Semantic Text Similarity for Civil Engineering Domain to Enhance Data Interoperability: A Domain-Specific Embedding Approach

Elham Farazdaghi, Hamed Asadollahi, Mojtaba Eslahi, Muhammad Ali Sammuneh, Rani El Meouche

***Abstract*—This paper investigates the application of semantic text similarity embedding models to address data interoperability challenges in the construction and civil engineering sectors. We focus on improving embedding models to produce high-quality domain-specific embeddings that precisely capture complex terminology and relationships among various building systems, including Building Information Modeling, Building Energy Modeling, and Building Management Systems. Our methodology includes a comprehensive data collection and preprocessing pipeline, followed by a multi-stage fine-tuning approach that employs continuous pre-training, task-specific fine-tuning, and parameter-efficient adaptation algorithms. We assess three parameter-efficient fine-tuning techniques: Low-Rank Adaptation, Rank-Stabilized Low-Rank Adaptation, and Weight-Decomposed Low-Rank Adaptation. Our evaluation demonstrates that domain-adapted models, especially, substantially surpass baseline language models in domain-specific tasks, including understanding technical terms and data alignment across domains. The results show significant performance enhancements with minimal resource investment, making these approaches practical for real-world applications in the building industry.**

***Keywords*—Building information systems, Data interoperability, Domain-specific embedding models, Semantic text similarity.**

I. Introduction

The building industry has become increasingly complex, involving creating and managing various models for design, construction, and operation. These models, developed using distinct tools and managed by separate teams, often lack standardization, leading to significant challenges in data coordination. For instance, Building Information Modeling (BIM) focuses on physical and functional characteristics of a structure, Building Energy Modeling (BEM) evaluates its energy performance, and Building Management Systems (BMS) oversee operational controls and automation.

Elham Farazdaghi, Hamed Asadollahi, Mojtaba Eslahi, Muhammad Ali Sammuneh, Rani El Meouche, Institut de Recherche de la construction, ESTP, 28 Avenue du Président Wilson, F-94230, Cachan, France,

The heterogeneity of these data sources hinders the integration of information critical to the development of smart building applications, impacting the efficiency and scalability of intelligent systems in the industry. In response to these challenges, recent advancements in semantic technologies have emerged as a promising approach to harmonize and integrate disparate data sources within the building domain. Semantic web technologies, including ontologies like Brick and query languages like SPARQL, provide a structured way to represent relationships between various data points. However, these technologies are still evolving and often require domain-specific adaptation to realize their full potential. One approach to enhancing semantic integration is the application of embedding models that capture semantic text similarity, particularly when fine-tuned on domain-specific data. While general-purpose Large Language Models (LLMs) have shown remarkable capabilities in various Natural Language Processing (NLP) tasks, there is a growing need for domain-specific models tailored to specific industries or fields. Domain-specific models can better understand industry-specific terminology, concepts, and context, developing deep expertise in their target domains and enabling them to handle complex tasks that require specialized knowledge. In this paper, we refine general LLMs to explore the use of embedding models for semantic text similarity in the building domain. We focus on how these models can be adapted to address the distinct needs of the building domain, demonstrating the potential for improved integration of BIM, BEM, and BMS systems, facilitating more cohesive data management and analysis.

II. State of the Art

The architecture, engineering, and construction (AEC) industry is characterized by its inherently complex and multidisciplinary nature, requiring the coordination of numerous stakeholders and interdependent processes throughout a building's lifecycle [1]. This complexity is particularly evident in the diverse data models and systems employed across the industry, creating significant challenges for seamless information exchange and comprehensive data analysis [2]. The collaborative ecosystem encompasses architects, structural engineers, contractors, and facility managers, with each professional contributing specialized

https://doi.org/10.17758/URUAE27.UR0625304

knowledge and introducing unique data requirements to projects [3]. These diverse professional needs have fostered a fragmented landscape of data models, formats, and storage systems ranging from Computer-Aided Design (CAD) drawings and Building Information Modelling (BIM) platforms to spreadsheets and text-based documentation [4]. According to Vanlande et al. [5], this heterogeneity of data creates substantial barriers to interoperability within the industry. Research indicates that each stakeholder frequently captures similar concepts using different terminologies or classification systems, significantly complicating the integration and analysis of information across platforms [6]. Berard and Karlshoej [7] argue that this lack of semantic standardisation impedes critical workflows, including design coordination, construction scheduling, and facility management. Furthermore, as noted by Bilal et al [8], these interoperability challenges limit the potential for implementing advanced analytics and data-driven decision-making processes in building projects. The fragmentation of information systems represents a significant obstacle to achieving greater efficiency and sustainability in the built environment [9].

A. Semantic Text Similarity and Embedding Models

Semantic Text Similarity (STS) emerged as a critical area of research in Natural Language Processing (NLP) as researchers sought more sophisticated ways to measure meaningful relationships between texts beyond simple lexical matching [10] .STS aims to capture the semantic content and conceptual connections between texts, regardless of surface-level variations in their wording or structure [11]. According to Reimers and Gurevych [12], these capabilities have proven essential for numerous applications including information retrieval, question answering, and text summarization. The evolution of STS has been closely tied to developments in distributional semantics [13]. Early approaches relied on statistical methods such as Latent Semantic Analysis (LSA), which represented words as vectors based on their co-occurrence patterns in large text corpora [14]. As noted by Mikolov et al. [15], these techniques were foundational in establishing the computational framework for semantic similarity. A paradigm shift occurred with the introduction of neural word embeddings, particularly Word2Vec [16] and GloVe [17], which revolutionised how machines represent and understand language. These approaches leveraged neural networks to learn dense vector representations that effectively captured semantic relationships between words [18]. Building on these foundations, the development of sentence-level encoders such as the Universal Sentence Encoder [19] and Sentence-BERT [12] further enhanced the ability to capture sentence-level semantic relationships directly, improving performance in tasks requiring deeper linguistic understanding. As demonstrated by Yang et al. [20], these models enabled more nuanced comparison of textual content across domains and applications. The emergence of contextual embeddings represents one of the most significant paradigm shifts in natural language processing [21]. While traditional embedding models like Word2Vec assigned each word a single, static vector regardless of context, contextual embeddings generate representations that change based on the surrounding context, capturing the subtle, shifting nature of language [22]. ELMo (Embeddings from Language Models) marked the first major breakthrough in contextual embeddings, using a bidirectional LSTM architecture to generate word representations on the fly [21]. BERT (Bidirectional Encoder Representations from Transformers) further revolutionized NLP capabilities through its transformer architecture and self-attention mechanism that could consider all words in a sentence simultaneously, creating deeply contextualized representations [23]. These advances fundamentally changed how semantic similarity could be computed between texts by incorporating contextual nuances previously inaccessible to computational systems [24].

B. Domain Adaptation of Embedding Models

Domain adaptation addresses the critical challenge that general-purpose language representations often perform suboptimally when applied to specialized domains with unique terminology, semantic relationships, and linguistic patterns , [25]. As observed by Ruder and Plank [26], the distributional shift between pre-training corpora and domain-specific applications creates substantial challenges for model generalization, necessitating specialized adaptation techniques to achieve optimal performance in target domains. Several methodologies have emerged to adapt embedding models to specific domains, each offering distinct advantages depending on computational resources, available data, and adaptation objectives [27]. These approaches can be categorized as follows:

1. **Continued pre-training approaches**, which extend the training of existing general models on domain-specific corpora [28], [29]. This approach leverages transfer learning principles to efficiently adapt pre-trained representations to new domains while retaining general linguistic knowledge. Lee et al. [30] demonstrated that this approach is particularly effective when the target domain shares some linguistic features with the general domain.
2. **Feature-based approaches** that integrate domain knowledge through supplementary features [21]. According to Akbik et al. [31], these methods enhance model performance by incorporating domain-specific indicators alongside general language representations, creating hybrid models that benefit from both general and specialized knowledge.
3. **Knowledge-enhanced methods** that explicitly incorporate structured domain knowledge [32], [33]. As noted by Yang et al. [34], these approaches integrate external knowledge bases, ontologies, or semantic networks to enrich language representations with domain-specific concepts and relationships.
4. **Domain-specific pre-training** that builds embeddings from scratch using exclusively in-domain texts [35], [25]. While computationally intensive, this approach, as described by Kerner and Koto et al. [36], [37] , can be optimal when the target domain differs substantially from

general language in vocabulary, syntax, or discourse patterns.

The effectiveness of domain adaptation manifests in significant performance improvements on downstream tasks within the target domain [28]. Adapted models demonstrate superior ability to disambiguate domain-specific terminology, recognize specialized relationships, and generate more relevant text completions compared to their general-purpose counterparts [38]. As empirically demonstrated by Gu et al. [35], domain-adapted models can achieve performance gains of 3-15% on specialized tasks such as biomedical entity recognition, legal document classification, and scientific relation extraction. There has been a significant advancement in the evolution of fine-tuning approaches for large language models in recent years. Full fine-tuning involves updating all parameters of a pre-trained model for downstream tasks, requiring substantial computational resources, particularly as model sizes increase [39], [40]. According to Kaplan et al. [41], the computational requirements scale linearly with the number of parameters, making full fine-tuning prohibitively expensive for many research groups and practitioners as models exceed hundreds of billions of parameters. To address these limitations, parameter-efficient fine-tuning (PEFT) methods have emerged as a more resource-conscious alternative [42], [43]. These approaches substantially reduce memory and computational requirements while preserving adaptation capabilities, as demonstrated in comprehensive studies by Hu et al. [44]and Lialin et al. [45].

Low-Rank Adaptation (LoRA) minimizes computational overhead by approximating weight updates through low-rank decomposition, achieving comparable performance to full fine-tuning while training less than 1% of the model's parameters [44]. As noted by Dettmers et al. [46], LoRA operates by decomposing weight updates into products of smaller matrices, thereby significantly reducing memory footprint during training. Experimental results from Li et al. [47] demonstrate that LoRA can achieve 95-99% of full fine-tuning performance across various natural language understanding benchmarks while requiring only a fraction of the computational resources.

Rank-Stabilized Low-Rank Adaptation (RSLoRA) enhances performance in structurally complex engineering tasks by mitigating rank collapse issues, particularly beneficial for specialized construction terminology and regulatory compliance tasks [48]. According to Zhang et al. [49], rank collapse occurs when the effective rank of update matrices diminishes during training, limiting representational capacity. RSLoRA addresses this challenge through orthogonality constraints and adaptive regularization techniques, enabling more robust adaptation to specialized domains. Empirical evaluations by [50] demonstrate that RSLoRA outperforms standard LoRA by 3-5% on domain-specific tasks requiring precise technical knowledge.

Weight-Decomposed Low-Rank Adaptation (DoRA) demonstrated superior performance in civil engineering contexts by separately updating magnitude and direction components of weight matrices, enabling more nuanced adaptation to the precise technical specifications common in structural analysis and construction planning [51]. This innovative approach, as explained by Nie et al. [52], decouples the optimization of directional and scalar components, allowing for more effective parameter updates. Experimental results indicate that DoRA achieves 5-8% improvement over conventional LoRA in tasks involving complex numerical reasoning and specialized terminology, while maintaining similar computational efficiency.

These PEFT methods have democratized model adaptation by enabling the customization of state-of-the-art models on consumer hardware while preserving most of the performance benefits associated with full fine-tuning [53], [45]. As documented by Xu et al. [54], researchers and practitioners with limited computational resources can now fine-tune multi-billion parameter models on single GPUs with 16-32GB of memory, a task previously requiring specialized hardware clusters. This accessibility has significantly expanded the application of LLMs across diverse domains and use cases, fostering innovation and specialized implementations across academic and industrial settings.

III. METHODOLOGY

In this research, we selected the pre-trained embedding model BAAI/bge-large-en-v1.5 [55] due to its optimal balance of performance and computational efficiency. This model provided an excellent starting point for our domain-specific adaptation.

A. Data Collection and Pre-processing

We compiled a comprehensive dataset from multiple sources within the building domain, including BIM documentation, BEM specifications, BMS manuals, and technical vocabulary from specialized dictionaries. The dataset was designed to be comprehensive, covering terminology spanning from classical to green architecture, traditional materials to modern products, and various aspects of building services. To ensure representativeness, we curated data reflecting diverse building types, climates, and regulatory contexts. The pre-processing stage involved refining the text data, removing irrelevant information, standardizing formats, and correcting errors. Special care was taken to preserve domain-specific abbreviations, acronyms, and technical terms. For text extraction, we employed pattern recognition for each document to collect terms and definitions, complemented by Named Entity Recognition (NER) with a pre-trained model. After extraction, we mapped synonymous terms and concepts. For data augmentation, we applied back translation, random swap, insert, and delete operations. The final dataset was carefully balanced to ensure adequate representation of different sub-domains within the building industry.

B. Model Architecture and Fine-tuning

After careful comparative analysis against other leading embedding models, we selected the BAAI/bge-large-en-v1.5

model as our foundation. Our domain adaptation strategy focuses on three parameter-efficient fine-tuning techniques:

Low-Rank Adaptation (LoRA)

LoRA introduces the concept of freezing pre-trained weights and injecting trainable rank decomposition matrices. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{dxk}$, LoRA defines the adapted weights as:

$$W = W_0 + \Delta W = W_0 + BA \quad (1)$$

where $A \in \mathbb{R}^{rxd}$, $B \in \mathbb{R}^{dxr}$, and $r \ll \min(d, k)$.

Rank-Stabilized Low-Rank Adaptation (RSLoRA)

RSLoRA extends LoRA by incorporating regularization techniques to prevent rank collapse during training:

$$W = W_0 + BA, \text{ with } R(A, B) \leq \varepsilon \quad (2)$$

where R represents the rank stabilization regularizer and ε is a threshold parameter.

Weight-Decomposed Low-Rank Adaptation (DoRA)

DoRA decomposes the weights into magnitude and direction components:

$$W = m \odot D \quad (3)$$

with m representing magnitude vectors and D normalized direction matrices ($\|D_i\|_2 = 1$). The adaptation process then becomes:

$$W = (m_0 + \Delta m) \odot (D_0 + \Delta D)/\|D_0 + \Delta D\|_2 \quad (4)$$

where $\Delta D = BA$ with constrained rank r.

Our training configuration implemented a comprehensive hyperparameter optimization protocol with a multidimensional parameter space exploration. We integrated an advanced pruning mechanism that dynamically calibrated model parameters according to performance metrics, while concurrently deploying callback functions for comprehensive performance monitoring.

IV. Experimental Results and Discussion

Our experiments revealed significant improvements in semantic text similarity tasks within the building and civil engineering domain when using our fine-tuned models. The evaluation demonstrated that all three parameter-efficient techniques substantially improved over the baseline BAAI/bge-large-en-v1.5 model.

V. Performance Evaluation

Our experiments revealed significant improvements in semantic text similarity tasks within the building and civil engineering domain when using our fine-tuned models. The evaluation demonstrated that all three parameter-efficient techniques substantially improved over the baseline BAAI/bge-large-en-v1.5 model.

A. Performance Evaluation

We evaluated the models using both intrinsic and domain-specific benchmarks. Table 1 shows performance comparisons across various metrics.

TABLE I: Performance of PEFT Methods on Different Metrics

Task Type	Base BGE	LoRA	DoRA	RSLoRA
Spearman Correlation	0.828	0.863	0.879	0.885
Precision	0.781	0.817	0.836	0.843
Recall	0.769	0.809	0.831	0.837
F1 Score	0.775	0.813	0.833	0.840
MRR	0.809	0.842	0.860	0.866
Training Time (hours)	-	25	26	38

TABLE II: Performance of PEFT Models on Different Datasets

Task Type	Base BGE	LoRA	DoRA	RSLoRA
HVAC Terminology	76.3%	84.7%	88.9%	87.5%
Building Codes	71.8%	82.3%	85.1%	84.2%
Material Specifications	74.5%	83.1%	87.4%	86.0%
System Components	73.9%	81.8%	86.2%	85.0%

We also tested the models on specific building domain datasets (see Table 2).

B. Comparative Analysis

DoRA consistently outperformed both LoRA and RSLoRA across all evaluation metrics, achieving the highest scores in Spearman correlation, precision, recall, F1 score, and Mean Reciprocal Rank. The performance gains were particularly significant in specialized building engineering subdomains, with DoRA showing a 12.6% improvement over the baseline in HVAC terminology understanding and a 13.3% improvement in building codes interpretation. However, this superior performance came with a computational cost trade-off. DoRA required approximately 35% more training time than LoRA, taking 38 hours compared to LoRA's 25 hours. RSLoRA offered a balanced middle ground, with performance closer to DoRA but with a training time like LoRA (26 hours).

C. Error Analysis

Our systematic error analysis revealed distinct error patterns across the different fine-tuning methods:

1. LoRA demonstrated pronounced difficulties with closely related technical terms, showing a 23.7% error rate for closely related technical pairs (e.g., "variable refrigerant flow" vs. "variable air volume").

2. RSLoRA showed systematic under-adaptation to emerging construction terminology, with a 22.8% higher error rate on terms introduced within the past 2 years compared to established terminology.
3. DoRA, while achieving the best overall performance, exhibited a pattern of high-confidence errors, particularly with emerging terminology. Unlike other methods, DoRA's errors showed high confidence scores (avg. 0.82) and an elevated false positive rate (17.3%) for emerging terminology classification.

The most challenging areas across all methods were context-dependent meanings (e.g., "slab" in structural versus geotechnical contexts) and domain-specific nuances (e.g., distinguishing between "compressive strength" and "tensile strength").

VI. Conclusions and Future Work

Our study demonstrates the effectiveness of parameter-efficient fine-tuning techniques in adapting embedding models to the building and civil engineering domain. The key findings include:

1. Domain-specific fine-tuning significantly improves semantic text similarity performance, with an average improvement of 12.8% using the DoRA approach.
2. DoRA provides the best performance with a 15% improvement over the baseline model although it requires around 35% more training time compared to LoRA.
3. RSLoRA offers a good balance of improvement and training efficiency, making it suitable for scenarios with limited computational constraints.
4. All models struggle with emerging terminology and context-dependent meanings, suggesting areas for future improvement.

These findings have important implications for addressing data interoperability challenges in the building industry. The fine-tuned models can significantly enhance the integration of information across BIM, BEM, and BMS systems, leading to more cohesive data management and analysis.

Future work will focus on expanding the training dataset to include a broader range of technical documents, exploring the integration of unsupervised learning techniques prior to supervised training, and developing a comprehensive benchmark database for standardized evaluation of models in the construction domain. Our ultimate vision is to develop a comprehensive framework that seamlessly bridges the semantic and structural gaps across heterogeneous data schemas used in the construction and building industry, serving as a universal translator that enables different systems to communicate effortlessly and data exchange between systems without loss of meaning or context.

References

[1] C. M. Eastman, Ed., BIM handbook: a guide to building information modeling for owners, managers, designers, engineers and contractors, 2. ed. Hoboken, NJ: Wiley, 2011.

[2] R. Sacks, C. Eastman, G. Lee, and P. Teicholz, BIM Handbook: A Guide to Building Information Modeling for Owners, Designers, Engineers, Contractors, and Facility Managers, 1st ed. Wiley, 2018. doi: 10.1002/9781119287568. https://doi.org/10.1002/9781119287568

[3] B. Hardin and D. McCool, BIM and construction management: proven tools, methods, and workflows, Second edition. Indianapolis, Indiana: Sybex, a Wiley brand, 2015.

[4] A. Aksamija, Integrating innovation in architecture: design, methods and technology for progressive practice and research. in AD smart, no. 04. West Sussex, United Kingdom: John Wiley & Sons, Ltd, 2016. https://doi.org/10.1002/9781119164807

[5] R. Vanlande, C. Nicolle, and C. Cruz, "IFC and building lifecycle management," Autom. Constr., vol. 18, no. 1, pp. 70–78, Dec. 2008, doi: 10.1016/j.autcon.2008.05.001. https://doi.org/10.1016/j.autcon.2008.05.001

[6] P. Pauwels, S. Zhang, and Y.-C. Lee, "Semantic web technologies in AEC industry: A literature overview," Autom. Constr., vol. 73, pp. 145–165, Jan. 2017, doi: 10.1016/j.autcon.2016.10.003. https://doi.org/10.1016/j.autcon.2016.10.003

[7] O. Berard and J. Karlshoej, "INFORMATION DELIVERY MANUALS TO INTEGRATE BUILDING PRODUCT INFORMATION INTO DESIGN".

[8] M. Bilal et al., "Big Data in the construction industry: A review of present status, opportunities, and future trends," Adv. Eng. Inform., vol. 30, no. 3, pp. 500–521, Aug. 2016, doi: 10.1016/j.aei.2016.07.001. https://doi.org/10.1016/j.aei.2016.07.001

[9] Y. Arayici, T. Fernando, V. Munoz, and M. Bassanino, "Interoperability specification development for integrated BIM use in performance based design," Autom. Constr., vol. 85, pp. 167–181, Jan. 2018, doi: 10.1016/j.autcon.2017.10.018. https://doi.org/10.1016/j.autcon.2017.10.018

[10] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity".

[11] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation," 2017, doi: 10.48550/ARXIV.1708.00055. https://doi.org/10.18653/v1/S17-2001

[12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," 2019, arXiv. doi: 10.48550/ARXIV.1908.10084.

[13] P. D. Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," 2010, doi: 10.48550/ARXIV.1003.1141. https://doi.org/10.1613/jair.2934

[14] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.," Psychol. Rev., vol. 104, no. 2, pp. 211–240, Apr. 1997, doi: 10.1037/0033-295X.104.2.211. https://doi.org/10.1037/0033-295X.104.2.211

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013, arXiv. doi: 10.48550/ARXIV.1301.3781.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," 2013, arXiv. doi: 10.48550/ARXIV.1310.4546.

[17] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162 https://doi.org/10.3115/v1/D14-1162

[18] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," 2016, arXiv. doi: 10.48550/ARXIV.1607.04606. https://doi.org/10.1162/tacl_a_00051

[19] D. Cer et al., "Universal Sentence Encoder," 2018, arXiv. doi: 10.48550/ARXIV.1803.11175.

[20] Y. Yang et al., "Multilingual Universal Sentence Encoder for Semantic Retrieval," 2019, arXiv. doi: 10.48550/ARXIV.1907.04307. https://doi.org/10.18653/v1/2020.acl-demos.12

[21] M. E. Peters et al., "Deep contextualized word representations," Mar. 22, 2018, arXiv: arXiv:1802.05365. Accessed: Nov. 06, 2024. [Online]. Available: http://arxiv.org/abs/1802.05365

[22] K. Ethayarajh, "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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China: Association for Computational Linguistics, 2019, pp. 55–65. doi: 10.18653/v1/D19-1006.
https://doi.org/10.18653/v1/D19-1006

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, arXiv: arXiv:1810.04805. Accessed: Sep. 23, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805

[24] K. Ethayarajh, "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings," Sep. 02, 2019, arXiv: arXiv:1909.00512. doi: 10.48550/arXiv.1909.00512.

[25] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," Sep. 10, 2019, arXiv: arXiv:1903.10676. Accessed: Sep. 24, 2024. [Online]. Available: http://arxiv.org/abs/1903.10676

[26] S. Ruder and B. Plank, "Strong Baselines for Neural Semi-Supervised Learning under Domain Shift," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1044–1054. doi: 10.18653/v1/P18-1096.
https://doi.org/10.18653/v1/P18-1096

[27] Z. wan, Y. Zhang, Y. Wang, F. Cheng, and S. Kurohashi, "Reformulating Domain Adaptation of Large Language Models as Adapt-Retrieve-Revise: A Case Study on Chinese Legal Domain," Aug. 26, 2024, arXiv: arXiv:2310.03328. doi: 10.48550/arXiv.2310.03328.
https://doi.org/10.18653/v1/2024.findings-acl.299

[28] S. Gururangan et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," May 05, 2020, arXiv: arXiv:2004.10964. doi: 10.48550/arXiv.2004.10964.
https://doi.org/10.18653/v1/2020.acl-main.740

[29] J. Howard and S. Ruder, "Universal Language Model Fine-tuning for Text Classification," May 23, 2018, arXiv: arXiv:1801.06146. doi: 10.48550/arXiv.1801.06146.
https://doi.org/10.18653/v1/P18-1031

[30] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
https://doi.org/10.1093/bioinformatics/btz682

[31] A. Akbik, T. Bergmann, and R. Vollgraf, "Pooled Contextualized Embeddings for Named Entity Recognition," in Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 724–728. doi: 10.18653/v1/N19-1078.
https://doi.org/10.18653/v1/N19-1078

[32] T. Zhang et al., "DKPLM: Decomposable Knowledge-Enhanced Pre-trained Language Model for Natural Language Understanding," Proc. AAAI Conf. Artif. Intell., vol. 36, no. 10, pp. 11703–11711, Jun. 2022, doi: 10.1609/aaai.v36i10.21425.
https://doi.org/10.1609/aaai.v36i10.21425

[33] M. E. Peters et al., "Knowledge Enhanced Contextual Word Representations," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China: Association for Computational Linguistics, 2019, pp. 43–54. doi: 10.18653/v1/D19-1005.
https://doi.org/10.18653/v1/D19-1005

[34] J. Yang, X. Hu, G. Xiao, and Y. Shen, "A Survey of Knowledge Enhanced Pre-trained Language Models," ACM Trans. Asian Low-Resour. Lang. Inf. Process., p. 3631392, Mar. 2024, doi: 10.1145/3631392.
https://doi.org/10.1145/3631392

[35] Y. Gu et al., "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," ACM Trans. Comput. Healthc., vol. 3, no. 1, pp. 1–23, Jan. 2022, doi: 10.1145/3458754.
https://doi.org/10.1145/3458754

[36] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," Sep. 10, 2021, arXiv: arXiv:2109.04607. doi: 10.48550/arXiv.2109.04607.
https://doi.org/10.18653/v1/2021.emnlp-main.833

[37] T. Kerner, "Domain-Specific Pretraining of Language Models: A Comparative Study in the Medical Field," Jul. 28, 2024, arXiv: arXiv:2407.14076. doi: 10.48550/arXiv.2407.14076.

[38] C. Ling et al., "Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey," Mar. 29, 2024, arXiv: arXiv:2305.18703. doi: 10.48550/arXiv.2305.18703.

[39] T. B. Brown et al., "Language Models are Few-Shot Learners," Jul. 22, 2020, arXiv: arXiv:2005.14165. Accessed: Aug. 28, 2024. [Online]. Available: http://arxiv.org/abs/2005.14165

[40] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Sep. 19, 2023, arXiv: arXiv:1910.10683. doi: 10.48550/arXiv.1910.10683.

[41] J. Kaplan et al., "Scaling Laws for Neural Language Models," Jan. 22, 2020, arXiv: arXiv:2001.08361. Accessed: Aug. 28, 2024. [Online]. Available: http://arxiv.org/abs/2001.08361

[42] N. Houlsby et al., "Parameter-Efficient Transfer Learning for NLP," Jun. 13, 2019, arXiv: arXiv:1902.00751. Accessed: Aug. 30, 2024. [Online]. Available: http://arxiv.org/abs/1902.00751

[43] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "AdapterFusion: Non-Destructive Task Composition for Transfer Learning," Jan. 26, 2021, arXiv: arXiv:2005.00247. Accessed: Aug. 30, 2024. [Online]. Available: http://arxiv.org/abs/2005.00247
https://doi.org/10.18653/v1/2021.eacl-main.39

[44] S. Hu et al., "Predicting Emergent Abilities with Infinite Resolution Evaluation," Apr. 17, 2024, arXiv: arXiv:2310.03262. Accessed: Aug. 28, 2024. [Online]. Available: http://arxiv.org/abs/2310.03262

[45] V. Lialin, V. Deshpande, and A. Rumshisky, "Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning," Mar. 27, 2023, arXiv: arXiv:2303.15647. Accessed: Jul. 03, 2024. [Online]. Available: http://arxiv.org/abs/2303.15647

[46] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," May 23, 2023, arXiv: arXiv:2305.14314. Accessed: Sep. 24, 2024. [Online]. Available: http://arxiv.org/abs/2305.14314

[47] C. Li and Z. Cao, "LoRa Networking Techniques for Large-scale and Long-term IoT: A Down-to-top Survey," ACM Comput. Surv., vol. 55, no. 3, pp. 1–36, Mar. 2023, doi: 10.1145/3494673.
https://doi.org/10.1145/3494673

[48] D. Kalajdzievski, "A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA," Nov. 28, 2023, arXiv: arXiv:2312.03732. Accessed: Nov. 12, 2024. [Online]. Available: http://arxiv.org/abs/2312.03732

[49] Q. Zhang et al., "AdaLoRA: Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning," Dec. 20, 2023, arXiv: arXiv:2303.10512. Accessed: Jul. 03, 2024. [Online]. Available: http://arxiv.org/abs/2303.10512

[50] K. P. V. Srinivasan, P. Gumpena, M. Yattapu, and V. H. Brahmbhatt, "Comparative Analysis of Different Efficient Fine Tuning Methods of Large Language Models (LLMs) in Low-Resource Setting," May 21, 2024, arXiv: arXiv:2405.13181. doi: 10.48550/arXiv.2405.13181.

[51] S.-Y. Liu et al., "DoRA: Weight-Decomposed Low-Rank Adaptation," Jul. 09, 2024, arXiv: arXiv:2402.09353. Accessed: Sep. 24, 2024. [Online]. Available: http://arxiv.org/abs/2402.09353

[52] A. Nie, C.-A. Cheng, A. Kolobov, and A. Swaminathan, "The Importance of Directional Feedback for LLM-based Optimizers," Jun. 20, 2024, arXiv: arXiv:2405.16434. doi: 10.48550/arXiv.2405.16434.

[53] V. Lialin, V. Deshpande, X. Yao, and A. Rumshisky, "Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning," Nov. 22, 2024, arXiv: arXiv:2303.15647. doi: 10.48550/arXiv.2303.15647.

[54] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, "Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment," Dec. 19, 2023, arXiv: arXiv:2312.12148. doi: 10.48550/arXiv.2312.12148.

[55] H. Cao, "Recent advances in text embedding: A Comprehensive Review of Top-Performing Methods on the MTEB Benchmark," Jun. 19, 2024, arXiv: arXiv:2406.01607. doi: 10.48550/arXiv.2406.01607.