine_with_restarts, warmup 0.10, label smoothing 0.02, gradient accumulation 4
Configuration	Early stopping	patience 2 on evaluation loss
Configuration	Precision mode	bf16 on Ampere+ GPUs, fp32 otherwise
Configuration	Fold weighting	inverse validation loss
Configuration	Prompting	CTASP
Configuration	Retrieval	multi-hop RAG over Disease Ontology, Wikipedia, and SNOMED CT
Configuration	Decoding	SCDR, 4 nucleus candidates + 1 beam candidate
Configuration	LLM judge	Qwen3-32B, used in judge-based variants
Configuration	Test-time concept source	CCTE extraction from the 12 Task B test abstracts

Self-Consistency Decoding with Re-ranking, or SCDR. For each concept, four candidate definitions are generated with nucleus sampling at temperatures 0.6, 0.7, 0.8, and 0.9, using `top_p=0.92` and `top_k=50`. A fifth candidate is generated with beam search using `num_beams=5` and `length_penalty=1.2`. Candidates are then re-ranked using a composite score that combines evidence faithfulness, inter-candidate agreement, concept-type coverage, length control, and a circularity penalty. For selected variants, we add an LLM judge based on Qwen3-32B. The judge generates an additional candidate and re-ranks the available set. If the external API is unavailable, the pipeline returns to the underlying neural prediction. Finally, all generated definitions are post-processed with fine-tuned prompting to better align with ISO 1087 and ISO 704 standards. This stage enforces several constraints: leading determiners are removed, trailing full stops are stripped, and any definitions that redundantly include the defined term (or tautological expressions) are penalized during the candidate selection process, thereby promoting non-circular formulations.

2.4. Evaluation protocol

Task A is evaluated with Micro-F1 and Type-F1 under case-insensitive exact matching, without lemmatization, following the DETECH annotation protocol. This choice is particularly important for coordinated expressions and singular/plural variation, because predicted terms must match the gold span surface form rather than a normalized lemma. Task B is evaluated with BLEU-4 and BERTScore, allowing us to assess both lexical similarity and semantic similarity between generated and reference definitions. In addition, our implementation follows the key DETECH protocol constraints directly in the pipeline. Single-word and multi-word terms are both extracted through BIO tagging. Coordination is handled through CTD, shared modifiers through HMA, term matching is case-insensitive, and no stemming or lemmatization is applied. For Task B, the generator is explicitly encouraged toward ISO-style intensional

definitions, and post-processing removes leading determiners and trailing sentence-final punctuation, while reducing circularity.

The system addresses synonymy and terminological variation at two levels. At training time, MeSH and UMLS synonym substitution (Table 1) exposes both BiomedBERT-large and BioLinkBERT-large to the same concept under varied surface forms, encouraging robust boundary prediction across paraphrases. At inference time, the ATR module recovers abbreviation–expansion pairs that the tagger may miss, linking short forms to their expanded equivalents. However, the evaluation protocol uses case–insensitive exact surface matching without lemmatization or synonym normalization. Consequently, if a gold annotation uses one variant and the system predicts a synonymous form, the prediction is scored as a false positive. This is a known limitation of strict exact-match evaluation in biomedical term extraction; partial-credit or synonym-aware evaluation would give a more complete picture of system coverage.

3. Results and Discussion

3.1. Task A. Automatic Term Extraction

Table 3 reports the five selected Task A runs. The runs differ in their use of boundary correction, confidence filtering, lexical filtering, and abbreviation recovery. The strongest Task A configuration is the raw Confidence-Calibrated Token Ensemble, which achieves the highest Micro-F1, Type-F1, and precision. This indicates that the weighted combination of BiomedBERT-large and BioLinkBERT-large is already well calibrated for this biomedical term extraction setting [7, 8]. The result is not merely recall-driven: the CCTE-only system also obtains the highest precision, making it the most reliable upstream extractor for Task B. Adding Biomedical Span Boundary Correction and linguistic rules (CTD and HMA) lowers overall F1 while leaving recall largely unchanged. In our setting, this suggests that boundary repair and decomposition may recover plausible spans, but also introduce false positives or slightly misaligned boundaries that are penalized under exact-match evaluation [10]. The same pattern remains visible when DATCF is added on top of BSBC, CTD, and HMA. In fact, the identical scores obtained by CCTE + BSBC and CCTE + BSBC + DATCF indicate that confidence filtering brings no measurable benefit in the reported subset, at least under the present thresholding and decoding setup. Qualitative output analysis demonstrated that a system with additional linguistic components remains viable, while highlighting the sensitivity of exact-match metrics.

By conducting a systematic error analysis of the output, it becomes evident that although the decompositions are linguistically sound, they are penalized by exact-match scoring. Given that multi-word expressions constitute more than 42% of the extracted training terms, this penalty can be substantial. When processing *diagnostic and molecular biomarkers*, the CTD correctly preserves the full entity while additionally generating the decomposed terms *diagnostic biomarkers* and *molecular biomarkers*. The same applies to *preclinical and clinical studies*, which are decomposed into *preclinical studies* and *clinical studies*, as well as *central nervous system diseases and disorders*, which are decomposed into *central nervous system diseases* and *central nervous system disorders*. However, if the gold annotation contains only the complete form without these decompositions, the system is penalized despite producing linguistically valid outputs.

The question of whether decomposed or nested terms should be retrieved during term extraction depends largely on the intended application and may therefore fall within the broader scope of term extraction itself. If term extraction is intended to support hierarchical ontologies that enable knowledge inference, question-answering, or applications requiring fine-grained semantic reasoning [11, 12], then handling the compositional meaning captured by nested concepts and recognizing implicit hierarchies can substantially benefit downstream tasks. Decomposition, for example, supports the identification of subordinate and sibling concepts. In medical domains, multi-level hierarchical information helps distinguish between broader categories (*disease*) and more specific manifestations (*Parkinson's disease* and *Alzheimer's disease*). However, not all extraction tasks require nested concepts. Flat term lists may be sufficient for large-scale text mining tasks where rapid extraction is prioritized, for domains with

Table 3
Selected Task A configurations and their evaluation results on the 15% held-out split of the official training corpus (cross-validation setting). Scores do not correspond to the official shared-task test set.

Run	BSBC	DATCF	GLF	ATR	Micro-F1	Type-F1	Precision	Recall
Run 1	No	No	No	No	0.9545	0.9624	0.9500	0.9591
Run 2	Yes	No	No	No	0.9423	0.9460	0.9268	0.9584
Run 3	Yes	Yes	Yes (0.75/2)	No	0.9376	0.9403	0.9222	0.9536
Run 4	Yes	Yes	No	No	0.9423	0.9460	0.9268	0.9584
Run 5	Yes	No	No	Yes	0.8917	0.8778	0.8322	0.9604

relatively homogeneous concept structures, for terminology databases or quick-reference resources, for applications focused primarily on terminology identification without deeper semantic reasoning, or for term extraction targeting non-hierarchical ontologies.

Since the scope of term extraction in this study is not explicitly defined, the modularity and ablation analysis of CTD and HMA become particularly important. Given that no cases of over-decomposition or incorrect head assignment were identified during output evaluation, it would be valuable to introduce a secondary evaluation metric that rewards linguistically plausible decompositions rather than penalizing them. The current evaluation could be complemented with a partial-credit scoring scheme in which correctly decomposed terms receive credit even when they diverge from the gold-standard surface form. Alternative F1 scores based on normalized variants could also be computed; for example, *diagnostic and molecular biomarkers*, *diagnostic biomarkers*, and *molecular biomarkers* could be treated as equivalent during evaluation. The resulting precision–recall trade-offs between strict and relaxed evaluation metrics could then be compared.

Alternatively, separate F1 breakdowns could be reported specifically for coordinated structures. In addition, it would be informative to quantify what proportion of false positives produced by CTD/HMA variants consist of linguistically well-formed decompositions as opposed to genuinely spurious extractions. If decompositions account for, for example, 40% of the observed penalty, the interpretation shifts from viewing post-processing as reducing performance to recognizing it as a valuable component for targeted term extraction under an integrated evaluation protocol.

The more engineered neural pipeline with GLF (0.75/2) also underperforms the raw ensemble. Tuning GLF (0.60, 0.65, 0.70, 0.75) and applying different `wordfreq` thresholds (2, 3, 5, 7) have been shown to have only marginal effects, keeping the neural pipeline inherently less effective than the others. From a discussion perspective, this is an important result. While post-processing and filtering are commonly explored in NER and term extraction pipelines, our Task A results show that, in this shared-task setting, a strong neural ensemble is more effective than heavier rule-based correction [13, 14]. Once the base extractor is already highly accurate, additional constraints may over-correct biomedical expressions rather than improve them.

The abbreviation-oriented configuration, CCTE + BSBC + ATR, exhibits the clearest precision-recall trade-off. It reaches the highest recall among the reported variants, but at a substantial precision cost, resulting in the lowest Micro-F1 and Type-F1. This suggests that abbreviation recovery increases coverage, but in its current form, it may be too permissive. Since abbreviation identification, expansion, and disambiguation are known to be challenging in biomedical text, this behavior is not surprising [15, 16, 17]. Under exact-match evaluation, that precision loss outweighs the recall gain [10].

Overall, the Task A results support the following conclusion: the most effective submission strategy is to rely on the confidence-calibrated ensemble itself, and to treat post-processing as optional rather than central. The representative runs show that structural and abbreviation-oriented rules can change recall behavior, but none of them improve on the raw CCTE system. This also motivates our downstream Task B design, where selecting the most reliable extractor is more valuable than maximizing term count through aggressive expansion.

Table 4
Selected Task B configurations and their evaluation results.

Run	Folds	RAG	LLM Judge	BLEU-4	BERTScore-F1	BERTScore-R	Avg. Length	Length Ratio
Run 1	Top-3	No	No	0.0668	0.5809	0.5788	14.9500	0.8650
Run 2	Best	Yes	No	0.0714	0.5828	0.5661	11.5700	0.6690
Run 3	Top-3	No	Yes	0.0498	0.5733	0.6041	21.2900	1.2310
Run 4	All 3	No	Yes	0.0498	0.5733	0.6041	21.2900	1.2310
Run 5	Best	Yes	Yes	0.0500	0.5764	0.6044	21.1700	1.2250

3.2. Task B. Definition Generation

Table 4 reports the five selected Task B runs. The runs differ along three dimensions: fold selection, retrieval augmentation, and use of an LLM judge. The strongest Task B configuration is the best-fold BioBART model with retrieval augmentation. It achieves the highest BLEU-4 and BERTScore-F1 among the selected variants, making it the most balanced definition-generation setting. This result suggests that retrieval-supported generation helps the model produce definitions that are both lexically closer to the references and semantically better aligned with them. Because all variants share the same BioBART-v2-large backbone, the observed difference is not attributable to architecture alone, but to the interaction between fold quality and external knowledge support [6, 4, 18, 5].

An important aspect of this best-performing variant is its shorter output profile. The best-fold RAG model produces the shortest definitions on average and the lowest length ratio, yet still obtains the strongest overall quality scores. This indicates that better performance is not achieved by verbosity, but by producing compact and better-targeted definitional statements. In the context of biomedical terminology, this is desirable, since the shared task requires concise formulations rather than verbosity. The result is also consistent with the ISO-style intensional format adopted in our pipeline, where a concise genus–differentia structure is preferred over descriptive elaboration.

The non-RAG neural ensemble remains competitive. The Top-3 ensemble without retrieval reaches a BLEU-4 of 0.0668 and a BERTScore-F1 of 0.5809, which places it very close to the best RAG configuration. This shows that the base generator is already strong even without external evidence. At the same time, the RAG-based model still yields the best overall scores, which suggests that ontology-enhanced retrieval is beneficial when paired with the strongest fold and controlled generation [4, 5, 19]. In other words, retrieval is helpful, but only when it remains tightly aligned with concept-level evidence and does not introduce unnecessary lexical drift.

The judge-based variants show a different pattern. Their recall-oriented semantic coverage is slightly higher, with the best BERTScore-R obtained by the best-fold RAG + Judge variant. However, this gain does not translate into stronger overall performance. Both BLEU-4 and BERTScore-F1 remain below the best neural-only RAG system, while the generated definitions become substantially longer. This suggests that the judge tends to favor broader, more inclusive formulations that capture more semantic material, but at the cost of lexical precision and terminological compactness. Under shared-task evaluation, such verbosity is not rewarded.

A further observation is that the Top-3 ensemble + LLM Judge and the All-folds + LLM Judge variants produce identical scores in the reported evaluation. This indicates that, in the no-RAG judge setting, expanding the ensemble from Top-3 to all folds has no measurable effect on the final output. The dominant factor appears to be the judge's re-ranking behavior rather than the specific ensemble width. This is analytically useful because it shows that the judge is not simply selecting the best candidate from a richer pool, but imposing a relatively stable preference pattern across similar candidate sets.

The CTASP prompting strategy is designed to elicit genuine genus–differentia formulations by injecting a semantic type label as a genus hint. Whether generated definitions follow a strict intensional structure depends, however, on which system variant is considered. The LLM-judge runs produce well-formed genus–differentia definitions with identifiable superordinate classes and differentiating properties. For example, *gut dysbiosis* is defined as *dysbiosis characterized by an altered composition and reduced diversity of microorganisms in the intestinal tract, often involving imbalanced ratios of bacterial*

taxa and linked to inflammatory and neurological conditions; *gut-brain axis* as *communication system characterized by bidirectional neural, hormonal, and immune-mediated signaling between the central nervous system and the gastrointestinal tract, integrating microbial and host-derived signals to coordinate physiological responses*; and *oxidative stress* as *physiological stress characterized by an excess of reactive oxidants and insufficient antioxidant capacity, leading to potential damage to cellular macromolecules and disruption of redox signaling pathways*. In each case the genus phrase opens the definition and the differentiating properties follow in a subordinate clause. By contrast, the best-fold BioBART RAG variant (Run 2), which scores highest on BLEU-4 and BERTScore-F1, frequently maps ontology fragments to incorrect target concepts, producing definitions that are syntactically plausible but semantically misaligned—for instance, assigning an epilepsy definition to *gut-brain axis* or a biodiversity definition to *Parkinson's disease*. The ISO-style post-processing removes leading determiners, strips trailing full stops, and penalises circular formulations, but it cannot correct concept-level retrieval errors. This reveals a limitation of the automatic evaluation: BLEU-4 rewards surface overlap with the reference, and a short ontology snippet that happens to share lexical items with the gold definition can outscore a structurally correct but differently phrased intensional definition. BERTScore-R is a better proxy for this quality dimension, and it is indeed highest in the judge-enhanced runs. Systematic structural evaluation—for example, automatic detection of the genus phrase and differentiating properties—is left for future work.

Overall, the Task B results mirror the pattern observed in Task A. The best system is not the most complex one, but the most controlled one. The strongest configuration combines the best BioBART fold with retrieval augmentation, producing compact definitions with the best BLEU-4 and BERTScore-F1 scores. In contrast, LLM-judge variants increase output length and semantic recall, but this does not translate into stronger overall performance. These results suggest that, for this task, controlled retrieval-grounded generation is preferable to broader judge-driven expansion [4, 5, 19].

4. Conclusion and Limitations

We presented TermHunter, a cascaded system for biomedical term extraction and definition generation in the DETECH 2026 shared task. For *Task A*, the best-performing configuration on the held-out split was the raw confidence-calibrated ensemble of BiomedBERT-large and BioLinkBERT-large, reaching 0.9545 Micro-F1 and 0.9624 Type-F1. Post-processing modules were useful for exploring boundary correction, abbreviation recovery, and structural decomposition, but they did not improve exact-match F1 in the reported runs. This suggests that, in this setting, a strong neural ensemble is more reliable than aggressive post-hoc correction. For *Task B*, the best result was obtained with the best-fold BioBART-v2-large model combined with retrieval augmentation. This configuration produced the most balanced definitions, achieving the highest BLEU-4 and BERTScore-F1 while keeping outputs concise. LLM-judge variants increased output length and semantic recall, but reduced overall lexical and semantic alignment.

The main limitation of the system is its sensitivity to exact surface matching. Coordinated expressions, nested concepts, shared modifiers, abbreviations, and left-headed multi-word expressions remain difficult, especially when linguistically plausible outputs differ from the gold surface form. For definition generation, the main limitations are factual consistency, definition compactness, and variation in specificity. Definition evaluation is challenging when based on comparison with a gold standard, as semantically equivalent definitions are difficult to detect under strict structural constraints. Future work should therefore focus on more conservative structural decomposition, better abbreviation disambiguation, and tighter integration between retrieved evidence and definitional generation.

Code Availability

The implementation and supplementary resources for TermHunter are available at: https://github.com/NinaKivanani/TermHunter

Acknowledgments

This work was also supported by the LuxVoice project, funded by the Luxembourg National Research Fund (FNR) under project reference 19205922.

The presentation of this work at DETECH 2026 in Zadar, Croatia, on 24 June 2026 was supported by COST Action CA22126, **European Network on Lexical Innovation (ENEOLI)**, through a Dissemination Grant awarded to Rossella Resi.

Declaration on Generative AI

Generative AI tools (Grammarly and GPT-4o) were used only for language editing and improving readability during the preparation of this manuscript. These tools were not used to generate core scientific ideas, experimental data, or technical contributions. All authors have thoroughly reviewed and approved the final manuscript and take full responsibility for the integrity of its entire content.

References

[1] H. T. H. Tran, M. Martinc, J. Caporusso, A. Doucet, S. Pollak, The recent advances in automatic term extraction: A survey, arXiv preprint arXiv:2301.06767 (2023).

[2] J. A. Lossio-Ventura, R. Sun, S. Boussard, T. Hernandez-Boussard, Clinical concept recognition: evaluation of existing systems on ehrs, Frontiers in Artificial Intelligence 5 (2023) 1051724.

[3] HEREDITARY Project, Detech 2026 challenge tasks, 2026. Official challenge description for DETECH 2026.

[4] F. Remy, K. Demuynck, T. Demeester, Automatic glossary of clinical terminology: a large-scale dictionary of biomedical definitions generated from ontological knowledge, in: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, 2023, pp. 265–272.

[5] D. Shlyk, T. Groza, M. Mesiti, S. Montanelli, E. Cavalleri, Real: A retrieval-augmented entity linking approach for biomedical concept recognition, in: Proceedings of the 23rd workshop on biomedical natural language processing, 2024, pp. 380–389.

[6] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, S. Yu, Biobart: Pretraining and evaluation of a biomedical generative language model, in: Proceedings of the 21st Workshop on Biomedical Language Processing, 2022, pp. 97–109.

[7] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.

[8] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 8003–8016.

[9] F. Vezzani, G. M. Di Nunzio, V. Bonato, G. Silvello, Overview of the international workshop on definition and term extraction challenge (detech) 2026, in: Proceedings of the International Workshop on Definition and Term Extraction Challenge (DETECH) 2026, CEUR.org, Zadar, Croatia, 2026.

[10] R. T.-H. Tsai, S.-H. Wu, W.-C. Chou, Y.-C. Lin, D. He, J. Hsiang, T.-Y. Sung, W.-L. Hsu, Various criteria in the evaluation of biomedical named entity recognition, BMC bioinformatics 7 (2006) 92.

[11] B. Portelli, S. Scaboro, E. Santus, H. Sedghamiz, E. Chersoni, G. Serra, Generalizing over long tail concepts for medical term normalization, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 8580–8591.

[12] R. Yin, Z. Zhou, Z. Gao, A joint model for hierarchical nested information extraction, IEEE Access 10 (2022) 50985–50995.

[13] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning

models, in: Proceedings of the 27th international conference on computational linguistics, 2018, pp. 2145–2158.
[14] J. M. Giorgi, G. D. Bader, Transfer learning for biomedical named entity recognition with neural networks, Bioinformatics 34 (2018) 4087–4094.
[15] A. S. Schwartz, M. A. Hearst, A simple algorithm for identifying abbreviation definitions in biomedical text, in: Biocomputing 2003, World Scientific, 2002, pp. 451–462.
[16] D. Movshovitz-Attias, W. Cohen, Alignment-hmm-based extraction of abbreviations from biomedical text, in: BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, 2012, pp. 47–55.
[17] Q. Jin, J. Liu, X. Lu, Deep contextualized biomedical abbreviation expansion, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 88–96.
[18] Y. Guo, W. Qiu, G. Leroy, S. Wang, T. Cohen, Retrieval augmentation of large language models for lay language generation, Journal of Biomedical Informatics 149 (2024) 104580.
[19] M. Muhetaer, A. Yusupu, W. Yifan, M. Mutalipu, F. Hao, Medical qa dialogue datasets in rag systems performance evaluation and chatgpt optimization, Scientific Reports 15 (2025) 44467.