improving text embeddings with large language models

improving text embeddings with large language models is a pivotal advancement in the field of natural language processing (NLP). Text embeddings are numerical representations of text that capture semantic meaning, enabling machines to understand and process language more effectively. Large language models (LLMs), with their extensive training on diverse datasets, have revolutionized the quality and applicability of these embeddings. This article explores how improving text embeddings with large language models enhances various NLP tasks, including search, classification, and recommendation systems. It delves into techniques used to optimize embeddings, the benefits of leveraging large-scale models, and practical considerations for implementation. The discussion also covers challenges and future prospects in embedding technology powered by LLMs. Following this introduction, a detailed table of contents outlines the key areas to be addressed.

Understanding Text Embeddings

Text embeddings are fundamental components in NLP that translate words, phrases, or entire documents into dense vector representations. These vectors capture syntactic and semantic information, enabling algorithms to perform tasks like similarity measurement, clustering, and classification. Traditional embedding methods, such as Word2Vec and GloVe, rely on co-occurrence statistics but often lack contextual understanding. Improving text embeddings with large language models introduces contextualized representations that dynamically adjust based on surrounding text, resulting in more accurate and meaningful embeddings.

Types of Text Embeddings

Various types of text embeddings exist, each with specific characteristics and use cases. Static embeddings assign a fixed vector to each word, regardless of context, while contextual embeddings vary depending on sentence structure and meaning. Large language models predominantly generate contextual embeddings, which better capture nuances in language.

- **Static Embeddings:** Generated by models like Word2Vec and GloVe; efficient but context-agnostic.

- **Contextual Embeddings**: Produced by models such as BERT and GPT; context-sensitive and dynamic.
- **Sentence and Document Embeddings**: Represent larger text units, often aggregated from word or token embeddings.

Importance of High-Quality Embeddings

The quality of text embeddings directly impacts the performance of downstream NLP applications. High-quality embeddings enable better semantic understanding, improving tasks such as information retrieval, sentiment analysis, and machine translation. Improving text embeddings with large language models leads to richer representations that capture subtle meanings and relationships within text data.

The Role of Large Language Models in Enhancing Embeddings

Large language models, trained on massive datasets with billions of parameters, have transformed the generation of text embeddings. These models leverage deep learning architectures, such as transformers, to understand context and semantic relationships at scale. Improving text embeddings with large language models results in vectors that reflect deeper linguistic knowledge and world understanding, surpassing traditional embedding methods.

Architecture of Large Language Models

Most state-of-the-art large language models utilize transformer architectures characterized by self-attention mechanisms. This design allows models to weigh the importance of different words in a sequence when generating embeddings, facilitating context-aware representations. The depth and size of these models enable learning complex patterns and dependencies in language data.

Contextualization and Dynamic Representation

One key advantage of large language models is their ability to produce contextual embeddings that vary depending on the sentence or paragraph. Unlike static embeddings, which assign a fixed vector to each word, LLMs consider the entire input context, leading to more accurate semantic representations. This dynamic nature is crucial for tasks requiring a nuanced understanding of language.

Techniques for Improving Text Embeddings with LLMs

Several techniques leverage large language models to improve text embeddings, enhancing their semantic richness and applicability. These methods optimize the embedding generation process, ensuring that the resulting vectors better capture meaning and relevance.

Fine-Tuning Pretrained Models

Fine-tuning involves adapting a pretrained large language model to a specific domain or task by continuing its training on relevant datasets. This technique refines embeddings so that they better reflect specialized vocabulary and context, improving performance in niche applications.

Prompt Engineering for Embedding Generation

Prompt engineering strategically crafts input text to guide large language models in producing more informative embeddings. By designing prompts that emphasize certain aspects of the text, embeddings can be tailored to highlight the desired semantic features, enhancing downstream task efficiency.

Multi-Task and Contrastive Learning Approaches

Multi-task learning trains models on several related tasks simultaneously, which helps produce embeddings that generalize better across applications. Contrastive learning techniques, such as SimCSE, improve embeddings by encouraging similar texts to have closer vectors and dissimilar texts to be farther apart in the embedding space.

Utilizing Sentence and Document-Level Embeddings

Improving text embeddings with large language models also involves generating embeddings for longer text units beyond single words. Techniques like mean pooling, max pooling, or specialized sentence transformers aggregate token embeddings to create meaningful sentence or document vectors suitable for complex NLP tasks.

Applications and Benefits of Enhanced Text Embeddings

The advancements in text embeddings driven by large language models have broad applications across industries and NLP tasks. Improving text embeddings with large language models enables more accurate, efficient, and scalable solutions in multiple domains.

Information Retrieval and Search Engines

Improved embeddings allow search systems to better understand query intent and content semantics, leading to more relevant and precise results. Semantic search benefits significantly from contextual embeddings, which interpret user queries beyond keyword matching.

Natural Language Understanding and Generation

Enhanced embeddings facilitate better comprehension of input text in tasks like sentiment analysis, summarization, and question answering. They also contribute to more coherent and contextually appropriate text generation by language models.

Recommendation Systems and Personalization

In recommendation engines, embeddings represent user preferences and item attributes, enabling personalized suggestions. Large language models improve these embeddings by capturing subtle semantic relationships and user intent more effectively.

Benefits of Improved Text Embeddings with LLMs

- **Higher Accuracy**: More precise semantic understanding improves task outcomes.
- **Context Awareness**: Captures nuanced language features and polysemy.
- **Domain Adaptability**: Fine-tuning enables specialization for diverse industries.
- **Scalability**: Efficient embeddings support large-scale applications.
- **Robustness**: Better handling of noisy or ambiguous text inputs.

Challenges and Future Directions

Despite significant progress, improving text embeddings with large language models presents challenges that require ongoing research and innovation. Addressing these obstacles is critical to unlocking the full potential of embedding technologies.

Computational Complexity and Resource Requirements

Large language models demand substantial computational power and memory, making embedding generation costly and less accessible for some applications. Optimizing models and developing efficient inference techniques are active areas of research to mitigate these constraints.

Bias and Fairness in Embeddings

Embeddings generated by large language models can inherit biases present in training data, potentially leading to unfair or discriminatory outcomes. Techniques to detect, quantify, and reduce bias in embeddings are essential for ethical NLP deployment.

Interpretability and Explainability

Understanding why certain embeddings represent text in specific ways remains a challenge. Improving interpretability helps build trust and facilitates debugging and refinement of NLP systems.

Future Prospects

Emerging techniques such as few-shot learning, continual learning, and multimodal embeddings promise to further enhance text representation quality. Integration of knowledge graphs and symbolic reasoning with LLM-generated embeddings may also expand capabilities, enabling more sophisticated language understanding and reasoning.

Frequently Asked Questions

What are text embeddings and why are they important in NLP?

Text embeddings are numerical vector representations of text that capture semantic meaning, enabling machines to understand and process language. They are important because they allow algorithms to perform tasks like similarity comparison, clustering, and classification effectively.

How do large language models improve the quality of text embeddings?

Large language models improve text embeddings by leveraging vast amounts of training data and deep architectures to capture complex semantic and syntactic relationships, resulting in richer and more context-aware vector representations.

What techniques are commonly used to generate embeddings with large language models?

Common techniques include using pretrained transformer-based models like BERT, GPT, or RoBERTa to extract embeddings from specific layers, fine-tuning these models on domain-specific data, and employing methods like sentence-transformers to produce sentence-level embeddings.

Can fine-tuning large language models improve embeddings for specific tasks?

Yes, fine-tuning large language models on task-specific or domain-specific datasets helps the embeddings capture relevant nuances and improves performance on specialized tasks such as sentiment analysis or information retrieval.

How do contextual embeddings differ from traditional static embeddings?

Contextual embeddings generated by large language models dynamically change based on the surrounding text, capturing word sense and meaning in context, whereas traditional static embeddings like Word2Vec assign a single fixed vector per word regardless of context.

What challenges exist when using large language models for improving text embeddings?

Challenges include high computational costs, large memory requirements, potential biases in pretrained models, and the need for large amounts of labeled data for effective fine-tuning or adaptation to specific domains.

How can improved text embeddings benefit real-world applications?

Improved text embeddings enable more accurate search and recommendation systems, better sentiment and intent analysis, enhanced machine translation, and more effective chatbots and virtual assistants by providing deeper understanding of user inputs and textual data.

Additional Resources

1. *Enhancing Text Embeddings with Large Language Models*
This book explores the foundations and advanced techniques of generating high-quality text embeddings using large language models (LLMs). It covers the theoretical underpinnings, practical algorithms, and real-world applications. Readers will learn how to leverage LLMs to improve semantic understanding and downstream NLP tasks effectively.

2. *Deep Learning for Text Representation: From Word Embeddings to Large Language Models*
Focusing on the evolution of text representation, this book traces the journey from traditional word embeddings to cutting-edge LLM-based embeddings. It provides comprehensive insights into architectures like transformers and their role in enhancing semantic vector spaces. Practical examples and code snippets help readers implement these methods in their projects.

3. *Optimizing Large Language Models for Semantic Embedding Generation*
This work dives into optimization strategies for large language models aimed at producing superior text embeddings. It discusses fine-tuning, prompt engineering, and transfer learning techniques that maximize embedding quality. The book also presents case studies demonstrating significant improvements in information retrieval and recommendation systems.

4. *Text Embeddings in the Era of Large Language Models*
Covering the latest advances, this book presents how large language models have revolutionized text embeddings. It examines different embedding techniques, including contextual and static embeddings, and their comparative effectiveness. Researchers and practitioners will find guidance on selecting and adapting embeddings for diverse NLP applications.

5. *Practical Guide to Building Text Embeddings with Transformers*
A hands-on manual for developers and data scientists, this guide focuses on using transformer-based LLMs to create robust text embeddings. It includes step-by-step tutorials, coding examples, and best practices for embedding extraction and evaluation. The book also addresses challenges like scalability and model interpretability.

6. *Advanced Techniques in Text Embedding with Large Language Models*
Delving deeper into sophisticated methods, this book covers techniques such as embedding fusion, dimensionality reduction, and contrastive learning with LLMs. It highlights recent research trends and experimental results that push the boundaries of semantic representation. The content is suited for advanced practitioners aiming to innovate in NLP.

7. *Large Language Models and Their Impact on Text Embedding Quality*
This book analyzes the transformative impact of large language models on the quality and applicability of text embeddings. It presents comparative studies, benchmarks, and performance analyses across various LLM architectures. Readers gain an understanding of how to harness these models for enhanced semantic similarity and clustering tasks.

8. *From Bag-of-Words to Contextual Embeddings: Leveraging LLMs for Text Understanding*
Tracing the historical progression of text embeddings, this book emphasizes the shift from simple bag-of-words models to rich, contextual embeddings powered by LLMs. It explains the conceptual differences and demonstrates practical improvements in tasks like sentiment analysis and question answering. The narrative provides a solid foundation for newcomers and experienced NLP practitioners alike.

9. *Building Scalable Text Embedding Pipelines with Large Language Models*
This book addresses the engineering challenges of deploying large language model-based text embedding systems at scale. It covers topics such as distributed computing, efficient indexing, and real-time embedding generation. Readers will learn how to build robust pipelines that support large-scale applications in search engines and recommendation platforms.

Improving Text Embeddings With Large Language Models

improving text embeddings with large language models: Computer Vision – ECCV 2024 Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, Gül Varol, 2024-10-27 The multi-volume set of LNCS books with volume numbers 15059 up to 15147 constitutes the refereed proceedings of the 18th European Conference on Computer Vision, ECCV 2024, held in Milan, Italy, during September 29–October 4, 2024. The 2387 papers presented in these proceedings were carefully reviewed and selected from a total of 8585 submissions. They deal with topics such as computer vision; machine learning; deep neural networks; reinforcement learning; object recognition; image classification; image processing; object detection; semantic segmentation; human pose estimation; 3d reconstruction; stereo vision; computational photography; neural networks; image coding; image reconstruction; motion estimation.

improving text embeddings with large language models: Neural Information Processing Mufti Mahmud, Maryam Doborjeh, Kevin Wong, Andrew Chi Sing Leung, Zohreh Doborjeh, M. Tanveer, 2025-07-15 The sixteen-volume set, CCIS 2282-2297, constitutes the refereed proceedings of the 31st International Conference on Neural Information Processing, ICONIP 2024, held in Auckland, New Zealand, in December 2024. The 472 regular papers presented in this proceedings set were carefully reviewed and selected from 1301 submissions. These papers primarily focus on the following areas: Theory and algorithms; Cognitive neurosciences; Human-centered computing; and Applications.

improving text embeddings with large language models: Advances in Information Retrieval Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, Nicola Tonellotto, 2025-04-03 The five-volume set LNCS 15572, 15573, 15574, 15575 and 15576 constitutes the refereed conference proceedings of the 47th European Conference on Information Retrieval, ECIR 2025, held in Lucca, Italy, during April 6–10, 2025. The 52 full papers, 11 findings, 42 short papers and 76 papers of other types presented in these proceedings were carefully reviewed and selected from 530 submissions. The accepted papers cover the state-of-the-art in information retrieval and recommender systems: user aspects, system and foundational aspects, artificial intelligence and machine learning, applications, evaluation, new social and technical challenges, and other topics of direct or indirect relevance to search and recommendation.

improving text embeddings with large language models: New Trends in Theory and Practice of Digital Libraries Wolf-Tilo Balke, Koraljka Golub, Yannis Manolopoulos, Kostas Stefanidis, Zheying Zhang, Trond Aalberg, Paolo Manghi, 2025-10-28 This book constitutes the proceedings of the workshops held at the 29th International Conference on Theory and Practice of Digital Libraries, TPDL 2025, which took place in Tampere, Finland, during September 23-26, 2025. The 20 short papers, 8 Demo papers and 9 workshop papers included in this book were carefully reviewed and selected from 103 paper submissions (52 full papers, 40 short papers and 11demos). TPDL has established itself as an important international forum focused on digital libraries and associated technical, practical, and social issues.

improving text embeddings with large language models: Artificial Intelligence XLI Max Bramer, Frederic Stahl, 2024-11-28 This two-volume set, LNAI 15446 and LNAI 15447, constitutes the refereed proceedings of the 44th SGAI International Conference on Artificial Intelligence, AI 2024, held in Cambridge, UK, during December 17–19, 2024. The 36 full papers and 18 short papers presented in these two volumes were carefully reviewed and selected from 80 submissions. Part I includes papers from the Technical stream, whereas Part II includes papers from the Application stream. These volumes are organized into the following topical sections: - Part I: Neural nets; Deep learning; Large language models; Machine learning; Evolutionary and genetic algorithms; Knowledge management, Short Technical Papers. Part II: Machine vision; Evaluation of AI systems; Applications of machine learning; Other AI applications, Short Application Papers.

improving text embeddings with large language models: Advanced Data Mining and Applications Quan Z. Sheng, Gill Dobbie, Jing Jiang, Xuyun Zhang, Wei Emma Zhang, Yannis Manolopoulos, Jia Wu, Wathiq Mansoor, Congbo Ma, 2024-12-13 This six-volume set, LNAI 15387-15392, constitutes the refereed proceedings of the 20th International Conference on Advanced Data Mining and Applications, ADMA 2024, held in Sydney, New South Wales, Australia, during December 3–5, 2024. The 159 full papers presented here were carefully reviewed and selected from 422 submissions. These papers have been organized under the following topical sections across the different volumes: - Part I : Applications; Data mining. Part II : Data mining foundations and algorithms; Federated learning; Knowledge graph. Part III : Graph mining; Spatial data mining. Part IV : Health informatics. Part V : Multi-modal; Natural language processing. Part VI : Recommendation systems; Security and privacy issues.

improving text embeddings with large language models: Linking Theory and Practice of Digital Libraries Apostolos Antonacopoulos, Annika Hinze, Benjamin Piwowarski, Mickaël Coustaty, Giorgio Maria Di Nunzio, Francesco Gelati, Nicholas Vanderschantz, 2024-09-24 This book constitutes the refereed proceedings of the 28th International Conference on Linking Theory and Practice of Digital Libraries, TPDL 2024, held in Ljubljana, Slovenia, during September 24–27. The 13 full papers, 19 short papers and 11 papers of other types included in this book were carefully reviewed and selected from 83 submissions. Over the years, TPDL has established itself as an important international forum focused on digital libraries and associated technical, practical, and social issues. In 2024, TPDL expanded its scope to prominently include Document Analysis/Recognition and Information Retrieval, acknowledging the vital role of those research areas in the creation (by means of digitization and information extraction from heterogeneous sources), access, discovery, and dissemination of digital content.

improving text embeddings with large language models: *Intelligent Computing* Kohei Arai, 2025-08-13 This book compiles a curated selection of insightful, rigorously researched, and state-of-the-art papers presented at the Computing Conference 2025, hosted in London, UK, on June 19–20, 2025. Drawing submissions from across the globe, the conference received 473 papers, each subjected to a stringent double-blind peer-review process. Of these, 169 papers were accepted for inclusion, reflecting exceptional scholarship and innovation across disciplines such as IoT, artificial intelligence, computing, data science, networking, data security, and privacy. Researchers, academics, and industry leaders converged to share pioneering ideas, transformative methodologies, and practical solutions to real-world challenges. By bridging academic theory and industrial application, the conference catalyzed opportunities for knowledge synthesis and interdisciplinary progress. The diverse contributions within this proceedings not only address contemporary technological issues but also anticipate future trends, offering frameworks for continued exploration. We trust this collection will serve as an indispensable reference for researchers, practitioners, and policymakers navigating the evolving landscapes of computing and digital innovation. As we reflect on the conference’s outcomes, we are confident that the insights and collaborations forged here will inspire sustained advancements in these critical fields. May the ideas within these pages spark further inquiry, drive technological evolution, and contribute meaningfully to solving the challenges of our interconnected world.

improving text embeddings with large language models: *Linking Theory and Practice of Digital Libraries* Wolf-Tilo Balke, Koraljka Golub, Yannis Manolopoulos, Kostas Stefanidis, Zheying Zhang, 2025-09-22 This book constitutes the refereed proceedings of the 29th International Conference on Theory and Practice of Digital Libraries on Linking Theory and Practice of Digital Libraries, TPDL 2025, held in Tampere, Finland, during September 23–26, 2025. The 14 full papers and 11 finding papers included in this book were carefully reviewed and selected from 103 submissions. The papers are organized in the following topical: Keynotes; Large Language Models; Scholarly Issues; Citation Management; Digital Archives; Digital Humanities and Cultural Heritage; Knowledge Graphs.

improving text embeddings with large language models: World Conference of AI-Powered Innovation and Inventive Design Denis Cavallucci, Stelian Brad, Pavel Livotov, 2024-10-28 This book constitutes the proceedings of the 24th IFIP WG 5.4 International TRIZ Future Conference on AI-Powered Innovation and Inventive Design, TFC 2024, held in Cluj-Napoca, Romania, during November 6–8, 2024. The 42 full papers presented were carefully reviewed and selected from 72 submissions. They were organized in the following topical sections: Part I - AI-Driven TRIZ and Innovation Part II - Sustainable and Industrial Design with TRIZ; Digital Transformation, Industry 4.0, and Predictive Analytics; Interdisciplinary and Cognitive Approaches in TRIZ; Customer Experience and Service Innovation with TRIZ.

improving text embeddings with large language models: Information and Communication Technology Wray Buntine, Morten Fjeld, Truyen Tran, Minh-Triet Tran, Binh Huynh Thi Thanh, Takumi Miyoshi, 2025-04-25 This four-volume set, CCIS 2350-2353, constitutes the referred proceedings of the 13th International Symposium on Information and Communication Technology, SOICT 2024, held in Danang, Vietnam in December 2024. The 88 full papers and 68 poster papers presented here were carefully reviewed and selected from 229 submissions. The papers presented in these volumes are organized in the following topical sections: Part I: Multimedia Processing; Operations Research. Part II: AI Applications; Cyber Security. Part III: AI Foundations and Big Data; Human-Computer Interaction. Part IV: Lifelog and Multimedia Retrieval; Generative AI; Software Engineering.

improving text embeddings with large language models: Breaking Barriers with Generative Intelligence. Using GI to Improve Human Education and Well-Being Azza Basiouni, Claude Frasson, 2024-07-25 The book constitutes the proceedings for the First International Conference on Breaking Barriers with Generative Intelligence, BBGI 2024, held in Thessaloniki, Greece, on June 10, 2024. This Workshop is part of the 20th International Conference on Intelligent Tutoring Systems (ITS2024) which was held in Thessaloniki, from June 10 to June 13, 2024. The 19 full papers and 1 short paper included in this volume were carefully reviewed and selected from a total of 24 submissions. Breaking Barriers with Generative Intelligence delves into how GI in AI improves human education and well-being. This interdisciplinary event brought together professionals from academia, industry, and government to address AI ethics, human-AI interaction, and the societal implications of GI. Participants learned to tackle social concerns and promote diversity in research and development through keynote presentations, panel discussions, and interactive workshops.

improving text embeddings with large language models: Building Generative AI Applications with Open-source Libraries Srikannan Balakrishnan, 2025-03-27 Generative AI is revolutionizing how we interact with technology, empowering us to create everything from compelling text to intricate code. This book is your practical guide to harnessing the power of open-source libraries, enabling you to build cutting-edge generative AI applications without needing extensive prior experience. In this book, you will journey from foundational concepts like natural language processing and transformers to the practical implementation of large language models. Learn to customize foundational models for specific industries, master text embeddings, and vector databases for efficient information retrieval, and build robust applications using LangChain. Explore open-source models like Llama and Falcon and leverage Hugging Face for seamless implementation. Discover how to deploy scalable AI solutions in the cloud while also understanding crucial aspects of data privacy and ethical AI usage. By the end of this book, you will be equipped with technical skills and practical knowledge, enabling you to confidently develop and deploy your own generative AI applications, leveraging the power of open-source tools to innovate and create. WHAT YOU WILL LEARN ● Building AI applications using LangChain and integrating RAG. ● Implementing large language models like Llama and Falcon. ● Utilizing Hugging Face for efficient model deployment. ● Developing scalable AI applications in cloud environments. ● Addressing ethical considerations and data privacy in AI. ● Practical application of vector databases for information retrieval. WHO THIS BOOK IS FOR This book is for aspiring tech professionals, students, and creative minds seeking to build generative AI applications. While a basic understanding of programming and an interest in AI are beneficial, no prior generative AI expertise is required. TABLE OF CONTENTS 1. Getting Started with Generative AI 2. Overview of Foundational Models 3. Text Processing and Embeddings Fundamentals 4. Understanding Vector Databases 5. Exploring LangChain for Generative AI 6. Implementation of LLMs 7. Implementation Using Hugging Face 8. Developments in Generative AI 9. Deployment of Applications 10. Generative AI for Good

improving text embeddings with large language models: Hands-On Large Language Models Jay Alammar, Maarten Grootendorst, 2024-09-11 AI has acquired startling new language capabilities in just the past few years. Driven by the rapid advances in deep learning, language AI systems are able to write and understand text better than ever before. This trend enables the rise of new features, products, and entire industries. With this book, Python developers will learn the practical tools and concepts they need to use these capabilities today. You'll learn how to use the power of pre-trained large language models for use cases like copywriting and summarization; create semantic search systems that go beyond keyword matching; build systems that classify and cluster text to enable scalable understanding of large amounts of text documents; and use existing libraries and pre-trained models for text classification, search, and clusterings. This book also shows you how to: Build advanced LLM pipelines to cluster text documents and explore the topics they belong to Build semantic search engines that go beyond keyword search with methods like dense retrieval and rerankers Learn various use cases where these models can provide value Understand the architecture of underlying Transformer models like BERT and GPT Get a deeper understanding of how LLMs are trained Understanding how different methods of fine-tuning optimize LLMs for specific applications (generative model fine-tuning, contrastive fine-tuning, in-context learning, etc.)

improving text embeddings with large language models: AI Frameworks Enabled by Blockchain Vikram Dhillon, David Metcalf, Max Hooper, 2025-07-17 Blockchain technology offers a powerful foundation for building trust, privacy and verifiability into AI frameworks. This book will focus on how a blockchain can enable AI frameworks and applications to scale in a responsible fashion, reshaping the future of numerous industries from financial markets to healthcare and education. You'll see that in the next wave of AI products, blockchain can provide a "Trust Layer," a fundamental feature previously only implemented for parties within a blockchain network. The provable consensus algorithms and oracles previously implemented in blockchains can be extended to autonomous agents that are integrated with large language models (LLMs) and future applications. Finally, you'll learn that safety is a major concern for practical applications of AI and blockchain can help mitigate threats due to the decentralized nature. As such, there will be significant discourse on how blockchain can provide enhanced security against prompt injections, LLM-hijacking for dangerous information and privacy. These ideas were studied rigorously when large financial institutions were releasing their own blockchains and distributed ledger protocols with a heavy focus privacy. AI is undergoing a Cambrian explosion this year with foundational models emerging for all major domains of study, however, most such models lack the capacity to externally validate for the "correctness" of a fact, or reply made by the LLM. Similarly, there are no definitive methods to distinguish between meaningful insights and hallucination. These challenges remain at the forefront of AI research, and AI Frameworks Enabled by Blockchain aims to translate technical literature into actionable and practical tips for the AI domain. What You Will Learn !-- [if !supportLists]--· !--[endif]--Bring a layer of accuracy to generative AI where a non-generative component behaves as guardrails !-- [if !supportLists]--· !--[endif]--Protect users from harmful biases as well as hallucinations. !-- [if !supportLists]--· !--[endif]--See how blockchain plays a role in aligning AI with human interests. !-- [if !supportLists]--· !--[endif]--Review use-cases and real-world applications from parties that have invested a significant amount in building technology stacks utilizing both. Who This Book Is For Enterprise users and policy makers in the field of Professional and Applied Computing

improving text embeddings with large language models: Proceedings of International Conference on Theoretical and Applied Computing Lisa Mathew, K. G. Subramanian, Atulya K. Nagar, 2025-02-13 This book presents research papers presented at the International Conference on Theoretical and Applied Computing 2023, held during September 13–15, 2023. ICTAC 2023 is organized by Amal Jyothi College of Engineering, India. This book covers topics, such as theoretical foundations of computing, algorithms and data structures, computer systems and architecture, computer networks and communications, graph theory, algorithms and complexity, quantum computation theory, computational geometry, software engineering and programming languages, human-computer interaction, artificial intelligence and machine learning, data mining and knowledge discovery, parallel and distributed computing, grid and cloud computing, bioinformatics/biomedical applications, data mining, evolutionary computation, fuzzy logic, genetic algorithms, natural language processing and image processing.

improving text embeddings with large language models: Knowledge Science, Engineering and Management Zhi Jin, Yuncheng Jiang, Robert Andrei Buchmann, Yaxin Bi, Ana-Maria Ghiran, Wenjun Ma, 2023-08-08 This volume set constitutes the refereed proceedings of the 16th International Conference on Knowledge Science, Engineering and Management, KSEM 2023, which was held in Guangzhou, China, during August 16–18, 2023. The 114 full papers and 30 short papers included in this book were carefully reviewed and selected from 395 submissions. They were organized in topical sections as follows: knowledge science with learning and AI; knowledge engineering research and applications; knowledge management systems; and emerging technologies for knowledge science, engineering and management.

improving text embeddings with large language models: *Pattern Recognition* Björn Andres, Florian Bernard, Daniel Cremers, Simone Frintrop, Bastian Goldlücke, Ivo Ihrke, 2022-09-22 This book constitutes the refereed proceedings of the 44th DAGM German Conference on Pattern Recognition, DAGM GCPR 2022, which was held during September 27 – 30, 2022. The 37 papers presented in this volume were carefully reviewed and selected from 78 submissions. They were organized in topical sections as follows: machine learning methods; unsupervised, semi-supervised and transfer learning; interpretable machine learning; low-level vision and computational photography; motion, pose estimation and tracking; 3D vision and stereo; detection and recognition; language and vision; scene understanding; photogrammetry and remote sensing; pattern recognition in the life and natural sciences; systems and applications.

improving text embeddings with large language models: *Intelligent and Fuzzy Systems* Cengiz Kahraman, Selcuk Cebi, Basar Oztaysi, Sezi Cevik Onar, Cagrı Tolga, Irem Ucal Sari, Irem Otay, 2025-07-25 Artificial Intelligence in Human-Centric, Resilient & Sustainable Industries This book focuses on benefiting artificial intelligent tools in our business and social life under emerging conditions. Human-centric, resilient, and sustainable industries are built on ideals like human-centricity, ecological advantages, or social benefits. The mission of human-centric artificial intelligence is to improve people’s lives by offering solutions that boost productivity, accessibility to resources, security, well-being, and general quality of life. The latest intelligent methods and techniques on human-centric, resilient, and sustainable industries are introduced by theory and applications. This book covers the chapters of world-wide known experts on machine learning, medical image processing, process intelligence, process mining, and others. The intended readers are intelligent systems researchers, lecturers, M.Sc. and Ph.D. students trying to develop approaches giving human needs, values, and viewpoints top priority through artificial intelligent systems.

improving text embeddings with large language models: *Building Neo4j-Powered Applications with LLMs* Ravindranatha Anthapu, Siddhant Agarwal, 2025-06-20 A comprehensive guide to building cutting-edge generative AI applications using Neo4j's knowledge graphs and vector search capabilities Key Features Design vector search and recommendation systems with LLMs using Neo4j GenAI, Haystack, Spring AI, and LangChain4j Apply best practices for graph exploration, modeling, reasoning, and performance optimization Build and consume Neo4j knowledge graphs and deploy your GenAI apps to Google Cloud Purchase of the print or Kindle book includes a free PDF eBook Book DescriptionEmbark on an expert-led journey into building LLM-powered applications using Retrieval-Augmented Generation (RAG) and Neo4j knowledge graphs. Written by Ravindranatha Anthapu, Principal Consultant at Neo4j, and Siddhant Agrawal, a Google Developer Expert in GenAI, this comprehensive guide is your starting point for exploring alternatives to LangChain, covering frameworks such as Haystack, Spring AI, and LangChain4j. As LLMs (large language models) reshape how businesses interact with customers, this book helps you develop intelligent applications using RAG architecture and knowledge graphs, with a strong focus on overcoming one of AI’s most persistent challenges—mitigating hallucinations. You'll learn how to model and construct Neo4j knowledge graphs with Cypher to enhance the accuracy and relevance of LLM responses. Through real-world use cases like vector-powered search and personalized recommendations, the authors help you build hands-on experience with Neo4j GenAI integrations across Haystack and Spring AI. With access to a companion GitHub repository, you’ll work through code-heavy examples to confidently build and deploy GenAI apps on Google Cloud. By the end of this book, you’ll have the skills to ground LLMs with RAG and Neo4j, optimize graph performance, and strategically select the right cloud platform for your GenAI applications.What you will learn Design, populate, and integrate a Neo4j knowledge graph with RAG Model data for knowledge graphs Integrate AI-powered search to enhance knowledge exploration Maintain and monitor your AI search application with Haystack Use LangChain4j and Spring AI for recommendations and personalization Seamlessly deploy your applications to Google Cloud Platform Who this book is for This LLM book is for database developers and data scientists who want to leverage knowledge graphs with Neo4j and its vector search capabilities to build intelligent search and recommendation systems. Working knowledge of Python and Java is essential to follow along. Familiarity with Neo4j, the Cypher query language, and fundamental concepts of databases will come in handy.

Related to improving text embeddings with large language models

Improving Text Embeddings with Large Language Models In this paper, we introduce a novel and simple method for obtaining high-quality text embeddings using only synthetic data and less than 1k training steps

Improving Text Embeddings with Large Language Models This paper shows that the quality of text embeddings can be substantially enhanced by exploiting LLMs. We prompt proprietary LLMs such as GPT-4 to generate diverse synthetic

The State of Embedding Technologies for Large Language Models We are observing cutting-edge, transformer-derived embeddings, supercharged by LLM pre-training and innovative contrastive learning methodologies, delivering unparalleled

Embedding Models Explained, How To Use Them & 10 Tools In short, embedding models aren’t just about text—they’ve become a universal way to represent all kinds of data. Whether you’re working with documents, images, or even

Improving Text Representations with Large Language Models Recent advances in large language models (LLMs) have significantly improved the qual-ity of text representations, enabling break-throughs in dense retrieval, semantic search, and a range of

LLM-Enhanced Semantic Text Segmentation - MDPI 6 days ago This study investigates semantic text segmentation enhanced by large language model (LLM) embeddings. We assess how effectively embeddings capture semantic

Understanding Amazon Titan: Large Language Models for AWS Understand Amazon Titan LLMs on AWS Bedrock: enterprise-ready models for text, embeddings, image AI, semantic search, and responsible generative AI workflows

Language model - Wikipedia A language model is a model of the human brain's ability to produce natural language. [1][2] Language models are useful for a variety of tasks, including speech recognition, [3] machine

Improving Text Embeddings with Large Language Models We posit that generative language modeling and text embeddings are the two sides of the same coin, with both tasks requiring the model to have a deep understanding of

The widespread adoption of large language model-assisted Large language models (LLMs), such as ChatGPT, represent a transformative development in artificial intelligence (AI), significantly impacting how individuals, businesses,

Improving Text Embeddings with Large Language Models In this paper, we introduce a novel and simple method for obtaining high-quality text embeddings using only synthetic data and less than 1k training steps

Improving Text Embeddings with Large Language Models This paper shows that the quality of text embeddings can be substantially enhanced by exploiting LLMs. We prompt proprietary LLMs such as GPT-4 to generate diverse synthetic

The State of Embedding Technologies for Large Language Models We are observing cutting-

edge, transformer-derived embeddings, supercharged by LLM pre-training and innovative contrastive learning methodologies, delivering unparalleled

Embedding Models Explained, How To Use Them & 10 Tools In short, embedding models aren’t just about text—they’ve become a universal way to represent all kinds of data. Whether you’re working with documents, images, or even

Improving Text Representations with Large Language Models Recent advances in large language models (LLMs) have significantly improved the qual-ity of text representations, enabling break-throughs in dense retrieval, semantic search, and a range of

LLM-Enhanced Semantic Text Segmentation - MDPI 6 days ago This study investigates semantic text segmentation enhanced by large language model (LLM) embeddings. We assess how effectively embeddings capture semantic

Understanding Amazon Titan: Large Language Models for AWS Understand Amazon Titan LLMs on AWS Bedrock: enterprise-ready models for text, embeddings, image AI, semantic search, and responsible generative AI workflows

Language model - Wikipedia A language model is a model of the human brain's ability to produce natural language. [1][2] Language models are useful for a variety of tasks, including speech recognition, [3] machine

Improving Text Embeddings with Large Language Models We posit that generative language modeling and text embeddings are the two sides of the same coin, with both tasks requiring the model to have a deep understanding of

The widespread adoption of large language model-assisted writing Large language models (LLMs), such as ChatGPT, represent a transformative development in artificial intelligence (AI), significantly impacting how individuals, businesses,

Improving Text Embeddings with Large Language Models In this paper, we introduce a novel and simple method for obtaining high-quality text embeddings using only synthetic data and less than 1k training steps

Improving Text Embeddings with Large Language Models This paper shows that the quality of text embeddings can be substantially enhanced by exploiting LLMs. We prompt proprietary LLMs such as GPT-4 to generate diverse synthetic

The State of Embedding Technologies for Large Language Models We are observing cutting-edge, transformer-derived embeddings, supercharged by LLM pre-training and innovative contrastive learning methodologies, delivering unparalleled

Embedding Models Explained, How To Use Them & 10 Tools In short, embedding models aren’t just about text—they’ve become a universal way to represent all kinds of data. Whether you’re working with documents, images, or even

Improving Text Representations with Large Language Models Recent advances in large language models (LLMs) have significantly improved the qual-ity of text representations, enabling break-throughs in dense retrieval, semantic search, and a range of

LLM-Enhanced Semantic Text Segmentation - MDPI 6 days ago This study investigates semantic text segmentation enhanced by large language model (LLM) embeddings. We assess how effectively embeddings capture semantic

Understanding Amazon Titan: Large Language Models for AWS Understand Amazon Titan LLMs on AWS Bedrock: enterprise-ready models for text, embeddings, image AI, semantic search, and responsible generative AI workflows

Language model - Wikipedia A language model is a model of the human brain's ability to produce natural language. [1][2] Language models are useful for a variety of tasks, including speech recognition, [3] machine

Improving Text Embeddings with Large Language Models We posit that generative language modeling and text embeddings are the two sides of the same coin, with both tasks requiring the model to have a deep understanding of

The widespread adoption of large language model-assisted writing Large language models

(LLMs), such as ChatGPT, represent a transformative development in artificial intelligence (AI), significantly impacting how individuals, businesses,

Related to improving text embeddings with large language models

Self-improving language models are becoming reality with MIT's updated SEAL technique (18h) Researchers at the Massachusetts Institute of Technology (MIT) are gaining renewed attention for developing and open sourcing a technique that allows large language models (LLMs) — like those

Self-improving language models are becoming reality with MIT's updated SEAL technique (18h) Researchers at the Massachusetts Institute of Technology (MIT) are gaining renewed attention for developing and open sourcing a technique that allows large language models (LLMs) — like those

Large Language Models Rival Humans in Learning Logical Rules, New Study Finds (The Debrief4d) New research shows large language models rival humans in learning logic-based rules, reshaping how we understand reasoning

Large Language Models Rival Humans in Learning Logical Rules, New Study Finds (The Debrief4d) New research shows large language models rival humans in learning logic-based rules, reshaping how we understand reasoning