Diagnosable ColBERT: Debugging Late-Interaction Retrieval Models Using a Learned Latent Space as Reference

Remy François
Parallia Healthcare
francois.remy@parallia.eu

Abstract

Reliable biomedical and clinical retrieval requires more than strong ranking performance: it requires a practical way to find systematic model failures and curate the training evidence needed to correct them. Late-interaction models such as ColBERT (Khattab and Zaharia, 2020) provide a first solution thanks to the interpretable token-level interaction scores they expose between document and query tokens. Yet this interpretability is shallow: it explains a particular document–query pairwise score, but does not reveal whether the model has learned a clinical concept in a stable, reusable, and context-sensitive way across diverse expressions. As a result, these scores provide limited support for diagnosing misunderstandings, identifying irreasonably distant biomedical concepts, or deciding what additional data or feedback is needed to address this. In this short position paper, we propose Diagnosable ColBERT, a framework that aligns ColBERT token embeddings to a reference latent space grounded in clinical knowledge and expert-provided conceptual similarity constraints. This alignment turns document encodings into inspectable evidence of what the model appears to understand, enabling more direct error diagnosis and more principled data curation without relying on large batteries of diagnostic queries.

1 Introduction

Trustworthy artificial intelligence in high-stakes domains increasingly requires transparent reporting of model capabilities, limitations, and failure modes (World Health Organization, 2021; European Parliament and Council of the European Union, 2017; Oberst et al., 2024). Finding systematic model failures is a central requirement for deploying retrieval systems in clinical and biomedical settings. When a retrieval model confuses nearby concepts, overgeneralizes across contexts, or fails to encode clinically salient negation, the consequence is not merely a lower ranking metric, but a weaker basis for downstream decision making and a poorer signal for curating corrective training data (Harkema et al., 2009; Lee et al., 2010; Wang et al., 2020). For this reason, debugging should ideally reveal not only that a model failed on a particular document–query pair, but also *what* the model appears to understand, *where* that understanding breaks down, and *which* additional examples are needed to improve it.

Late-interaction models in the style of ColBERT are appealing in this regard because they expose fine-grained interaction scores between query and document tokens (Khattab and Zaharia, 2020). However, this evidence remains retrospective and local: it helps explain why a given query matched a given document, but not whether the model has learned clinically meaningful distinctions robustly across paraphrases, contexts, and compositional variants (Huang and Baldwin, 2023; Kang et al., 2025). Explaining a match is not the same as gauging and diagnosing misunderstandings.

In clinical retrieval, the most consequential failures often depend on context-sensitive factors such as negation, temporality, uncertainty, experiencer, historical status, and the distinction between mentioning a test and asserting a condition. These are precisely the kinds of distinctions that may be weakly expressed in local match scores while remaining decisive for retrieval quality.

We therefore argue that checking model understanding requires interpretation at multiple levels of contextualization. Many biomedical meanings are only partially specified at the token level and become clearer as progressively more context is introduced (Lee et al., 2010). For example, a model that appears reasonable on *cat* may still fail on *cat scratch disease*. Moreover, paragraph-level context can further influence whether a mention is current, historical, speculative, negated, or even concerns another person (Harkema et al., 2009; Wang et al., 2020). Interpretability must therefore extend

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 713–717

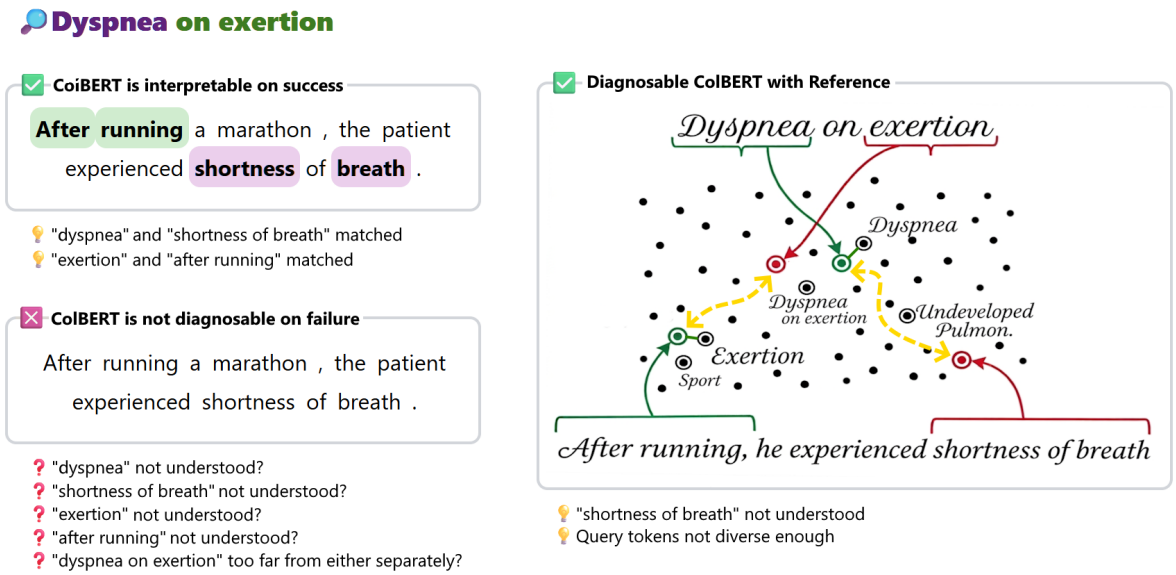

Figure 1: Standard ColBERT-style late interaction produces interpretable query–document scores, but provides little guidance when no meaningful match is found for a token, as illustrated on the left. By contrast, augmenting the retriever with a reference latent space, as on the right, allows testers of Diagnosable ColBERT models to diagnose failures more quickly and more actionably by separating latent-space semantics from in-context understanding.

beyond isolated token matches toward a graded view of understanding across term, sentence, and paragraph-level alignment. Accordingly, we treat diagnostic alignment as multi-factorial: it should reveal term-level concept identity, phrase-level composition, context-level qualifiers, and broader discourse elements that alter clinical interpretation.

This broader view of interpretation also motivates the use of a structured latent space rather than a fixed inventory of labels. If the embedding space is shaped to respect concept-level and sentence-level similarity constraints, then it can represent not only previously named clinical concepts, but also contextually enriched and partially novel combinations of them (Campbell et al., 2014). Prior work also suggests that such spaces are feasible to construct in practice, whether by grounding biomedical representations in ontological definitions and relations as in BioLORD (Remy et al., 2024), by aligning dense representations to structured label spaces (Decorte et al., 2025), or by grounding annotated mentions in a dynamic model-generated concept space when the concept inventory is large and evolving (Stepanov et al., 2026).

In this publication, we therefore propose Diagnosable ColBERT, a framework that leverages such a reference latent space learned from expert-specified similarity constraints over clinical concepts, and aligns late-interaction token embeddings to it via pre-projection adapters. The central proposal of this paper is therefore not a new retrieval scoring rule, but a new diagnostic lens: late-interaction representations should be interpreted against a clinically grounded reference space rather than only through pairwise match scores. The goal is not merely to explain a retrieval score after the fact, but to make document encodings themselves interpretable as evidence about what the model appears to know. This creates a path toward diagnosing misunderstandings directly from encoded documents, identifying confusable concepts, and curating training data in a more principled way. More generally, we advocate a shift from retrospective score inspection to concept-grounded diagnosis of model understanding. Such representations could support both direct human inspection and more automated auditing workflows.

2 Diagnostic Framework

Diagnosable ColBERT is organized around a pre-existing reference latent space that serves as a domain-specific diagnostic scaffold. This latent space needs to accommodate both concept names, clinical sentences, and paragraphs; see for example BioLORD (Remy et al., 2024). The purpose of this space is to make contextual token representations clinically legible: not only in terms of term-level concept identity, but also in terms of local composition and context-level qualifiers such as negation, temporality, uncertainty, or experiencer. Instead of only asking which query token matched which document token, we ask what clinically relevant factors a contextualized representation appears to encode and how they arrange geometrically.

In this view, alignment means mapping late-interaction token representations into a space where these factors can be inspected more directly. The retrieval representation should remain tied to this diagnosed representation, but need not be identical to it. A natural design is for retrieval embeddings to be learned as a lower-dimensional downprojection of the diagnosed representation, so that retrieval can reweight clinically relevant factors for ranking efficiency without discarding the richer structure needed for diagnosis. This keeps the framework centered on a simple idea: retrieval and diagnosis should stay coupled, but diagnosis should not be reduced to whatever geometry happens to be most convenient for fast search.

The specific architecture used to realize this idea is secondary to the paper's main claim. One can implement the diagnostic view with lightweight contextual adapters and task-specific projection heads; our essential contribution is the diagnostic framing itself: a clinically grounded reference geometry for inspecting what a late-interaction retriever appears to perceive from a context.

3 Practical Examples

A central limitation of standard ColBERT interpretability is that it is fundamentally relational: it explains why a query matched a document, but offers limited leverage when the goal is to determine why a relevant match failed to occur (see Figure 1). Diagnosable ColBERT is intended to address precisely this gap. Rather than treating retrieval failure as a single undifferentiated event, it seeks to identify which part of the representation pipeline failed to preserve the clinically relevant meaning.

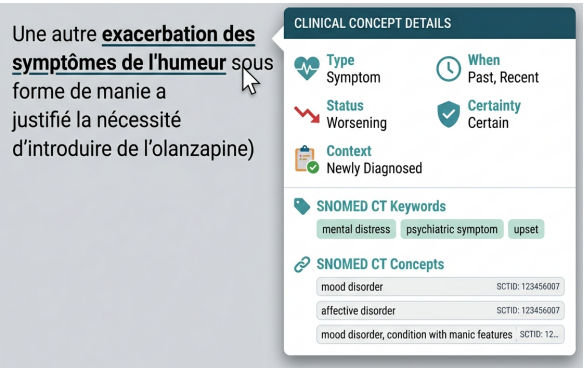

Figure 2: Proposed debugging interface for Diagnosable ColBERT. The interface exposes token-level query and document representations together with their placement in the reference latent space and the clinically meaningful concepts or qualifiers associated with nearby regions. This gives testers a practical way to distinguish weak interaction scores from deeper failures of concept grounding, abbreviation handling, or contextual interpretation, and to turn a retrieval miss into an actionable diagnosis.

Consider a case report retrieval system in which a tester issues the query *bartonellosis*. A relevant report is missed because the report mentions only *CSD*, the abbreviation for *cat scratch disease*. Standard query–document inspection establishes that the interaction between *bartonellosis* and *CSD* is too weak. But this finding alone is diagnostically incomplete. It does not tell us whether the query representation failed to capture the target disease family, whether the document representation failed to interpret the abbreviation correctly, or whether both encodings are deficient.

Diagnosable ColBERT resolves this ambiguity by grounding both sides in a reference latent space. The tester can inspect whether *bartonellosis* is already positioned near the relevant disease concept and, separately, whether *CSD* is mapped into that same region. If *bartonellosis* is correctly grounded but *CSD* is not, then the root cause is not weak interaction per se, but document-side abbreviation understanding. If both are misplaced, the problem lies deeper, at the level of concept representation. This kind of distinction is difficult to obtain from pairwise interaction scores alone, yet it directly guides remediation.

When the problem lies in abbreviation grounding, the natural intervention is to curate examples that explicitly connect abbreviated, expanded, and synonymous forms. When the problem lies in broader concept placement, the remedy instead concerns the shaping of the underlying semantic space.

The same diagnostic logic extends beyond query-conditioned retrieval. Some failures can be identified directly from the document encoding itself. Consider the utterance: *Are you allergic to any of the following drugs? — Yes, to ranitidine*. Here the relevant question is whether the token representation of *ranitidine* reflects not only the medication identity, but also the allergic relation introduced by context. If the encoding remains close to the drug concept while failing to capture the neighboring context of allergy, adverse reaction, or intolerance, then the model has encoded the entity but not the clinically decisive context.

This is a query-free diagnostic signal: the tester can identify the failure before retrieval is even attempted. Again, the diagnosis suggests a targeted intervention, namely the addition of training examples in which drug mentions appear indirectly in contexts related to allergy or adverse-reactions, and ensure that this context is getting properly picked up by the encoder model.

4 Real-World Usage

In our experience, debugging user interfaces for Diagnosable ColBERT models[1] have already proven useful as a practical way to inspect token-level phenomena and surface failures that would have been difficult to localize from ranking metrics alone. What made the interface valuable was that it helped spot remaining representational weaknesses that actually mattered in practice, including gaps in conceptual grounding, abbreviation and brand name handling, and preservation of clinical context.

Subjectively, this changed the development process. Instead of treating misretrievals as abstract evaluation outcomes, we were often able to trace issues back to recurring representational problems and then respond with more targeted curation and model changes. These observations informed the progression from ClinicalEncoder25 to ClinicalEncoder26AM and beyond, even if this workflow remains work in progress and should not be mistaken for a complete evaluation methodology.

5 Limitations and Scope

Diagnosable ColBERT is best understood as a diagnostic scaffold, not as a complete semantics of biomedical meaning. Its purpose is to make clinically relevant aspects of the model's internal organization more legible to a tester, not to claim that every retrieval-relevant distinction can be exhaustively captured in a single reference geometry. Any such space reflects design choices about which concepts, relations, and contextual qualifiers are made salient, and its usefulness depends on being well shaped by domain knowledge, expert constraints, and clinically meaningful similarity structure. The proposal therefore improves diagnosability without by itself guaranteeing retrieval quality: diagnostic alignment complements retrieval evaluation by making failure modes more interpretable and actionable, but useful systems will still need strong ranking behavior as well.

6 Conclusion

In this paper, we have argued that debugging late-interaction retrievers requires more than inspecting token-level query–document scores after a failure has already occurred. Our central claim is that retrieval models become substantially more diagnosable when their token representations are aligned to a clinically grounded reference latent space that makes concept grounding, contextual qualifiers, and representational gaps directly inspectable.

Diagnosable ColBERT is a concrete step in this direction. Rather than treating retrieval as a black box whose failures must be inferred indirectly from weak matches and downstream metrics, it provides a practical framework for localizing why a miss occurred and for inspecting difficult document segments even before a query is issued.

Much remains to be refined, especially in constructing strong reference spaces and turning diagnoses into robust improvement workflows. Even so, we view the direction as already promising in practice: clinically grounded diagnostic representations complement ranking evaluation by making retrieval behavior more legible and actionable.

[1] Public demos: http://demo26.parallia.eu/ and http://text2json.parallia.eu.

References

Walter S. Campbell, James R. Campbell, William W. West, James C. McClay, and Steven H. Hinrichs. 2014. Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings. *Journal of the American Medical Informatics Association*, 21(5):885–892.

Jens-Joris Decorte, Jeroen Van Hautte, Chris Develder, and Thomas Demeester. 2025. Efficient text encoders for labor market analysis. *IEEE Access*, 13:133596–133608.

European Parliament and Council of the European Union. 2017. Regulation (EU) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) no 178/2002 and regulation (EC) no 1223/2009 and repealing council directives 90/385/EEC and 93/42/EEC. Official Journal of the European Union, L 117, pp. 1–175. Adopted April 5, 2017; published May 5, 2017.

Hendrik Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics*, 42(5):839–851.

Yiran Huang and Timothy Baldwin. 2023. Robustness tests for automatic machine translation metrics with adversarial attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4649–4675.

Junmo Kang, Yunhyeok Ro, Junsie Heo, and Minjoon Seo. 2025. TRIAL: Token relations and importance aware late-interaction for accurate text retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15404–15427.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Dennis H. Lee, Francis Y. Lau, and Hue Quan. 2010. A method for encoding clinical datasets with SNOMED CT. *BMC Medical Informatics and Decision Making*, 10:53.

Michael Oberst, Davis Liang, and Zachary C. Lipton. 2024. Pioneering the science of AI evaluation. Whitepaper, published September 19, 2024; updated August 7, 2025.

François Remy, Kris Demuynck, and Thomas Demeester. 2024. BioLORD-2023: Semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855.

Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, and Oleksandr Lukashov. 2026. The million-label NER: Breaking scale barriers with GLiNER bi-encoder. *Preprint*, arXiv:2602.18487. Introduces GLiNKER, a modular framework for large-scale entity linking.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. MedSTS: A resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72.

World Health Organization. 2021. Ethics and governance of artificial intelligence for health: WHO guidance. Published June 28, 2021.