Journal of the American Medical Informatics Association, 2024, 1–12
https://doi.org/10.1093/jamia/ocae029
Research and Applications

INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD

Research and Applications

BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights

François Remy, Msc[1]*, Kris Demuynck, PhD[1], Thomas Demeester, PhD[1]

[1]Internet and Data Science Lab, imec, Ghent University, Ghent, Belgium

*Corresponding author: François Remy, Msc, Internet and Data Science Lab, imec, Ghent University, Ghent, Belgium (francois.remy@ugent.be)

Abstract

Objective: In this study, we investigate the potential of large language models (LLMs) to complement biomedical knowledge graphs in the training of semantic models for the biomedical and clinical domains.

Materials and Methods: Drawing on the wealth of the Unified Medical Language System knowledge graph and harnessing cutting-edge LLMs, we propose a new state-of-the-art approach for obtaining high-fidelity representations of biomedical concepts and sentences, consisting of 3 steps: an improved contrastive learning phase, a novel self-distillation phase, and a weight averaging phase.

Results: Through rigorous evaluations of diverse downstream tasks, we demonstrate consistent and substantial improvements over the previous state of the art for semantic textual similarity (STS), biomedical concept representation (BCR), and clinically named entity linking, across 15+ datasets. Besides our new state-of-the-art biomedical model for English, we also distill and release a multilingual model compatible with 50+ languages and finetuned on 7 European languages.

Discussion: Many clinical pipelines can benefit from our latest models. Our new multilingual model enables a range of languages to benefit from our advancements in biomedical semantic representation learning, opening a new avenue for bioinformatics researchers around the world. As a result, we hope to see BioLORD-2023 becoming a precious tool for future biomedical applications.

Conclusion: In this article, we introduced BioLORD-2023, a state-of-the-art model for STS and BCR designed for the clinical domain.

Key words: natural language processing; machine learning; knowledge bases; biological ontologies; semantics.

Introduction

Clinical and biomedical natural language processing (NLP) rose to prominence in the last few years,[1] debuting a long series of surveys[2–8] highlighting the potential and inherent challenges of harnessing the synergies between, firstly, the reliable insights originating from biomedical knowledge graphs (BKGs) and, secondly, the impressive generalization capabilities of cutting-edge deep learning techniques, specifically large language models (LLMs).

Such an insight fusion could have an immense impact across a wide range of applications, encompassing clinical case summarization, clinical decision support, patient diagnosis and triage, pharmacovigilance, disease subtyping, drug discovery, as well as help researchers build more explainable artificial intelligence (AI) systems.

In this work, we extend the line of work initiated by BioSyn,[9] SapBERT,[10] and BioLORD[11] by integrating knowledge graph information during and after the pretraining of semantic bidirectional language models.[12] These bidirectional models form a cornerstone of the modern retrieval-augmented generation (RAG) pipeline,[13,14] which is critical in most applications of LLMs in a real-world setting (both to enable LLMs to access accurate and up-to-date data before writing their answers, but also because it makes tracing and combating erroneous answers more tractable).

This article introduces several novel contributions to this existing body of work, aiming: firstly, at broadening the biomedical expertise of semantic models, secondly, at reducing the trade-off between the biomedical knowledge and the general language understanding of finetuned models, and thirdly at enabling more languages to benefit from the obtained improvements.

To this end, we present in this article a new model, BioLORD-2023, which builds upon the achievements of the original BioLORD model[11] but has novel characteristics, such as an improved training strategy and an updated training corpus. The original BioLORD will henceforth be referred to as BioLORD-2022, and compared against the new BioLORD-2023, to avoid confusion.

The first contribution of BioLORD-2023 concerns the usage of OpenAI LLMs[15] for converting into text the information contained in biomedical ontologies and their knowledge graphs,[16,17] a task where LLM's language fluency and latent knowledge of biomedical matters proved useful. This is important, as only 5% of clinical concepts possess human-written definitions in large biomedical meta-thesauri.[11,16] Fortunately, modern LLMs prompted with knowledge graph information have been shown to generate largely reliable, insightful, and fluent definitions for a vast majority of biomedical concepts.[18]

Received: November 28, 2023. **Revised:** January 16, 2024. **Editorial Decision:** January 31, 2024. **Accepted:** February 2, 2024

Yet, the practical benefits of such LLM-generated definitions have not been studied extensively so far, something we aim to address. In this article, we conclusively demonstrate that the existence of these artificial definitions for a large majority of biomedical concepts is able to substantially enhance the quality of textual representations obtained using the "Learning of Ontological Representations through Definitions and textual representations" strategy (LORD),[11] by including such definitions in the training data of our new model and assessing its impact on downstream tasks.

Our second contribution, a self-distillation approach, takes advantage of the existence of this broad set of definitions to accelerate the convergence process of the LORD training strategy, thereby achieving superior biomedical knowledge acquisition at a reduced loss of general language understanding capabilities.

Combining these 2 strategies, we train a new biomedical semantic model, BioLORD-2023. We evaluate our newly trained BioLORD-2023 model on a broad spectrum of downstream tasks, including biomedical concept representation (BCR), semantic textual similarity (STS), and named entity linking (NEL), with considerable gains across the entire range of tasks.

Our third and last contribution is the release of our first multilingual clinical language model, enabling the retrieval and the concept normalization of content in up to 50 languages, thanks to the cross-lingual distillation strategy described by Reimers et al.[19] and a multilingual alignment dictionary built from SNOMED—CT[17] using LaBSE.[20]

We evaluate this new multilingual model based on the test suite developed for multilingual-SapBERT,[21] a similar model which is widely considered as the current state of the art in the domain. We also evaluate the quality of the distillation process using our evaluation metrics for the English language.

To summarize, our 3 main contributions are as follows:

1) The expansion of our training corpus by supplementing its existing knowledge with new LLMs-generated definitions for 400 000 concepts, fusing knowledge graph and LLM insights inside the LORD pretraining.
2) The introduction of a novel self-distillation technique to speed up biomedical knowledge acquisition while preserving the language understanding capabilities of the BioLORD-type models.
3) The delivery of a state-of-the-art multilingual model for the biomedical domain, using a proven cross-lingual distillation technique.

Related works

Before delving into the details of our methodology, we provide a short description of the context in which this work finds its place, the technologies used in this article, and the previous efforts which made this work possible, starting with BKGs.

Biomedical knowledge graphs

BKGs are graph-based representations of biomedical data and knowledge, where nodes represent entities (eg, genes, diseases, drugs) and edges represent relationships (eg, interactions, associations, causations) between these entities. BKGs can be classified by their scope, curation strategy, and structure, and they can contain various types of information.[22] Several applications and tasks in biomedicine and healthcare have been shown to benefit from BKGs, both because BKGs can be used as reliable sources of information, and because they provide traceable explanations for answers which can be derived from them.

Unified Medical Language System (UMLS)[16] and Systematized Nomenclature of Medicine—Clinical Terms (SNOMED—CT)[17] are 2 examples of BKGs that are used to standardize health and clinical information. They differ from each other both in scope and structure. UMLS, as a conglomerate of biomedical vocabularies, aims for large coverage and encompasses over 3.7 million concepts from 200+ source vocabularies, including SNOMED-CT; however, it does not provide consistent views for all its concepts. Conversely, SNOMED-CT aims to provide a reliable gold standard for electronic health record (EHR) standardization, thanks to its meticulous formal logic-based structure and approximately 358 000 clinical concepts.

In this work, we maximally leverage the strengths of both UMLS and SNOMED-CT by using them in the learning phases for which they are best suited. For instance, in continuity with previous iterations of BioLORD, we employ the UMLS concepts and relationships as part of the contrastive pretraining, where the increased scope and diversity of described relationships is an advantage. However, we also employ definitions from the Automatic Glossary of Clinical Terminology (AGCT),[18] which leverages the more standardized and consistently annotated graph of SNOMED-CT to enhance the homogeneity and reliability of the produced definitions.

Large language models in healthcare

LLMs encompass various types of machine learning models which have been trained on vast amounts of text to either achieve some understanding of existing content or generate original content based on instructions. They have recently demonstrated remarkable capabilities in NLP tasks and beyond.

LLMs have been used in various clinical applications.[23–25] One such application is medical transcription and clinical coding using the International Classification of Diseases, where LLMs have been used to improve the accuracy and efficiency of converting spoken medical observations into written or structured EHRs.[26]

LLMs also show promise in various clinical data analysis tasks. For instance, they can analyze patient data such as medical records,[23] or interpret imaging studies and laboratory results.[27,28] These insights can support diagnosis by doctors and other healthcare professionals.[28] Finally, LLMs can also be used to identify clinical trial opportunities for patients by analyzing patient data such as medical records.[29]

Examples of large biomedical language models include Med-PALM,[24] Galactica,[30] ClinicalGPT,[31] BioMedLM,[32] BioGPT,[33] and others. Commercial language models such as ChatGPT[15] and GPT-4[34] have also shown great capabilities in Biomedical AI,[25] despite lacking dedicated finetuning procedures.

Retrieval-augmented generation

While LLMs have demonstrated remarkable capabilities in NLP tasks, they are prone to hallucinations, which can result in incorrect diagnoses and treatments, leading to adverse effects on patients. RAG is a method that can be used to reduce hallucinations in LLMs, as well as increase trust in AI-based tools by explicitly linking their output to external knowledge.

RAG involves augmenting LLMs with information retrieval (IR) systems, which can provide relevant content retrieved from external corpora as references. By incorporating external knowledge, retrieval-augmented LLMs can answer in-domain questions that cannot be answered by solely relying on the world knowledge stored in the model's own parameters.

IR systems often make use of pretrained models for dense passage retrieval (DPR) and baseline systems such as BM25, a traditional IR strategy that uses term frequency-inverse document frequency weighting to rank documents based on their relevance to a query. In this work, we focus on creating a dense concept and sentence representation model which is well suited for use, among other tasks, as a DPR model within a biomedical RAG pipeline.

Biomedical representation learning

To build a system capable of retrieving content relating to a concept by one of its names, the underlying models must be familiar with the vast biomedical terminology and the meaning of the underlying concepts. Because of the daunting scale of clinical terminology, in-domain pretraining is insufficient to cover long multiword expressions accurately. Therefore, a now large body of work attempts to produce better representations using BKGs as a source, due to their extensive coverage of biomedical concepts.

Although multiple variations of the strategy exist, most state-of-the-art models prior to 2022 focused on learning representations of biomedical entities based on the synonyms of entities, in a contrastive manner. Since then, BioLORD-2022 was introduced to avoid spurious token-based overfitting using more of the relationships found in the BKGs, and by using full definitions to extract more fine-grained information out of medical knowledge bases. However, BioLORD-2022 did not make use of LLMs to expand the training data, and it only used concept definitions during the pretraining phase. BioLORD-2022 models also suffer from a considerable performance decrease in general language understanding compared to the original pretrained language models they were based on, which is undesirable. In this work, we extend the BioLORD-2022 training strategy to fix these deficiencies.

Cross-lingual distillation of knowledge

In addition to this, we provide a multilingual variant of our new model. While some multilingual biomedical models already exist, such as Multilingual SapBERT, their performance has been lower than the monolingual English models. This is because they rely on the existence of sufficient non-English training data during pretraining, but the existence of such data for the biomedical domain is scarce, and results in subpar generalization.

One of the goals of this work is to improve upon this state of the art using proven techniques for cross-lingual distillation.[19] Cross-lingual distillation trains a multilingual model to produce the same output as a target English model irrespective of the language in which a concept name is provided, such that "*Fever*," "*Fiebre*," and "*Fieber*" produce the same output.

Methods

In this work, we train a concept and sentence representation model fine-tuned for biomedical content, following a novel 3-step strategy (see Figure 1), and evaluate its potential on several downstream tasks, all described in the following subsections. We also provide justifications for the changes made to the previous iterations of this strategy, with the use of ablation studies.

Training data

In addition to the algorithmic changes described in the next section, a core aspect of our work hinges on the new data we leverage during the training, which we describe further in this section.

In addition to the training data already used in previous works, this study makes use of a LLM to generate a large set of definitions of biomedical concepts (grounded using the information contained in the SNOMED-CT ontology as context, as well as the information stored in the weights of the LLM itself). We leave the precise description of the procedure and its expert evaluation to the paper introducing the dataset [18] but we provide the most relevant details in the following paragraph.

In this new study, the preexisting UMLS definitions are indeed complemented by 400 000 biomedical definitions from the AGCT, a large-scale biomedical dictionary of clinical concepts which we generated using the SNOMED-CT ontology and the GPT-3.5 language model. A subset of the generated definitions was evaluated by NLP researchers with biomedical expertise on 3 metrics: factuality, insight, and fluency; based on these metrics and a strict 6-grade quality rating, it was determined that more than 80% of the generated definitions would be usable for patient education, while more than 96% appeared useful for machine learning tasks. In this work, we set out to confirm whether that is truly the case in a practical scenario.

BioLORD-2023 also makes use of a newer version of the UMLS ontology (v2023AA) to generate its textual description, compared to BioLORD-2022 (which used v2020AB). This enables the new version of the model to become more aware of recent developments in the field, for example, including knowledge related to the COVID-19 pandemic.

Training strategy

Contrastive phase

To obtain our BioLORD-2023 model, we first make use of the contrastive objective devised by van den Oord et al,[35] with the goal of instilling biomedical knowledge into a base language model. More precisely, we make use of the LORD strategy[11] in which batches of concept names and their definitions are fed to the language model, and where the distance between the representation of a concept name and the representation of its definition should be minimized while maximizing the distance between a concept and the definition of the other concepts in the batch.

For instance, the representations by the language model of "ranitidine" and of "an H2-antagonist substance frequently used to treat peptic ulcer"[16] are encouraged to be as close as possible, while remaining far from the representations of "aspirin" and of "a synthetic compound used medicinally to relieve mild or chronic pain and to reduce fever and

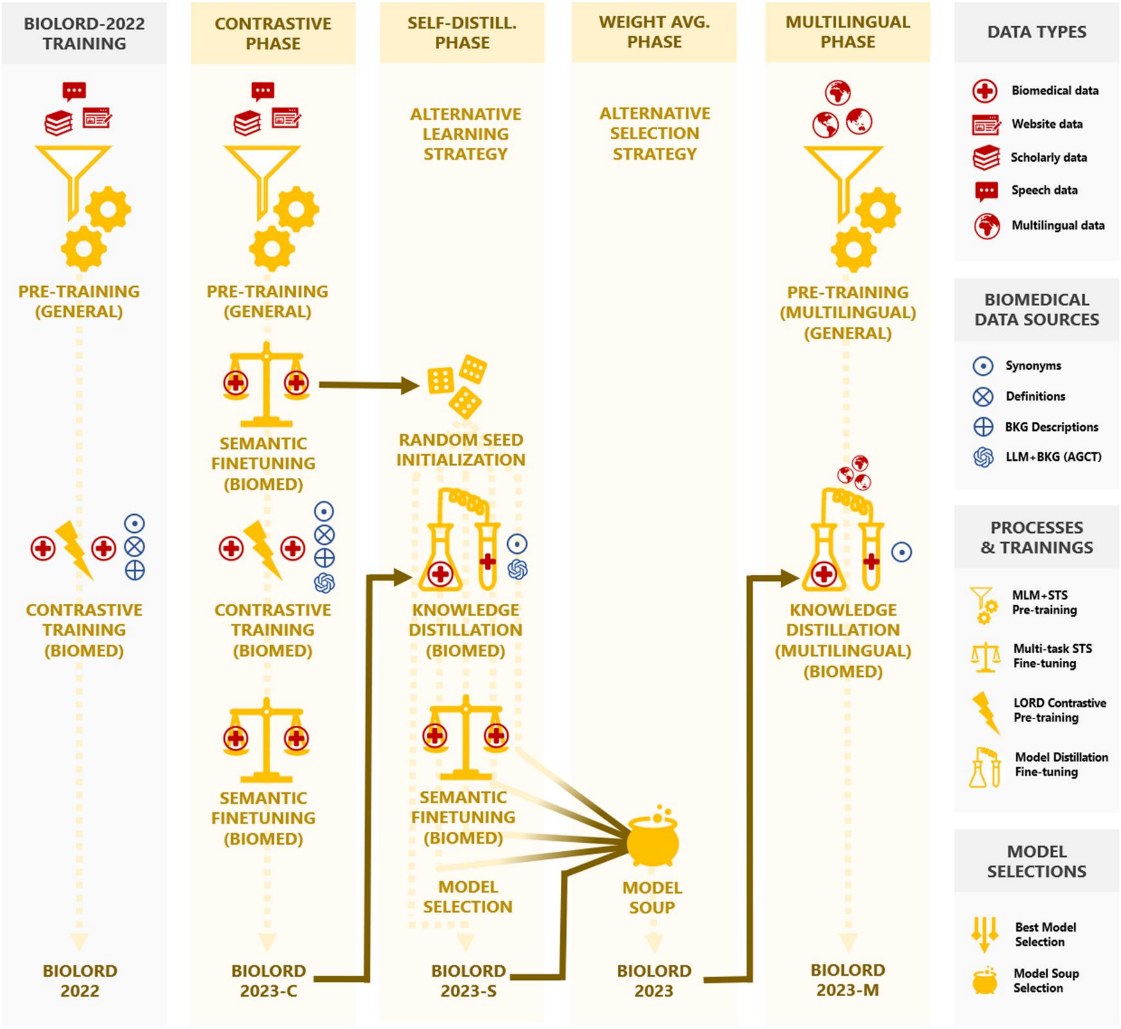

Figure 1. Compared to BioLORD-2022 (left), BioLORD-2023 involves a more advanced training strategy, composed of 3 phrases: a contrastive phase (further illustrated in Figure 2), a self-distillation phase (illustrated in Figure 3), and a weight-averaging phase (all further described in the following subsections).

inflammation",[16] which are related to another concept of the batch (as illustrated in Figure 2).

In this work, we make use of the improved initialization strategy developed by Remy et al,[36] which involves adapting the STAMB2 model[37] used for the initialization of BioLORD-2022 to the STS tasks of our benchmark in a multi-task setup, prior to applying our contrastive phase. Multi-task setups have been shown to be effective in scenarios where catastrophic forgetting is possible as a result of continual learning.[38] By aligning the STAMB2 model to the human preference of STS datasets prior to applying the contrastive learning phase of BioLORD, BioLORD-2023 produces representations of medical concepts which align more closely with human judgment.

We also insert this adaptation phase a second time, after the contrastive learning stage of BioLORD. This further enhances the BioLORD-2023 model's performance on STS tasks.

Self-distillation phase

The contrastive pretraining strategy described earlier was shown to yield excellent results in biomedical knowledge acquisition. We briefly summarize next the findings of BioLORD-2022[11] and refer to the paper for details.

Compared to their STAMB2 base,[37] models trained using the BioLORD-2022 methodology show a visible improvement in clinical sentence understanding and greatly improved BCR capabilities.

Unlike models trained using the SapBERT training strategy, BioLORD-2022 models remained proficient in the handling of sentence-level semantics. We attribute this to our choice of model initialization and the inclusion of concept definitions in the BioLORD-2022 training strategy, which succeeded in preventing a catastrophic forgetting of sentence parsing during the biomedical knowledge acquisition phase.

However, the addition of the definitions did not prove sufficient to avoid a measurable degradation of the performance of the model in general-purpose semantic similarity tasks. This degradation cannot be solely attributed to a loss of knowledge about general-domain concepts, as the performance degradation remained visible even when no general-domain knowledge was required for solving the semantic task.

We attribute a large part of this performance degradation to the semantic space distortion induced by the extensive

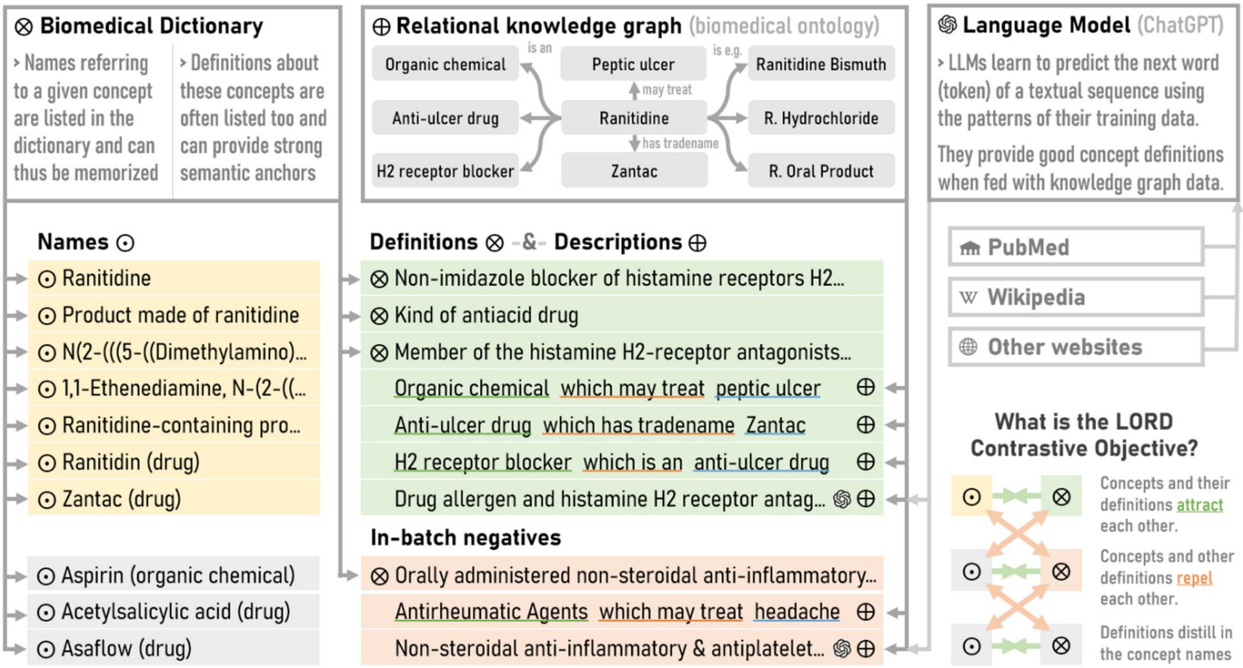
⊗ Biomedical Dictionary
› Names referring to a given concept are listed in the dictionary and can thus be memorized
› Definitions about these concepts are often listed too and can provide strong semantic anchors
⊕ Relational knowledge graph (biomedical ontology)
Organic chemical
Anti-ulcer drug
H2 receptor blocker
is an
Peptic ulcer
may treat
Ranitidine
has tradename
Zantac
is e.g.
Ranitidine Bismuth
R. Hydrochloride
R. Oral Product
Language Model (ChatGPT)
› LLMs learn to predict the next word (token) of a textual sequence using the patterns of their training data.
They provide good concept definitions when fed with knowledge graph data.
PubMed
Wikipedia
Other websites
Names ⊙
⊙ Ranitidine
⊙ Product made of ranitidine
⊙ N(2-(((5-((Dimethylamino)...
⊙ 1,1-Ethenediamine, N-(2-((...
⊙ Ranitidine-containing pro...
⊙ Ranitidin (drug)
⊙ Zantac (drug)
⊙ Aspirin (organic chemical)
⊙ Acetylsalicylic acid (drug)
⊙ Asaflow (drug)
Definitions ⊗ -&- Descriptions ⊕
⊗ Non-imidazole blocker of histamine receptors H2...
⊗ Kind of antiacid drug
⊗ Member of the histamine H2-receptor antagonists...
Organic chemical which may treat peptic ulcer ⊕
Anti-ulcer drug which has tradename Zantac ⊕
H2 receptor blocker which is an anti-ulcer drug ⊕
Drug allergen and histamine H2 receptor antag... ⊕
In-batch negatives
⊗ Orally administered non-steroidal anti-inflammatory...
Antirheumatic Agents which may treat headache ⊕
Non-steroidal anti-inflammatory & antiplatelet... ⊕
What is the LORD Contrastive Objective?
Concepts and their definitions attract each other.
Concepts and other definitions repel each other.
Definitions distill in the concept names

Figure 2. BioLORD aims to bring the representation of biomedical concept names (⊙) and their definitions (⊗) closer to each other, to ground the name representations with knowledge from the definitions. This is illustrated for the Ranitidine and Aspirin concepts from UMLS. Knowledge from the ontology's relational knowledge graph is injected by extending the set of known definitions with automatically generated definitions (⊕) from the Automatic Glossary of Clinical Terminology, as well as with simpler template-based descriptions sampled from UMLS relationships (⌘⊕). Contrastive learning is applied to attract the representations of compatible pairs (⊙, ⊗, or ⊕) and repel incompatible ones.

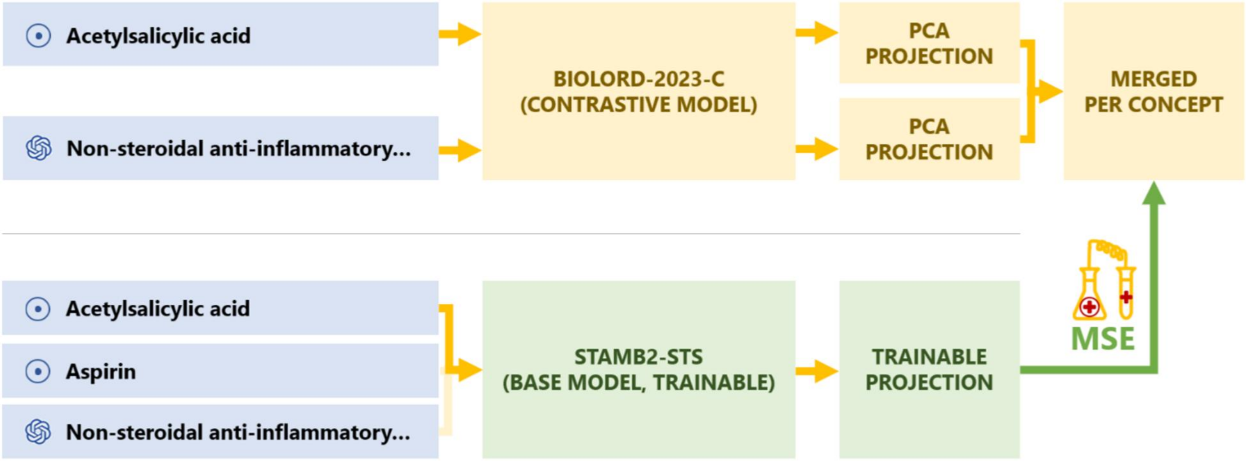
Acetylsalicylic acid
Non-steroidal anti-inflammatory...
BIOLORD-2023-C (CONTRASTIVE MODEL)
PCA PROJECTION
PCA PROJECTION
MERGED PER CONCEPT
Acetylsalicylic acid
Aspirin
Non-steroidal anti-inflammatory...
STAMB2-STS (BASE MODEL, TRAINABLE)
TRAINABLE PROJECTION
MSE

Figure 3. In the self-distillation phase, the knowledge acquired during the contrastive phase is imbued into the base model, using a more direct training strategy. The representation of each textual variant of a concept is trained to map the average of the contrastive model representation of its name and definition.

contrastive learning of concept names and their definitions, which only elicits a select few aspects of the STS task, blowing up the importance of these aspects considerably at the expense of other aspects of language understanding, thereby resulting in a loss of calibration of the model.

To counter-balance this, we propose to substitute the unsupervised contrastive learning phase by a supervised objective, taking into account the learnings of the contrastive phase without having to resort to a contrastive objective. To this end, we generate concept embeddings for the 4 million biomedical concepts contained in UMLS (using the BioLORD-2023-C model), and finetune a base model (which has not undergone the contrastive learning phase) to accurately predict these distilled concept embeddings via a learned projection followed by a standard mean-squared-error loss (see Figure 3).

We call this process the "self-distillation phase," as we distill the knowledge acquired by the contrastive model into a past version of itself in a supervised manner, hoping that this will better preserve its existing knowledge.

Like in the previous phase, we leverage the knowledge extracted from the knowledge graphs and the LLM by incorporating the concept definitions in the distillation process. We do this by first producing embeddings for concept names and their generated concept definitions using the model resulting from the contrastive phase of BioLORD-2023, and by subsequently averaging these 2 representations for each concept and using the result as the regression target during self-distillation, for both the concept name and its definition.

To improve the training speed, we reduce the latent space of produced embeddings to 64 dimensions through principal component analysis (PCA). Finally, we train a randomly initialized linear projection head on top of the base STAMB2-STS model to predict these 64-dimension embeddings (see Figure 3).

This supervised self-distillation phase possesses several key advantages: firstly, by including the concept definitions in the process to produce the concept embeddings, better representations can be learnt for biomedical concepts whose name is otherwise uninformative or difficult for the model to memorize; secondly, it is considerably faster than the contrastive learning phase, which is likely to cause less forgetting of the original task for an identical level of new knowledge acquisition[39]; thirdly, it further enables the language model to leverage its existing features to obtain the desired knowledge, without having to distort them to reduce the in-domain anisotropy,[40–42] and fourthly, it should be possible to apply this distillation phase on a different base model than the one used during the contrastive learning phase, taking advantage of improved base models at a limited training cost (we, however, leave the exploration of this aspect for future works).

Weight-averaging phase

An interesting aspect of the self-distillation phase described above is that it hinges on a randomly initialized projection head added on top of the base sentence representation model. As a result of this, different random seeds result in slightly different models, focusing on different aspects of the sentence embedding.

While a commonly used technique in this scenario, called hyperparameter tuning, aims to select the best model from multiple experiments based on a held-out validation set, Wortsman et al[43] discovered a better strategy, which they named "model soups," and which consists in the averaging of the weight of the parameters of several fine-tuned models.

Indeed, Wortsman et al. were able to show that it is often possible to improve the accuracy and robustness of the resulting system by averaging the weights of multiple models, each fine-tuned with different hyperparameter configurations. Unlike a traditional ensemble, no additional inference or memory costs are incurred as a result of the merge, irrespective of the number of models being merged, which makes this technique particularly attractive for DPR models, where a fast inference is highly desirable. We evaluate the impact of the weight averaging in Discussion section.

Cross-lingual distillation

Finally, we further make use in this work of the technique described by Reimers et al,[19] which consists of using a parallel corpus to distill a high-quality monolingual model into a multilingual model yielding similar representations for a text in that language and translations sourced from the corpus.

We cross-lingually distill the representations of our English model into the "paraphrase-multilingual-mpnet-base-v2" language model introduced in the same paper as the distillation technique, and which supports 50+ languages.

In particular, we make use of an aligned corpus generated from the regional releases of SNOMED-CT, using the Google LaBSE bi-text mining model,[20] and which we describe in more detail in our EmP 2022 publication.[44] This aligned corpus contains alignments for the following languages: English, Spanish, French, German, Dutch, Danish, and Swedish.

While we did not investigate this in the current version of BioLORD-2023-M, this corpus could be complemented by multilingual annotations from UMLS (and in particular its MESH subset), to increase language coverage. We leave this investigation to a follow-up work.

Evaluation methodology

As mentioned in the abstract, an extensive test suite is required to evaluate the capabilities of semantic models. The following paragraphs list and describe the various benchmarks used to evaluate BioLORD-2023, covering STS, concept representations, and entity linking, all in the biomedical and clinical domains.

Clinical semantic textual similarity

STS is an NLP task measuring the degree of semantic alignment between NLP models and human judgment, by assigning similarity scores to a pair of 2 sentences, usually from 0 to 5, and computing the correlation between scores obtained by expert human judgment and model-assigned scores.

We evaluate the STS capabilities of the models on 5 popular benchmarks: 3 biomedical or clinical ones (MedSTS,[45] MedNLI-S,[46] BIOSSES[47]) and 2 general purposes benchmarks (SICK[48] and STS-Benchmark[49]) For readers unfamiliar with these datasets, we provide a brief introduction for each of them in Appendix SC.

Biomedical concept representation

Known as BCR, this task concerns the mapping of biomedical concepts to a vector latent space, whose features enable classifying these concepts or deriving properties from them. It can be relevant for numerous biomedical tasks including disease subtype annotations.[50]

Following the approach of Kalyan and Sangeetha,[51] we evaluate our model using 4 benchmarks: EHR-RelB,[52] UMNSRS-Similarity,[53] UMNSRS-Relatedness,[53] and MayoSRS.[54] For readers unfamiliar with these datasets, we provide a brief introduction for each of them in Appendix SC.

Biomedical named entity linking

The task of biomedical concept name normalization, also referred to as NEL in the broader literature, concerns the mapping of free-form text describing clinical disorders or concepts to a fixed list of biomedical concepts, such as the elements of the biomedical ontology.

To showcase improvements in NEL, we reuse the evaluation setup devised by Portelli et al,[55] where biomedical language models were evaluated on a set of 5 datasets of varying levels of formality, listed here in the reverse order of formality (least formal first): TwiMed-Twitter,[56] SMM4H,[57] PsyTar,[58] CADEC,[59] and TwiMed-PubMed.[56] For readers unfamiliar with these datasets, we provide a brief introduction for each of them in Appendix SC.

Hyperparameters and training details

In order to focus on the effect of the inclusion of LLM-generated definitions in the training set, and on our improved training strategy including the novel self-distillation step, all the experiments that follow are finetuned from the same base model as BioLORD-2022.

This base model had a size identical to the other baseline models evaluated in this study, enabling a fair comparison between them. We also report some interesting findings about larger models in Appendix SA.

To facilitate the replication of our results, we release the code used in the various phases of our training jointly with this article, and we detail our choices of hyperparameters in Appendix SB.

Results

This section presents the empirical evaluation results of BioLORD-2023, in comparison to existing models. In order to gain insights into the modified training strategy, compared to BioLORD-2022, a number of ablation results are provided as well. Finally, our multilingual model is also evaluated.

For our evaluation of the English BioLORD-2023, we follow a structure similar to BioLORD-2022, and analyze in turn the suitability of the model for several tasks including clinical-STS, BCR, and NEL (as described in the previous section). To produce fair results, all models undergo our multitask finetuning, which was shown to improve results across the board.

We report these results in Table 1. Results for additional baseline models[23,60] are reported in Appendix SD.

We also conduct an ablation study, showing the effect of the various training phases of the BioLORD-2023 methodology. To compare the effect of the training strategies more effectively, we report the absolute improvements over the base model (STAMB2 in all cases, for BioLORD-2022 and 2023 models).

Results are shown in Table 2.

In addition, we separately evaluate our cross-lingual model using both the English test suite (to evaluate the distillation quality) and a multilingual NEL task, XL-BEL.

For XL-BEL, we report the 3 European languages on which both multilingual SapBERT and BioLORD-2023-M were finetuned: German, Spanish, and English.

We report those results in Table 3.

To support the research community, we release on HuggingFace the models we trained, at all 3 stages of our pipeline. This will also enable other researchers to pick the model that is best suited for their experiments.

https://huggingface.co/FremyCompany/BioLORD-2023

Discussion

This section provides insights into the results, structured according to the benchmark tasks, covering the absolute metrics reported in Table 1 and the impact of training strategies as reported in Table 2. After discussing the results for the English BioLORD-2023, we will demonstrate the quality of its multi-lingual counterpart, BioLORD-2023-M, by referring to Table 3.

Clinical semantic textual similarity

As can be seen in Table 1, our new model BioLORD-2023 demonstrates a considerably increased performance on biomedical tasks such as MedSTS (from 86.3 to 88.3), BIOSSES (from 84.0 to 86.1), and MedNLI-S (from 89.9 to 92.4) while also increasing its performance in general-purpose tasks like the STS-Benchmark (from 86.5 to 87.8) and SICK (from 89.3 to 90.3).

Our BioLORD training strategy therefore achieved significant gains in the biomedical domain, while maintaining a high performance in the general-purpose domain. Compared to the original STAMB2 model, our model only suffered a negligible drop of 0.2 and 0.4 points in STS-B and SICK, respectively. These tasks are not relevant to our main objective of enhancing the biomedical knowledge of the model, so we did not optimize our training for them. Therefore, we regard such a minor performance degradation as a success, as it reflects the (now excellent) trade-off between generalization and specialization that is inherent to any finetuning process.

The performance of biomedical semantic models on the general domain was shown in previous studies[36] to be a good indicator of the NEL performance of the models in less formal contexts, such as healthcare information posted on social media, a crucial information source for pharmacovigilance. We also believe that it is a good indication of model robustness, as not all clinical notes are written in formal and unambiguous medical language.

The above results demonstrate with confidence that BioLORD-2023 is the new state-of-the-art semantic model for the biomedical domain.

Biomedical concept representation

In this section, we analyze the performance of the embeddings produced by state-of-the-art models for the task of modeling biomedical concepts.

BioLORD-2023 obtains superior performance on all the considered BCR benchmarks, as reported in Table 1. It performs particularly well on the EHR-Rel-B benchmark (the most exhaustive and recent one) as well as the UMNSRS-Similarity benchmark.

The ablation results of Table 2 show that the evolution of the scores over the training phase follows a similar pattern to the clinical STS tasks already presented in the previous subsection, although there appears to be some trade-off between Similarity tasks and Relatedness tasks.

This time again, the results are consistent with our hypothesis that BioLORD-2023 is the current best biomedical embedder, even beating more complex systems that explicitly combine graphs and text embedders such as Kalyan and Sangeetha[51] and Mao and Fung.[61] We refer to these papers for more details on their results and methodology.

Biomedical named entity linking

Overall, the strong results for the biomedical NEL tasks confirm that BioLORD-2023 performs well both on more formal datasets such as TwiMed-PM and on informal medical datasets, such as TwiMed-TW.

Unlike the STS and BCR tasks, we do not notice a similarly clear pattern of decreased performance for the ablation studies. We suspect that NEL is a task that favors heavily contrastive learning strategies, making the gains of the self-distillation less relevant.

Table 1. Performance characteristics of state-of-the-art biomedical models on STS (Pearson correlation), BCR (Spearman correlation), and NEL (Top1 Accuracy).

		BioSyn[9]	SapBERT[10]	BioLORD-2022	BioLORD-2023
STS	MedSTS[45]	84.0	86.0	86.3	88.3
	MedNLI-S[46]	89.5	90.5	89.9	92.4
	BIOSSES[47]	92.1	89.3	84.0	86.1
	SICK[48]	86.7	80.3	89.3	90.3
	STS[49]	79.4	81.9	86.5	87.8
	(average)	86.3	85.6	87.2	89.0
BCR	EHR-Rel-B[52]	42.5	51.7	57.5	63.6
	UMNSRS-S[53]	43.6	53.0	56.0	59.2
	UMNSRS-R[53]	39.1	47.5	54.4	54.4
	MayoSRS-S[54]	45.1	62.5	74.7	74.4
	(average)	42.6	53.7	60.7	62.9
NEL	TwiMed-TW[56]	42.8	48.3	48.5	49.8
	SMM4H[57]	33.1	43.4	46.5	47.7
	PsyTAR[58]	52.4	64.8	64.7	66.3
	CADEC[59]	35.3	40.4	58.7	63.0
	TwiMed-PM[56]	65.3	70.1	70.4	69.4
	(average)	45.8	53.4	57.8	59.2

The following models are evaluated: BioSyn (state-of-the-art in 2020), SapBERT (state-of-the-art in 2021), BioLORD-2022 (our baseline), and BioLORD-2023 (our new model). Bolding and a color code indicate the best and second-best results for a given task.

Table 2. Performance characteristics of the BioLORD models obtained after each proposed training phase, relative to the STAMB2 performance (in percentage points), on STS (Pearson correlation), BCR (Spearman correlation), and NEL (Top1 Accuracy).

		STAMB2[37] (base)	BioLORD-2022	BioLORD-2023-C	BioLORD-2023-S	BioLORD-2023
STS	MedSTS[45]	85.9	+0.4	+0.4	+1.6	**+2.4**
	MedNLI-S[46]	89.4	+0.5	+2.5	+2.7	**+3.0**
	BIOSSES[47]	90.7	−6.7	**−4.5**	−5.3	−4.6
	SICK[48]	90.7	−1.4	−1.0	−0.8	**−0.4**
	STS[49]	88.0	−1.5	−1.1	−1.0	**−0.2**
	(average)	89.0	+0.5	+1.5	+2.2	+2.7
BCR	EHR-Rel-B[52]	47.1	+10.4	+11.2	+15.7	**+16.5**
	UMNSRS-S[53]	43.9	+12.1	+13.7	**+15.3**	**+15.3**
	UMNSRS-R[53]	46.7	+07.7	+08.0	**+09.1**	+07.7
	MayoSRS-S[54]	54.1	**+20.6**	+18.6	+15.9	+20.3
	(average)	48.0	+12.7	+12.9	+14.0	+15.0
NEL	TwiMed-TW[56]	44.1	+04.4	**+05.7**	+03.7	**+05.7**
	SMM4H[57]	38.4	+08.1	+08.5	+03.2	**+09.3**
	PsyTAR[58]	56.5	+08.2	+08.9	+05.2	**+09.8**
	CADEC[59]	36.9	+21.8	**+26.2**	+24.4	+26.1
	TwiMed-PM[56]	62.8	**+07.6**	+04.9	+05.7	+06.6
	(average)	47.7	+10.0	+10.8	+08.4	+11.5

The following models are evaluated: STAMB2 (our shared base model), BioLORD-2022 (our baseline), BioLORD-2023 (our new model), and an ablation study for each intermediary training phase (BioLORD-2023-C for the contrastive phase, and BioLORD-2023-S for the Self-Distillation phase; see Figure 1). Bolding and a color code indicate the best and second-best results for a given task.

Cross-lingual distillation

In this section, we evaluate the performance of the multilingual BioLORD model (referred to as BioLORD-2023-M) and further discuss its potential.

We evaluate BioLORD-2023-M in 2 ways. Firstly, we evaluate its performance on the same datasets used for our English evaluation, to determine the impact of the multilingual distillation on the model performance. Secondly, we assess its concept name normalization capabilities for languages other than English, such as Spanish and German, using the procedure used by Multilingual SapBERT.

In Table 3, we report the performance of multilingual variants of SapBERT and BioLORD on the English tasks on which their monolingual equivalents were evaluated before.

As a result of the distillation procedure, which focuses on biomedical concept names, BioLORD-2023-M achieves comparable or superior performance on BCR benchmarks compared to BioLORD-2023. When it comes to sentence-level STS tasks, some performance degradation is incurred, but BioLORD-2023-M remains well ahead of its SapBERT competitors. These results indicate that the multilingual distillation procedure was highly effective and displays strong performance on English tasks for which the original model excelled.

When we consider the performance of these multilingual models on NEL tasks in English. This time again, we find comparable performances between the original and the distilled model, albeit the monolingual model usually performs slightly better. In all cases, BioLORD-2023-M performs considerably better than multilingual SapBERT on these tasks.

Finally, we also report in Table 3 the performance of multilingual SapBERT and multilingual BioLORD on the clinical subset of XL-BEL, a dataset specifically developed for the

Table 3. Performance characteristics of state-of-the-art multilingual biomedical models on STS (Pearson correlation), BCR (Spearman correlation), NEL (Top1 Accuracy), and Multilingual NEL (Top1 Accuracy).

		SapBERT[10]	mSapBERT[21]	BioLORD-2023-M	BioLORD-2023
STS	MedSTS[45]	86.0	85.6	86.0	88.3
	MedNLI-S[46]	90.5	88.1	92.1	92.4
	BIOSSES[47]	89.3	90.0	75.4	86.1
	SICK[48]	80.3	87.0	89.1	90.3
	STS[49]	81.9	83.5	85.1	87.8
	(average)	85.6	86.8	85.5	89.0
BCR	EHR-Rel-B[52]	51.7	42.4	64.1	63.6
	UMNSRS-S[53]	53.0	34.2	60.1	59.2
	UMNSRS-R[53]	47.5	29.6	54.3	54.4
	MayoSRS-S[54]	62.5	45.2	74.8	74.4
	(average)	53.7	37.9	63.3	62.9
NEL	TwiMed-TW[56]	48.3	47.4	49.3	47.4
	SMM4H[57]	43.4	40.8	42.3	46.9
	PsyTAR[58]	64.8	51.5	63.3	66.3
	CADEC[59]	40.4	46.8	47.0	47.4
	TwiMed-PM[56]	70.1	63.9	67.4	69.4
	(average)	53.4	50.4	53.9	55.5
MNEL	German XLB[21]	N/A	51.5	57.7	N/A
	Spanish XLB[21]	N/A	52.7	53.1	N/A
	English XLB[21]	N/A	78.2	73.1	N/A
	(average)	N/A	58.1	59.4	N/A

Bolding indicates the best results among multilingual models for a given task. (English-only models are only provided for comparison purposes.)

evaluation of multilingual SapBERT. Because we noted that a large proportion of the evaluation set concerns the names of plant and animal species, which are not very relevant to the clinical setting these models were developed for, we filtered the XL-BEL test set to exclude these mention types (based on their UMLS Semantic Types). This enables the evaluation to focus on all the other semantic types, such as clinical disorders, drugs, and procedures.

Overall, BioLORD-2023-M achieves better results compared to multilingual SapBERT on both German and Spanish, the 2 non-English languages supported by both models. Interestingly, multilingual SapBERT keeps an edge when it comes to English data, but monolingual models would be better suited for this task than a multilingual model.

In summary, the results of this section demonstrate that our multilingual BioLORD-2023-M model achieves comparable performance on multilingual NEL as the multilingual SapBERT model while achieving considerably better results on the English STS and BCR tasks than both the English and Multilingual SapBERT on these datasets. This makes our multilingual model a solid choice for a large range of biomedical tasks in the supported languages.

Qualitative assessment in the clinical domain

To complement the quantitative evaluation of our model, we also performed a qualitative error analysis to assess the clinical validity of BioLORD-2023 as a foundation model for practical applications in the future. This is especially important in the clinical domain, where errors can have serious consequences for patient care and research outcomes.

Therefore, we conducted a qualitative error analysis of the NEL results on the PsyTar dataset [59], which contains patient-reported symptoms and adverse reactions related to psychological disorders and treatments. We randomly selected 200 entity mentions from the test set and manually evaluated the predictions of BioLORD-2023 and 3 other state-of-the-art models.

We present the details of our qualitative error analysis in Appendix SE, where we show that BioLORD-2023 not only achieves the highest exact match score by several points but also reduces the number of severe errors by a very large margin compared to the other models. Moreover, our Sankey diagrams show that most of the errors made by BioLORD-2023 are shared by the other models, suggesting that model ensembling would have limited potential.

We attribute these advantages to the improved training strategy and data of BioLORD-2023, which enabled it to learn more fine-grained and robust representations of biomedical concepts and sentences.

Conclusion

In this study, we introduced BioLORD-2023, a model offering state-of-the-art capabilities for clinical STS and concept representation.

Through the introduction of innovative techniques such as the inclusion of LLM-generated definitions in the training data, a supervised self-distillation phase, a robust model-weights averaging, and a state-of-the-art cross-lingual distillation, we were able to train a series of models which confidently demonstrates substantial performance improvements over their predecessors across a wide range of tasks.

These enhancements can have real-world impact, as they empower researchers to grasp the global picture of diseases by tackling the challenges of complex biomedical literature, uncover the secrets of genes, proteins, and pathways, thereby accelerating the pace of biomedical research and drug discovery.

Moreover, our models enable a nuanced understanding of patient records, extracting meaningful insights that may have otherwise remained buried. This better comprehension of clinical narratives can contribute to more accurate diagnoses, personalized treatment plans, and improved patient outcomes.

Limitations and future work

While our work showcases an impressive improvement over the previous state of the art, it is not without limitations. One such limitation is the bound of its knowledge, which is constrained by the knowledge contained in the knowledge graphs used to prompt the language model.

Many entities in the knowledge graph contain few pieces of information, and this is particularly the case for organisms and species, making BioLORD models inadequate to differentiate between such entities, as they acquire very similar representations. Another limitation is that knowledge graphs trail the literature and might not include all new pieces of knowledge mentioned in the published biomedical papers.

To address this problem, we foresee a future version of this model which would rely on a combination of the knowledge graph information and of the recent biomedical literature to prompt LLMs before generating definitions, thereby including even more recent and relevant information, and helping create a more precise definitions for rare or very specific concepts.

The retrieval of the relevant documents to include in the prompt will already benefit from our state-of-the-art BioLORD-2023 model. We hope to study the impact of grounding document retrieval on the generated definitions in an upcoming study.

We also foresee future versions of BioLORD or similar semantic models making use of the larger STS models that are starting to become available. While most of the successful models remain closed sourced so far and can therefore not be freely finetuned for the biomedical domain, new and larger models are expected to be open-sourced as time goes by. Applying the self-distillation phase of BioLORD to these models could prove an efficient way to leverage them.

Acknowledgments

This work would not have been possible without the joint financial support of the Vlaams Agentschap Innoveren & Ondernemen (VLAIO) and the RADar innovation center of the AZ Delta hospital group, through their ambitious joint venture in the Advanced Data-Aided Medicine project (ADAM).

I would, of course, like to extend special thanks to Karel D'Oosterlinck, Alfiya Khabibullina, Beatrice Portelli, Simone Scarboro, Peter De Jaeger, and other individuals who helped me make progress along the way toward this journal publication, often in the form of joint conference or workshop publications.

A particular thanks is also extended to the team maintaining and developing GPULab, the machine learning infrastructure for AI computing built in collaboration between UGent, UAntwerpen, and the imec research and development center.

Finally, I would also like to thank my co-supervisors, Thomas Demeester and Kris Demuynck, for their support and constructive advice during the ideation process, and all along the development of this project up to this very article.

Author contributions

We, the authors of this work, François Remy, Kris Demuynck, and Thomas Demeester, collectively contributed to every stage of this project in accordance with the ICMJE guidelines for authorship. In the initial phase, we collaboratively conceived and designed the project goals, agenda, and methodologies. Subsequently, François Remy assumed the role of lead researcher, spearheading the execution of experiments, while Kris Demuynck and Thomas Demeester provided oversight and guidance in a supervisory capacity, with biweekly meetings. Throughout the project's evolution, all authors actively participated in shaping refinements and improvements. We confirm that all authors reviewed and approved each revision of this publication.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the "ADAM" grant of VLAIO O&O (HBC.2020.3234).

Conflicts of interest

None declared.

Data availability

Our training datasets (BioLORD, AGCT) and models (BioLORD-2023, BioLORD-2023-M) are all available for download on HuggingFace.

References

1. Houssein EH, Rehab EM, Abdelmgeid AA. Machine learning techniques for biomedical natural language processing: a comprehensive review. *IEEE Access*. 2021;9:140628-140653.
2. Shi J, Yuan Z, Guo W, Ma C, Chen J, Zhang M. Knowledge-graph-enabled biomedical entity linking: a survey. *World Wide Web*. 2023;26(5):2593-2622.
3. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: a roadmap. arXiv. June 14, 2023, preprint: not peer reviewed.
4. Satvik G, Shivam P, Somya G. Navigating healthcare insights: a birds eye view of explainability with knowledge graphs. In: *2023 IEEE Sixth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. 2023:54-61.
5. Lin X, Dai L, Zhou Y, et al. Comprehensive evaluation of deep and graph learning on drug-drug interactions prediction. *Brief Bioinform*. 2023;24(4):1-20.
6. Wang B, Xie Q, Pei J, et al. Pre-trained language models in biomedical domain: a systematic survey. *ACM Comput Surv*. 2023;56(3):1-52.
7. Hu L, Liu Z, Zhao Z, Hou L, Nie L, Li J. A survey of knowledge enhanced pre-trained language models. *IEEE Tran Knowl Data Eng*. 2023. https://doi.org/10.1109/TKDE.2023.3310002
8. Feng Z, Ma W, Yu W, et al. Trends in integration of knowledge and large language models: a survey and taxonomy of methods, benchmarks, and applications. arXiv. November 10, 2023, preprint: not peer reviewed.
9. Sung M, Jeon H, Lee J, Kang J. Biomedical entity representations with synonym marginalization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020:3641-3650.
10. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. In: *Proceedings of the 2021 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2021:4228-4238.

11. Remy F, Demuynck K, Demeester T. BioLORD: learning ontological representations from definitions for biomedical concepts and their textual descriptions. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*, December 2022: 1454-1465.
12. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings *of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019:4171-4186.
13. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*, December 2020:9459-9474.
14. Kim J, Choi S, Amplayo RK, Hwang SW. Retrieval-augmented controllable review generation. In: *Proceedings of the 28th International Conference on Computational Linguistics*, December 2020:2284-2295.
15. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *Adv Neural Inform Proc Syst*. 2022;35:27730-27744.
16. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32 (Database issue):D267-D270.
17. Coté RA, Robboy S. Progress in medical information management. Systematized nomenclature of medicine (SNOMED). *JAMA*. 1980;243(8):756-762.
18. Remy F, Demuynck K, Demeester T. 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*, July 2023:265-272.
19. Reimers N, Gurevych I. Making monolingual sentence embeddings multilingual using knowledge distillation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November 2020:4512-4525.
20. Feng F, Yang Y, Cer D, Arivazhagan N, Wang W. Language-agnostic BERT sentence embedding. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, May 2022:878-891.
21. Liu F, Vulić I, Korhonen A, Collier N. Learning domain-specialised representations for cross-lingual biomedical entity linking. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 2, August 2021:565-574.
22. Cui H, Lu J, Wang S, et al. A survey on knowledge graphs for healthcare: resources, applications, and promises. arXiv. August 23, 2023, preprint: not peer reviewed.
23. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194.
24. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180.
25. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388 (13):1233-1239.
26. Venkatesh KP, Raza MM, Kvedar JC. Automating the overburdened clinical coding system: challenges and next steps. *NPJ Digit Med*. 2023;36(1):16.
27. Wu C, Lei J, Zheng Q, et al. Can GPT-4V(ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. arXiv. October 17, 2023, preprint: not peer reviewed.
28. Yan Z, Zhang K, Zhou R, He L, Li X, Sun L. Multimodal ChatGPT for medical applications: an experimental study of GPT-4V. arXiv. October 29, 2023, preprint: not peer reviewed.
29. Jin Q, Wang Z, Floudas CS, Sun J, Lu Z. Matching patients to clinical trials with large language models. arXiv. July 28, 2023, preprint: not peer reviewed.
30. Taylor R, Kardas M, Cucurull G, et al. Galactica: a large language model for science. arXiv. November 16, 2022, preprint: not peer reviewed.
31. Wang G, Yang G, Du Z, Fan L, Li X. ClinicalGPT: large language models finetuned with diverse medical data and comprehensive evaluation. arXiv. June 16, 2023, preprint: not peer reviewed.
32. Bolton E, et al. Stanford CRFM introduces PubMedGPT 2.7B. Stanford University, 2022. Accessed May 2023. https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b
33. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. 2022;23(6):bbac409.
34. OpenAI. GPT-4 technical report. arXiv. March 15, 2023, preprint: not peer reviewed.
35. Oord AV, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv. July 10, 2018, preprint: not peer reviewed.
36. Remy F, Scaboro S, Portelli B. Boosting adverse drug event normalization on social media: general-purpose model initialization and biomedical semantic text similarity benefit zero-shot linking in informal contexts. In: *Proceedings of the Eleventh International Workshop on Natural Language Processing for Social Media*, November 1, 2023:47-52.
37. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, November 2019:3982-3992.
38. Ribeiro J, Melo FS, Dias J. Multi-task learning and catastrophic forgetting in continual reinforcement learning. *EPiC Ser Comput*. 2019;65:163-175.
39. He T, Liu J, Cho K, et al. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, May 2021:1121-1133.
40. Gao T, Yao X, Chen D. SimCSE: simple contrastive learning of sentence embeddings. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, November 2021:6894-6910.
41. Li B, Zhou H, He J, Wang M, Yang Y, Li L. On the sentence embeddings from pre-trained language models. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November 2020:9119-9130.
42. Ethayarajh K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, November 2019:55-65.
43. Wortsman M, Ilharco G, Gadre SY, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: *International Conference on Machine Learning*, vol. 162, 2022:23965-23998.
44. Remy F, De Jaeger P, Demuynck K. Taming large lexicons: translating clinical text using medical ontologies and sentence templates. In: *EmP: 1st RADar Conference on Engineer Meets Physician*, 2022:1-5.
45. Wang Y, Afzal N, Fu S, et al. MedSTS: a resource for clinical semantic textual similarity. *Lang Resourc Eval*. 2020;54(1):57-72.
46. Romanov A, Shivade C. Lessons from natural language inference in the clinical domain. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, November 2018, 1586-1596.
47. Sogancioglu G, Öztürk H, Özgür A. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*. 2017;33(14):i49-i58.
48. Marelli M, Menini S, Baroni M, Bentivogli L, Bernardi R, Zamparelli R. A SICK cure for the evaluation of compositional

distributional semantic models. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, May 2014:216-233.

49. Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 Task 1: semantic textual similarity multilingual and crosslingual focused evaluation. In: *Proceedings of the 11th International Workshop on Semantic Evaluation*, August 2017:1-14.
50. Ofer D, Linial M. Automated annotation of disease subtypes. medRxiv. September 25, 2023), preprint: not peer reviewed.
51. Kalyan KS, Sangeetha S. Hybrid approach to measure semantic relatedness in biomedical concepts. arXiv. January 25, 2021, preprint: not peer reviewed.
52. Schulz C, Levy-Kramer J, Van Assel C, Kepes M, Hammerla N. Biomedical concept relatedness—a large EHR-based benchmark. In: *Proceedings of the 28th International Conference on Computational Linguistics*, December 2020:6565-6575.
53. Pakhomov SV, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic similarity and relatedness between clinical terms: an experimental study. *AMIA Annu Symp. Proc*. 2010;2010:572.
54. Pakhomov SV, Pedersen T, McInnes B, Melton GB, Ruggieri A, Chute CG. Towards a framework for developing semantic relatedness reference standards. *J Biomed Inform*. 2011;44(2):251-265.
55. Portelli B, Scaboro S, Santus E, Sedghamiz H, Chersoni E, Serra G. Generalizing over long tail concepts for medical term normalization. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, December 2022:8580-8591.
56. Alvaro N, Miyao Y, Collier N. TwiMed: twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR Public Health Surveill*. 2017;3(2):e24.
57. Gonzalez-Hernandez G, Klein AZ, Flores I, et al. Overview of the Fifth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at COLING 2020. In: *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*. Association for Computational Linguistics, December 2020:27-36.
58. Zolnoori M, Fung KW, Patrick TB, et al. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data Brief*. 2019;24:103838.
59. Karimi S, Metke-Jimenez A, Kemp M, Wang C. Cadec: a corpus of adverse drug event annotations. *J Biomed Inform*. 2015;55:73-81.
60. Jin Q, Kim W, Chen Q, et al. MedCPT: contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*. 2023;39(11): btad651.
61. Mao Y, Fung KW. Use of word and graph embedding to measure semantic relatedness between Unified Medical Language System concepts. *J Am Med Inform Assoc*. 2020;27(10):1538-1546.