A QTT-Informed System for Biomedical Term Extraction and Definition Generation*

Diego A. Burgos[1,*,†], Antonio Tamayo Herrera[2,†], Giovanni Díaz[3,†], and Carlos Mario Pérez-Pérez[4,†]

[1] *Wake Forest University, Winston Salem, NC, USA*
[2] *University of Antioqua, Medellin, Colombia*
[3] *University of Antioqua, Medellin, Colombia*
[4] *National Autonomous University of Mexico (UNAM), Mexico City, Mexico*

Abstract

We describe In2Lab-TNT's participation in DETECH 2026 on automatic term extraction (Task A) and definition generation (Task B) in the biomedical intersecting domains of Parkinson's disease and mental health. Our system is grounded in the quantum theory of terms (QTT), which we operationalize through contextual semantic-state induction and a collapse-based coherence signal. For Task A, we combine a fine-tuned PubMedBERT span extractor with a second-stage contextual refinement module that computes mention embeddings, induces latent states with KMeans, and estimates collapse from centroid distance. For Task B, we map extracted mentions to concepts using training-derived normalization resources and generate definitions with a large language model using collapse-guided contextual evidence. On local development for Task B, the QTT-enhanced generator improves BLEU 1 from 12.63 (baseline) to 17.37 while maintaining similar BERTScore (≈0.80). Official Task A results show that our strongest runs are the recall-oriented fine-tuned extraction systems, while official Task B results show that preserving longer concept units yields better lexical, structural, and semantic definition quality than more aggressive concept splitting, although splitting slightly improves concept recall.

Keywords

Biomedical terminology, automatic term extraction, definition generation, quantum theory of terms, PubMedBERT, large language models

1. Introduction

Term extraction and definition generation remain foundational tasks in computational terminology, biomedical natural language processing, and knowledge representation. In biomedical abstracts, the same conceptual unit may appear as a single word term, a long multi-word term, an acronym, or a morphologically variable form. This makes extraction difficult at the mention level and even more difficult when the goal is to move from surface realizations to normalized concepts and intensional definitions. DETECH 2026 [1] addresses both problems through two complementary tasks: automatic term extraction (Task A) and concept assignment plus definition generation (Task B).

Our participation was guided by the superposition and collapse principles of the quantum theory of terms (QTT) [2][3], which models terms as dynamic semantic objects whose properties become stabilized under contextual observation. In computational terms, we interpret contextual embeddings as realizations of possible semantic states and collapse as a coherence signal that indicates how prototypical or stable a contextual use is relative to other uses. This interpretation allows us to connect both tasks within one framework. In Task A, collapse is used as a contextual filtering signal

* *International Workshop on DEfinition and Term Extraction Challenge (DETECH) 2026, June 24, 2026, Zadar, Croatia*
* Corresponding author.
† These authors contributed equally.
burgosda@wfu.edu (D. Burgos); antonio.tamayo@udea.edu.co (A. Tamayo); giovanny.diaz@udea.edu.co (G. Díaz); carlos.perez@enallt.unam.mx (C. M. Pérez-Pérez)
0000-0002-5784-3952 (D. Burgos); 0000-0002-5984-7463 (A. Tamayo); 0000-0002-0666-6829 (G. Díaz); 0000-0001-6792-322x (C. M. Pérez-Pérez)

for candidate mentions. In Task B, collapse guides evidence selection for LLM-based definition generation.

This paper first reports on some recent relevant work on term extraction and definition generation. Then a brief description of the tasks is provided followed by a system overview and the methodological and experimental set up for both tasks. Next, we present our results, analysis and discussion, and we close with conclusions, limitations, and future work.

2. Related Work

Automatic term extraction and definition extraction have evolved from rule-based pattern matching to data-driven and neural architectures that model specialized language with increasing contextual sensitivity. In ATE, the conceptual foundation remains the distinction between general lexis and domain-specific designation, typically operationalized through the complementary notions of termhood and unithood and framed within broader terminological theories such as Cabré's Communicative Theory of Terminology [4] or ISO standards [5]. Recent surveys show that the field has progressed from linguistic and statistical pipelines to transformer-based sequence labeling and, more recently, prompt-based large language models (LLMs), which now constitute the dominant research frontier [6][7].

Methodologically, term extraction has moved through three major paradigms: linguistic approaches based on morphosyntactic patterns, statistical approaches based on frequency and association measures, and hybrid systems that combine candidate filtering with ranking. Although hybrid methods historically delivered strong results, current state-of-the-art systems rely increasingly on contextualized representations from transformer encoders and on LLM-based few-shot or zero-shot extraction, especially in low-resource or domain-specific settings [8][9]. At the same time, benchmark construction remains a central bottleneck. Resources such as ACTER have substantially improved reproducibility and multilingual evaluation, yet the literature still shows persistent disagreement about what counts as a term, how boundaries should be delimited, and how nested or partially specialized units should be treated [10].

Definition extraction has followed a partially parallel trajectory. Earlier work focused on definitional patterns and sentence-level detection, but contemporary definition extraction treats the task as a structured prediction problem involving sentence classification, span identification, and term–definition linking. The DEFT benchmark, introduced in SemEval-2020 Task 6, was decisive in this transition because it formalized DE as a multi-component task and exposed the complexity of naturally occurring definitions beyond canonical definitional templates [11]. Subsequent joint neural models strengthened this direction by integrating syntactic connection and semantic consistency into unified architectures for simultaneous classification and extraction [12]. Most recently, LLM-based pipelines such as SciDef have extended DE into scientific literature mining and from pure extraction to definition generation, showing that multi-step prompting can recover a high proportion of definitions, while also revealing that relevance selection and evaluation fidelity remain major open challenges [13].

Across both term and definition extraction, the strongest recent methodological insight is that model improvement alone does not guarantee comparable progress. Along these lines, [14] reinforces that extraction research remains limited by the absence of sufficiently standardized gold standards, operational definitions of the target unit, and stable boundary-annotation criteria. That argument, while formulated for term extraction, extends directly to definition generation, where disagreement may concern not only whether a sentence is definitional, but also the exact delimitation of the term and its definition. Accordingly, the current state of the art in both fields is best characterized as a convergence of transformer- and LLM-based extraction, manually validated benchmark design, BIO-style sequence labeling, span classification, and increasing emphasis on reproducible evaluation across domains and languages.

Lastly, work on quantum-inspired approaches to language, cognition, and information retrieval (see for example [15]) has attracted growing scholarly interest, leading to substantial advances in several interconnected areas. Yet, these developments have generally remained fragmented and have not converged into a unified theory of the term comparable to the quantum theory of terms that informed the present proposal. Within QTT, semantic superposition reframes the term as a semantic system in itself, rooted in the theoretical foundations of terminology while simultaneously challenging traditional models that assume stability, discreteness, and pre-established meaning. It conceptualizes the term as a dynamic semantic entity whose properties emerge only through contextual actualization rather than existing in a fully determined state beforehand; it is this framework that guides the methodology for the system proposed here.

3. Shared Task Overview and Data

DETECH 2026 targets biomedical terminology in the intersecting domains of Parkinson's disease and mental health. Task A requires mention-level extraction of single-word and multi-word terms from biomedical abstracts. Evaluation uses mention-level Micro-F1 and type-level Type-F1. Micro-F1 rewards exact extraction of occurrences in context, whereas Type-F1 evaluates lightly normalized unique term types.

Task B moves from surface mentions to concept assignment and intensional definition generation. Systems submit a mention-to-concept file and a concept-to-definition file for each domain. The task data provide aligned mention-to-concept mappings and gold definitions for a subset of concepts. These aligned layers were especially important for our mention-to-concept design because they revealed recurrent normalization patterns such as one-to-one mappings for canonical terms, plural-to-singular normalization, acronym expansion, and one-to-many mappings for coordinated or extended sequences. The same data also supplied the definitional material used for local evaluation and for estimating the granularity of latent semantic states that we used as instantiation of QTT's collapse.

4. Unified System Overview

Our system consists of four connected modules across both tasks, namely, neural mention extraction, contextual representation and state induction, concept assignment, and definition generation. A transversal QTT-collapse layer connects all stages. The workflow begins with high-recall extraction, then refines or interprets candidates using contextual coherence. For Task B, training-side concept evidence, including collapse-ranked contexts, is reused to enrich evidence use on LLM (GPT-4o-mini) prompts (see Figure 1).

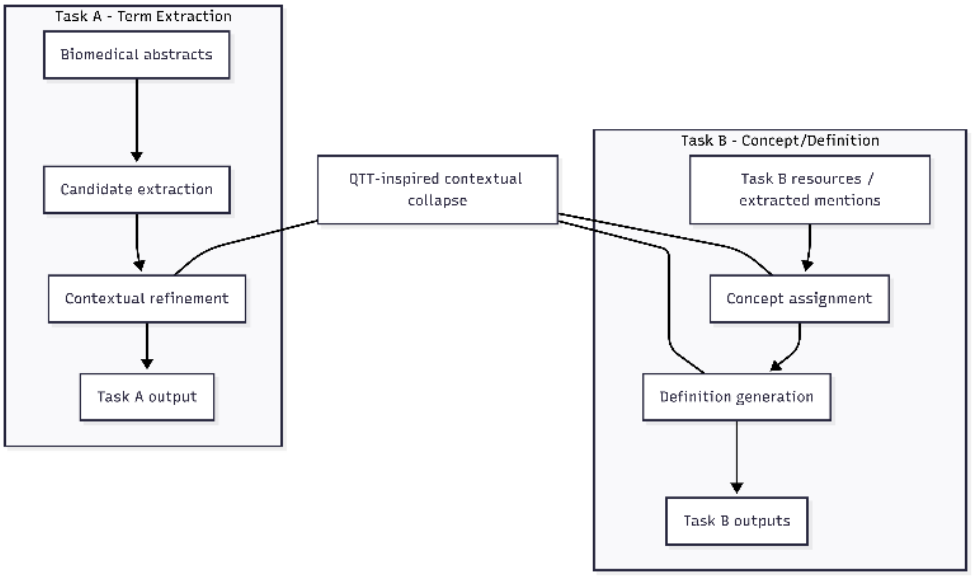

Figure 1: System overview diagram

Stage 1 of Task A predicts candidate term mentions together with confidence scores. Stage 2 re-encodes the full abstract with a frozen PubMedBERT model [16], extracts context-sensitive mention embeddings by mean pooling over the mention span, clusters these embeddings into latent semantic states, and computes collapse as a function of the mention's distance to its assigned centroid. For Task B, predicted mentions are mapped to normalized concepts using training-derived lookup resources, acronym expansion, and constrained decomposition heuristics. Definitions are then generated either by a baseline LLM-only prompt or by a stronger prompt that includes representative high-coherence mention forms, supporting contexts, and domain cues selected from the highest-collapse evidence.

5. Methodology

5.1. Task A

5.1.1. Neural Extraction and Contextual Refinement

We model Task A as a two-stage process. Stage 1 is a neural extractor based on PubMedBERT fine-tuning. We describe the model as a span boundary detector (see Figure 2). Given contextual token representations h_i, the model predicts the probability that token $_i$ begins or ends a term span:

$$p_{start}(i)=\sigma(W_s h_i + b_s), p_{end}(i)=\sigma(W_e h_i + b_e) \quad (1)$$

Candidate spans are created by pairing compatible predicted starts and ends. Each span receives a confidence score computed as the geometric mean of boundary probabilities:

$$confidence(i,j)=\sqrt{p_{start}(i) \cdot p_{end}(j)} \quad (2)$$

This score favors spans whose two boundaries are both reliable. To handle overlapping or nested outputs, candidate spans are sorted by confidence and pruned greedily, keeping the highest-confidence non-overlapping spans.

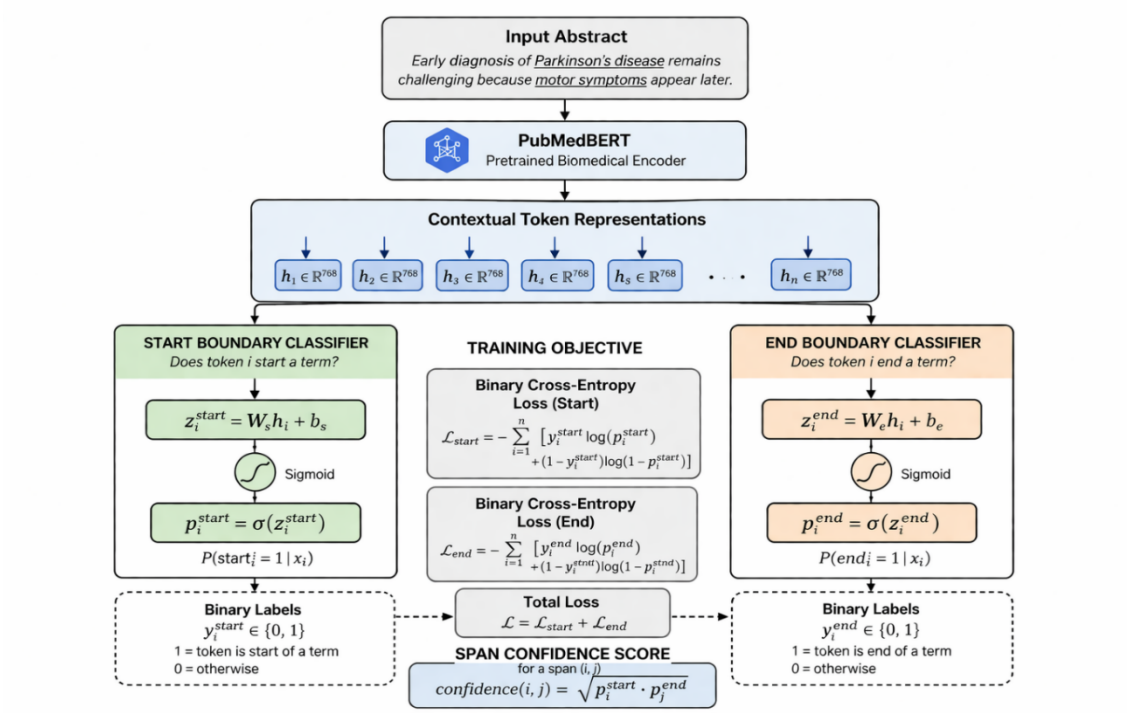

Figure 2: Stage 1 of Task A. Neural term extraction

Stage 2 does not fine-tune the encoder further. Instead, it performs frozen contextual feature extraction plus unsupervised state induction. Each abstract is encoded with PubMedBERT, the predicted mention span is located in its abstract through character offsets, and the token embeddings inside the span are mean-pooled as in (3):

$$v_m = \frac{1}{|S_m|} \sum_{t \in S_m} h_t \qquad (3)$$

This yields one contextual vector for each extracted occurrence, even if it is a multi-word sequence. The same surface term can therefore receive different vectors in different abstracts. We interpret this as contextual semantic variability, coherent with the QTT's superposition principle that multiple semantic states may coexist before contextual stabilization.

5.1.2. State Induction and Collapse

Mention embeddings are standardized and clustered with KMeans. Each cluster is interpreted as a latent semantic state. QTT-inspired collapse is computed from the distance between a mention embedding and its assigned centroid. Let d be this Euclidean distance so that collapse is defined as:

$$collapse = \exp\left(-\frac{d}{T}\right) \qquad (4)$$

where T is a robust temperature estimated from the interquartile range of distances. A higher collapse indicates that a mention is closer to the prototypical center of its state and is therefore treated as a more coherent or stable contextual realization of the mention.

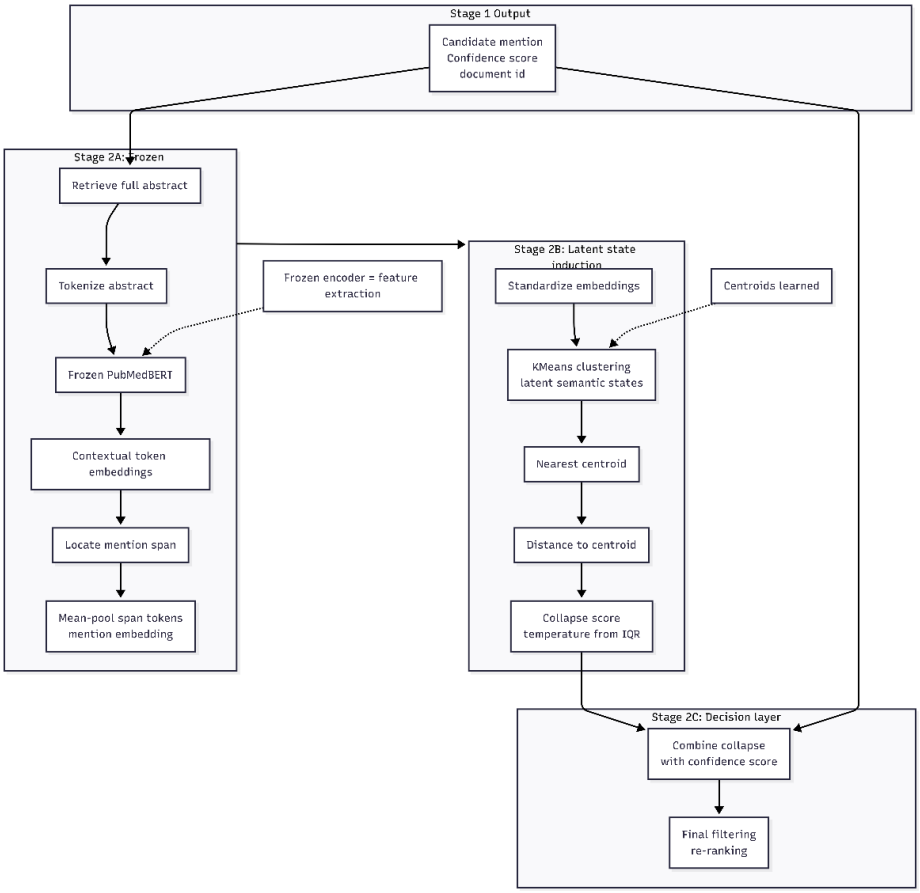

Figure 3: Stage 2 of Task A. Contextual refinement methodology

A central design choice is the number of latent states *K*. Rather than treating *K* as a purely arbitrary hyperparameter, we grounded it in an operative definition of what a term is by [17] and in assumptions about the average extent of specialized definitions proposed by [18]. An analysis of the Task B's training definitions suggested that a concept state is often delimited through approximately five co-defining terms or traits ($k_f = 5$). We also verified that each semantic frame *r* is partially shared by approximately three concepts ($r = 3$) and that a term participates in about five states or conceptual neighborhoods ($s = 5$). This led us to a theoretically motivated range near one hundred states, that is clusters, according to the equation shown in (5), where *N* is the number of concepts in the dataset:

$$K \approx \frac{N \cdot k_f}{r \cdot s} \qquad (5)$$

Empirically, K = 100 also yielded stronger collapse discrimination than smaller values and improved the influence of collapse in both Task A development and Task B prompting.

5.1.3. Decision Rule

We combine Stage 1 confidence and collapse into a unified scoring mechanism:

$$score = \alpha \cdot p_{term} + (1-\alpha) \cdot collapse \qquad (6)$$

Where α is a mixing parameter to control the balance between local neural extraction confidence (p_{term}) and contextual semantic coherence (*collapse*). The final selection of terms follows a two-branch

rule, namely, keep the candidate if $p_{term} \geq p_{keep}$. That is, we trust the neural extractor's performance, therefore we keep terms with confidence (p_{term}) higher than a tested threshold (p_{keep}) or $p_{term} \geq p_{min}$ AND $score \geq \tau$ so that we also set a minimum confidence threshold (p_{min}) and we rescue candidates with p_{term} over that minimum threshold if their score τ meets a minimum tested threshold. In order to determine the best parameters, we ran a grid search tuning.

This design aims at preservation of Stage 1 recall, controlled filtering of false positives, and limited but targeted influence of contextual signals. The system is intentionally conservative as Stage 1 provides coverage but Stage 2 refines only ambiguous cases. As we observed good performance by the neural extractor, the collapse signal is used as a tie-breaker and boundary adjustment mechanism, but not as a primary decision driver (See Figure 3).

5.2. Task B

5.2.1. Concept Assignment

Task B concept assignment is rule-based and training-informed. We build a mention-to-concept dictionary from the official training mappings, as well as an acronym map and a concept vocabulary. The prediction procedure first tries exact and normalized lookup. If a mention is an acronym observed in training, the system retrieves the expanded concept label, otherwise the acronym is kept. For longer sequences, we use constrained decomposition heuristics: we split on conjunctions and selected function words, and for segments longer than two words we optionally create modifier-plus-head candidates inside each segment. Segmentation is local, so modifiers from one segment are never combined with heads from another segment (see examples in Table 1).

We submitted two Task B variants. The long-concept run favored preserving longer concept units. The split-concept run applied more aggressive decomposition heuristics. This difference proved informative in the official results because it revealed a trade-off between concept recall and downstream definition quality.

Table 1
Extracted, long mentions vs split mentions

Mention	Extracted mention / Long concept run	Split mentions / Split concept run
1	reactive oxygen and nitrogen species	<reactive oxygen>
1	reactive oxygen and nitrogen species	<nitrogen species>
2	microbiota-derived extracellular vesicles	<microbiota-derived vesicle>
2	microbiota-derived extracellular vesicles	<extracellular vesicle>

5.2.2. Definition Generation

We implemented two definition generators. The baseline is a LLM-only prompt requesting a single intensional biomedical definition in an ISO, genus-differentia style. The stronger system is collapse-guided. It does not simply print numeric collapse values into the prompt; instead, it uses collapse to rank and select the evidence shown to the model. For each concept seen in training, we gather supporting mention variants, local contextual snippets, collapse values, and state assignments. We then select the highest-collapse contexts, extract frequent domain cues from them, and build the final prompt from representative mention forms, representative high-coherence contexts, and frequent domain cues. Thus, collapse influences generation indirectly through evidence curation.

At inference time, if a predicted test concept also exists in the training concept groups, the system uses the collapse-guided generator. Otherwise, it falls back to the baseline generator. This distinction matters when interpreting results as the strongest local gains were obtained on concepts with rich

training evidence, whereas the final submission necessarily mixes collapse-guided definitions for seen concepts with baseline definitions for previously unseen ones (see Figure 4).

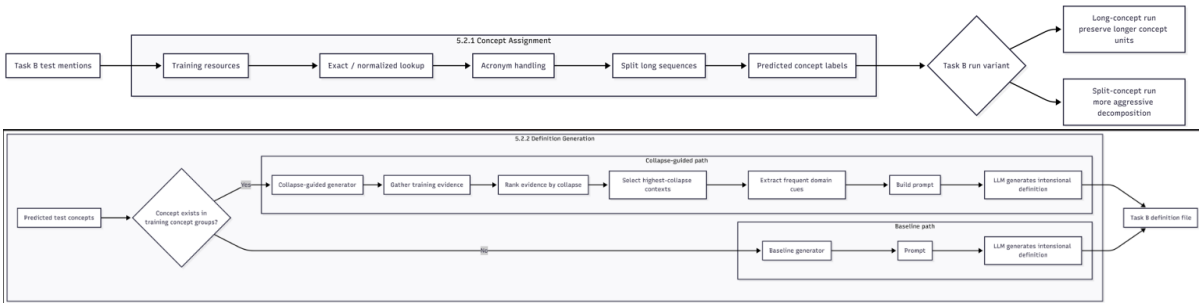

Figure 4: Methodology of Task B

6. Experimental Setup

For Task A, we experimented with several runs that varied in how nested terms and overlapping spans were handled. Run 4 corresponds to the fine-tuned PubMedBERT extractor with nested terms controlled by selecting the highest-confidence span when overlapping candidates were produced. This became our best official Task A run in both domains. For local development, we also evaluated collapse-guided filtering parameters to understand how contextual coherence reshaped the decision boundary.

For Task B, we conducted local evaluation on training concepts with gold definitions, comparing an LLM-only baseline to the QTT-enhanced generator. Official Task B submissions were then evaluated by the organizers with concept recall and matched-pair definition metrics: BLEU-1, BLEU-2, BLEU-3, ROUGE-1, ROUGE-2, ROUGE-L F1, and BERTScore F1. The comparison between long concepts and split concepts was particularly useful for understanding the impact of concept assignment on downstream definition quality.

7. Results

7.1. Task A Official Results

The most important Task A result is that Run 4 was clearly the strongest system on the official hidden evaluation. In Parkinson's disease, Run 4 reached Micro-F1 = 0.6522 and Type-F1 = 0.6600. In mental health, it reached Micro-F1 = 0.6561 and Type-F1 = 0.6703. Relative to the more conservative runs, Run 4 sacrificed precision but gained substantially in recall, and the recall gain was large enough to dominate overall performance. This indicates that, on the hidden set, a recall-oriented extraction strategy was more beneficial than a stricter precision-oriented filtering regime.

The official results also show a domain pattern already visible in development: mental health remains slightly harder overall than Parkinson's disease, especially in the more conservative runs. Still, the best run is strong in both domains and confirms the competitiveness of the fine-tuned extraction component.

7.2. Task A Local Development Observations

Our local collapse-oriented development experiments tell a complementary story. The collapse-guided system reached Micro-F1/Type-F1 of 0.8446/0.7636 on Parkinson's disease and 0.8153/0.7495 on mental health, with micro precision close to 0.90 in both domains. These development results support the claim that collapse is useful as a precision-oriented coherence signal. However, the official Task A results show that the hidden evaluation favored a broader recall profile. Taken together, these findings suggest that collapse is effective for reshaping borderline decisions, but its utility depends strongly on how recall is inherited from the neural extractor.

7.3. Task B Official Results

The official Task B results compare two runs: Run 1 with longer concept units and Run 2 with more aggressive concept splitting. The pattern is strikingly consistent across the two domains. In Parkinson's disease, Run 1 outperforms Run 2 on BLEU-1, BLEU-2, BLEU-3, ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore F1, while Run 2 slightly improves n_pairs and concept recall. The same happens in mental health. For example, in Parkinson's disease the long-concept run obtains BLEU-1 = 21.82, ROUGE-L = 0.232, and BERTScore F1 = 0.793, compared with 20.29, 0.223, and 0.790 for the split-concept run. In mental health, the long-concept run again dominates all definition-quality metrics, although the split-concept run yields a slightly higher concept recall.

This means that concept splitting improves coverage but slightly harms the quality of the resulting definitions. The likely reason is that aggressive splitting creates sub-concepts that are terminologically less stable, less canonical, or less well supported by training evidence. Longer concepts preserve more of the biomedical conceptual unit and therefore provide a better basis for intensional definition generation.

7.4. Task B Local Results

On local development, the stronger QTT-guided generator improved BLEU from 12.63 (baseline) to 17.37, a relative gain of approximately 37%, while BERTScore F1 remained close (0.8799 baseline vs. 0.8742 QTT). This is important because it shows that collapse-guided evidence selection makes generated definitions more lexically and structurally similar to gold intensional definitions without sacrificing semantic adequacy. In our interpretation, collapse works here not as a direct semantic score but as a coherence prior that filters out noisier supporting contexts and privileges conceptually stable evidence.

Table 2
Official Task A results across runs
Parkinson's disease:

Run	Micro-P	Micro-R	Micro-F1	Type-P	Type-R	Type-F1
Run 1	0.8305	0.4261	0.5600	0.8352	0.5300	0.6500
Run 2	0.8293	0.4278	0.5645	0.8300	0.5322	0.6497
Run 3	0.8266	0.4348	0.5699	0.8311	0.5414	0.6600
Run 4	0.7224	0.5900	0.6522	0.8292	0.5450	0.6600

Mental health:

Run	Micro-P	Micro-R	Micro-F1	Type-P	Type-R	Type-F1
Run 1	0.8236	0.4212	0.5573	0.8284	0.5496	0.6608
Run 2	0.8233	0.4218	0.5578	0.8282	0.5504	0.6613
Run 3	0.8180	0.4306	0.5642	0.8225	0.5622	0.6679

Run 4	0.7128	0.6079	0.6561	0.8212	0.5662	0.6703

Table 3
Official Task B results (Run 1 = long concepts; Run 2 = split concepts)
Parkinson's disease:

Run	n_pairs	Concept Rec.	BLEU-1	BLEU-2	BLEU-3	ROUGE-1	ROUGE-2	ROUGE-L	BERT F1
Run 1	76	0.466	21.82	11.95	7.74	0.289	0.091	0.232	0.793
Run 2	78	0.479	20.29	11.31	7.49	0.269	0.087	0.223	0.790

Mental health:

Run	n_pairs	Concept Rec.	BLEU-1	BLEU-2	BLEU-3	ROUGE-1	ROUGE-2	ROUGE-L	BERT F1
Run 1	100	0.469	18.68	8.57	5.05	0.252	0.061	0.201	0.777
Run 2	104	0.488	17.68	8.31	4.94	0.244	0.055	0.194	0.776

8. Discussion and conclusions

We presented a unified QTT-informed system for DETECH 2026 covering both automatic term extraction and definition generation. The system combines biomedical neural extraction, contextual state induction, collapse-based coherence estimation, training-informed concept assignment, and LLM definition generation. Official Task A results show that our strongest runs are the recall-oriented PubMedBERT extraction systems, while local development confirms that collapse can improve precision and stabilize borderline decisions. Official Task B results show that longer concepts consistently yield better lexical, structural, and semantic definitions than split concepts, even though splitting slightly improves concept recall.

Across both tasks, the results suggest that collapse is most useful as a coherence-sensitive downstream mechanism rather than as a replacement for a strong base model. In Task A, it functions as a precision-oriented signal that helps reshape the boundary in ambiguous cases, but official performance is still driven primarily by recall inherited from the extractor. In Task B, collapse contributes more clearly through evidence selection for the LLM, especially when rich training-side concept evidence exists.

The comparison between long and split concepts also yields a broader methodological insight. Concept normalization is not only a recall problem; it is also a representational problem. More aggressive decomposition can increase the chance of matching some gold concepts, but it may simultaneously produce units that are less definable in stable biomedical terms. This is fully compatible with the QTT perspective as conceptual stabilization depends on coherent semantic states, and over-fragmentation can weaken that coherence.

This paper has several limitations. First, the Task A collapse module was primarily validated through local development analysis, whereas the official hidden-set results favored the higher-recall extraction regime. Second, the Task B definition generator applies collapse only when a predicted concept is supported by training-side concept evidence; unseen concepts revert to the baseline generator. Third, the concept assignment rules remain heuristic and could benefit from explicit contextual similarity or ontology-aware constraints.

Future work will therefore focus on three directions: stronger nested-term handling in Task A, context-aware concept assignment in Task B, and a fuller theoretical formalization of collapse as a terminology-oriented contextual coherence signal. Future revisions of this approach will also expand the theoretical discussion of QTT, add fuller ablation studies, and analyze collapse coverage and failure cases in greater depth.

9. Declaration on Generative AI

During the preparation of this work, generative AI tools were used to support coding and draft polishing. All theoretical and methodological choices as well as experimental interpretations, and final content are by the authors, who take full responsibility for the paper.

References

[1] Di Nunzio, G. M., Vezzani, F., Bonato, V., & Silvello, G. (2026). DETECH 2026: Overview of the Definition and Term Extraction Challenge. In *Proceedings of the First Definition and Term Extraction Challenge (DETECH 2026)*. CEUR.org.

[2] Burgos, D., Quiroz, G., & Pérez-Pérez, C. M. (2024). Antecedentes y principios para una teoría cuántica del término. In G. Quiroz, D. Burgos, & F. Zuluaga (Eds.), *Terminología del español: el término / Spanish Terminology: The Term*, (pp. 9-33). Routledge.

[3] Burgos, D. (2024). A Quantum Theory of Terms and New Challenges to Meaning Representation of Quanterms. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations* @ LREC-COLING 2024 (pp. 48-53).

[4] Cabré, M. T. (1999). *Terminology: Theory, methods and applications*. John Benjamins.

[5] International Organization for Standardization. (2019). *ISO 1087:2019: Terminology work and terminology science-Vocabulary*. ISO.

[6] Xu, K., Feng, Y., Li, Q., Dong, Z., & Wei, J. (2025). Survey on terminology extraction from texts. *Journal of Big Data*, 12, Article 29. doi:10.1186/s40537-025-01077-x

[7] Tran, H. T. H., Martinc, M., Caporusso, J., Delaunay, J., Doucet, A., & Pollak, S. (2026). Recent advances in automatic term extraction: A comprehensive survey. *ACM Computing Surveys*, 58(9), 1–35. doi:10.1145/3787584

[8] Banerjee, S., Chakravarthi, B. R., & McCrae, J. P. (2024). Large language models for few-shot automatic term extraction. In *Natural language processing and information systems* (pp. 137–150). Springer. doi:10.1007/978-3-031-70239-6_10

[9] Liu, S., Lefever, E., & Hoste, V. (2025). MariATE: Automatic term extraction using large language models in the maritime domain. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing* (pp. 663–673). INCOMA Ltd. doi:10.26615/978-954-452-098-4-077

[10] Rigouts Terryn, A., Hoste, V., & Lefever, E. (2020). In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, 54(2), 385–418. doi:10.1007/s10579-019-09453-9

[11] Spala, S., Miller, N., Dernoncourt, F., & Dockhorn, C. (2020). SemEval-2020 Task 6: Definition extraction from free text with the DEFT corpus. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 336–345). International Committee for Computational Linguistics. doi:10.18653/v1/2020.semeval-1.41

[12] Veyseh, A. P. B., Dernoncourt, F., Dou, D., & Nguyen, T. H. (2020). A joint model for definition extraction with syntactic connection and semantic consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5), 9098–9105. doi:10.1609/aaai.v34i05.6444

[13] Kučera, F., Mandl, C., Echizen, I., Timofte, R., & Spinde, T. (2026). SciDef: Automating definition extraction from academic literature with large language models [Preprint]. *arXiv*. doi:10.48550/arXiv.2602.05413

[14] Tamayo Herrera, A. J. (2024). Hacia la estandarización de la evaluación en la extracción de términos. In G. Quiroz, D. A. Burgos, & J. F. Zuluaga Molina (Eds.), *Terminología del español: el término / Spanish terminology: The term* (pp. 223–242). Routledge. doi:10.4324/9781003344339-17
[15] Van Rijsbergen, C. J. (2004). *The geometry of information retrieval.* Cambridge University Press.
[16] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* (HEALTH), 3(1), 1-23.
[17] Burgos, D., & Vásquez, D. (2024). El nombre terminológico. In G. Quiroz, D. Burgos, & F. Zuluaga (Eds.), *Terminología del español: el término / Spanish terminology: The term* (pp. 47-71). Routledge.
[18] Burgos, D. (2024). El término metafórico. In G. Quiroz, D. Burgos, & F. Zuluaga (Eds.), *Terminología del español: el término / Spanish terminology: The term* (pp. 47-71). Routledge.