Text Embeddings Should Capture Implicit Semantics, Not Just Surface Meaning

Anonymous Author(s)
Affiliation
Address
email

Abstract

This position paper argues that the text embedding research community should move beyond surface meaning and embrace implicit semantics as a central modeling goal. Text embedding models have become foundational in modern NLP, powering a wide range of applications and drawing increasing research attention. Yet, much of this progress remains narrowly focused on surface-level semantics. In contrast, linguistic theory emphasizes that meaning is often implicit, shaped by pragmatics, speaker intent, and sociocultural context. Current embedding models are typically trained on data that lacks such depth and evaluated on benchmarks that reward the capture of surface meaning. As a result, they struggle with tasks requiring interpretive reasoning, speaker stance, or social meaning. Our pilot study highlights this gap, showing that even state-of-the-art models perform only marginally better than simplistic baselines on implicit semantics tasks. To address this, we call for a paradigm shift: embedding research should prioritize more diverse and linguistically grounded training data, design benchmarks that evaluate deeper semantic understanding, and explicitly frame implicit meaning as a core modeling objective, better aligning embeddings with real-world language complexity.

1 Introduction

Text embedding models are designed to transform textual content, whether sentences, paragraphs, or full documents, into dense vectors in a high-dimensional space, where the proximity of embeddings reflects semantic similarity [123, 102]. These models have become foundational in modern NLP and are now widely deployed in a pre-trained, off-the-shelf manner across a wide range of downstream tasks such as clustering [41, 8], classification [102], information retrieval [148, 57], and retrieval-augmented generation (RAG) [76]. In response, the research community has dedicated extensive effort to improving model architectures [123, 86, 12, 101], training strategies [38, 150, 82, 163], and evaluation benchmarks [102, 44, 34].

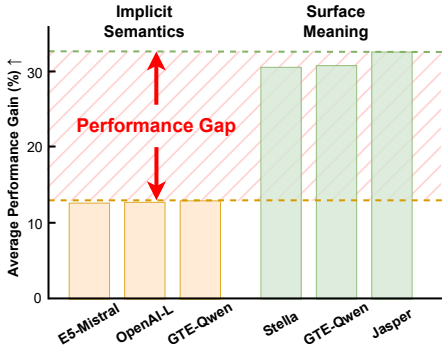

Figure 1: Average performance gains of top embedding models over the Bag-of-Tokens baseline on two evaluation sets: implicit semantics (averaged over seven datasets from Table 1) and surface meaning (averaged over MTEB classification tasks [102]).

The Overlooked Implicit Semantics Despite substantial advances in text embedding research, a critical gap remains: most embedding models are designed to capture surface-level semantics, such

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

as lexical overlap, syntactic variation, and topical similarity, while largely neglecting the deeper, implicit layers of meaning that are fundamental to human communication. Decades of linguistic theory have shown that meaning is often shaped not just by what is explicitly stated, but by what is implied, presupposed, or embedded within cultural and social context [50, 91, 63, 131, 17]. These Implicit meanings (e.g., pragmatic intent, speaker stance, and ideological framing) play a crucial role in how language is interpreted, shaping meaning in ways that go far beyond surface form.

Why Current Models Miss Implicit Meaning Yet, current embedding models are not designed to capture these rich and nuanced aspects of meaning. This limitation stems from two core issues: training data rarely provides supervision for implicit meaning, and benchmarks do not evaluate or reward its capture. Most embedding models are trained on datasets optimized for surface-level similarity, particularly those derived from information retrieval tasks [9, 70], which offer little opportunity to learn context-sensitive or socially grounded semantics. Compounding this issue, widely adopted benchmarks rarely test for deeper interpretive capabilities [148, 102], further disincentivizing the development of models that aim to go beyond shallow semantic matching. As a result, even state-of-the-art embeddings often fall short in capturing the implicit dimensions of language that are essential for human-like understanding.

The Performance Divide To investigate this limitation, we conduct a pilot study using a suite of linguistically informed datasets covering three tiers of implicit meaning: (1) utterance level (pragmatic inference), (2) speaker level (stance), and (3) society level (political and social bias). The empirical results reveal that state-of-the-art embedding models, despite excelling on conventional benchmarks, perform only marginally better than the Bag-of-Tokens baseline on tasks requiring implicit understanding. As illustrated in Figure 1, there is a substantial performance gap between models' capabilities to capture surface meaning versus implicit semantics.

Our Position **We argue that the text embedding research community must move beyond surface-level semantics and explicitly embrace implicit meaning as a core modeling objective.** This position paper calls for a shift in research priorities–toward curating more linguistically grounded training data, developing benchmarks that evaluate deeper semantic and social understanding, and building embedding models that more faithfully reflect the complexity of human communication.

2 Linguistic Foundations of Implicit Meaning

To sharpen our understanding, we first revisit the linguistic foundations of implicit meaning through a three-tier framework: utterance (pragmatics), speaker (stance-taking), and society (sociolinguistics).

Research Question: Implicit Semantics

How do linguistic theories shape implicit meaning?

2.1 Utterance Level: Linguistic Signals of Implicit Meaning

Pragmatics investigates how utterances derive meaning from context, bridging the gap between literal surface semantics and the speaker's intended message [40, 50, 91]. It foregrounds what is left unsaid yet successfully communicated, revealing interpretive layers that semantic analysis alone cannot fully capture. This perspective has significantly influenced NLP, particularly in tasks requiring deeper contextual reasoning [47, 18].

At the heart of pragmatics is the insight that meaning emerges from broader situational, social, and cultural contexts, including shared background knowledge and prevailing norms, which collectively guide interpretation [50, 91]. Within this framework, speakers frequently rely on **implicature**, indirect cues inferred rather than explicitly stated [40, 119, 48, 91]. For instance, the sentence "*Bart managed to pass the test*" subtly suggests his success was unexpected, though not logically entailed.

Another key construct is **presupposition**, where utterances embed background assumptions required for comprehension [119, 91]. A statement like "*Sam quit smoking*" presupposes that Sam smoked before, an assumption that persists under negation or interrogation. Together, these phenomena demonstrate how implicit meaning arises not only from what is said, but also from what is assumed or inferred–posing a foundational challenge for text embeddings aiming to model such nuance.

2.2 Speaker Level: Cognitive Processes in Implicit Meaning

While pragmatics focuses on utterances in context, the concept of **stance** emphasizes the speaker's internal positioning–expressing attitudes, evaluations, and degrees of alignment or commitment [63]. Stance-taking is crucial to implicit meaning, as it reveals emotional and social orientation through subtle linguistic cues. Kiesling's model formalizes stance through three dimensions: evaluation (positive or negative appraisal), alignment (social positioning relative to others), and investment (the degree of speaker commitment) [32, 75].

Sociolinguistic variation often reflects stance. Forms like *-in'* vs. *-ing* sever not only as dialectal variants but as markers of toughness, informality, or solidarity [62, 152]. Over time, such forms become enregistered–decoupled from specific groups and reused more broadly to index stance. For example, the word *dude* has shifted from a gendered term to a marker of casual camaraderie [61]. Quantitative studies further reveal that stance fluctuates across discourse, with dynamic shifts in speaker intent and alignment observed in corpora like Reddit [64]. In short, stance introduces a relational, affective, and indexical layer to meaning–complementing pragmatics and posing a challenge for embeddings to capture speaker intent and social positioning.

2.3 Society Level: Cultural Shaping of Implicit Meaning

Beyond individual cognition and utterances, **sociolinguistics** explores how meaning is shaped by identity, power, and culture. Variation in pronunciation, grammar, or vocabulary, such as dropping the *g* in *workin'*, regional vowel shifts, or particles like *lah* in Singapore English, serves as a social index, signaling class, peer-group belonging, or regional identity [131]. These features are culturally contingent: the same form may index friendliness in one context and lack of education in another. Embedding models that collapse such variation into surface-level representations risk erasing these nuanced social signals.

Language ideologies further complicate this picture by privileging certain varieties while stigmatizing others [15]. As high-status registers dominate pretraining corpora, embeddings often reflect and amplify social hierarchies. For example, African-American Vernacular English may be marginalized relative to Standard American English, encoding structural inequalities as statistical artifacts. Speakers also fluidly shift styles–alternating registers, dialects, or slang–to perform identity and negotiate relationships [17]. These shifts carry implicit social meaning, signaling inclusion, authority, or deference. Yet static embeddings, which average across usage, struggle to capture the fast-paced recalibration of meaning. To reflect the social dimension of language, embeddings must account for the implicit cultural cues embedded in linguistic variation.

Takeaway: Linguistic Layers of Implicit Meaning

Implicit meaning unfolds across three interconnected linguistic tiers: (1) **pragmatics** captures what is implied but unsaid at the utterance level; (2) **stance-taking** reveals the speaker's evaluative, relational, and intentional positioning; and (3) **sociolinguistics** exposes how language encodes identity, culture, and power. Together, these layers illustrate that meaning is not fixed or literal, but deeply contextual, socially embedded, and dynamically performed, posing a significant challenge for embedding models still anchored in surface-level representations.

3 Text Embedding Models

Text embedding, the task of mapping text into dense vector representations, has long been central to NLP and now underpins many state-of-the-art applications. This section surveys the evolution of embedding models, highlights active research directions, and critically examines the field's current limitations. Figure 2 provides an overview of the major model classes and trending topics shaping the current embedding landscape.

Research Question: Research Focus

What is the current state of research on text embedding models?

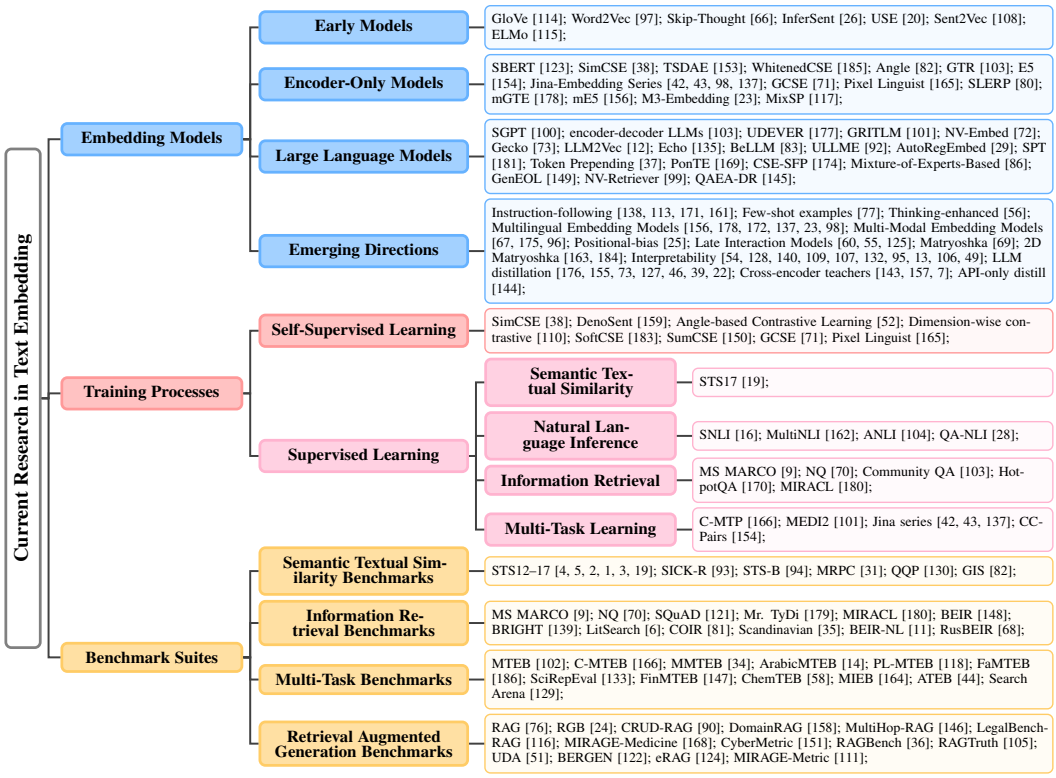

Figure 2: Taxonomy of current research in text embedding community.

Early Models Initial approaches rely on static word vectors pooled into sentence-level representations, such as Word2Vec and GloVe [97, 114]. Later models like Skip-Thought [66], InferSent [26], Sent2Vec [108], and the Universal Sentence Encoder [20] explore recurrent, transformer, or bilinear architectures. ELMo [115] marks a shift toward contextualized embeddings, dynamically encoding word meaning based on surrounding context.

Encoder-Only Models Pretrained encoder-only Transformers like BERT [30] and RoBERTa [88] enable context-aware sentence embeddings via [CLS] or mean pooling, often optimized with contrastive or denoising objectives. Subsequent advances, including Sentence-BERT [123], SimCSE [38], TSDAE [153], and E5 [154], improved embedding quality through new training strategies and architecture refinements [185, 82, 67, 71, 165, 80, 117, 137, 23, 98, 42, 43].

Large Language Models (LLMs) LLMs have recently been adapted for embedding tasks using both decoder-only and encoder-decoder designs. Research has explored fine-tuning, prompting, and other strategies to adapt general-purpose LLMs for dense, semantically rich representations. This includes efforts to repurpose decoder-only LLMs [100], adapt encoder-decoder frameworks [103], or build new LLM-based embedding models [101, 72, 73, 12, 135, 83, 92, 29, 181, 37, 169, 174, 86]. While these models offer strong semantic capabilities, their large size and high inference cost limit real-world deployment, sustaining demand for lighter encoder-based alternatives.

Emerging Directions Several trends have shaped recent progress. Instruction-following [138, 113, 171, 161], few-shot embedding [77], and "thinking-enhanced" representations [56] aim to improve adaptability. Multilingual and cross-lingual models [156, 178, 172, 137, 23, 98] expand embedding utility across languages. New forms of architectural and training design continue to appear, including sentence embeddings with hypernetworks [171], mitigation of positional bias [25], and efficiency-oriented models like ColBERT [60, 125, 55], Matryoshka [69], and 2D-layered embeddings [163, 184]. Interpretability remains a growing concern, with work exploring human-understandable embeddings [54, 128, 140, 109, 107, 132, 95, 13, 106, 49, 141].

Another emerging direction involves distilling knowledge from LLMs into lightweight sentence embedding models. This includes generating or augmenting training data with LLMs [176, 155, 73, 127, 46, 39], as well as distilling from more accurate but slower cross-encoder models [143, 157, 7].

Some approaches further leverage sentence summaries [150] or distill from proprietary APIs [144]. While these techniques expand supervision and improve performance, they mostly reinforce surface-level semantics, with limited attention to implicit meaning.

Open Questions: What Should Embeddings Capture? Despite rapid progress, a central question remains underexplored: what should text embeddings truly capture? While current models excel at encoding surface-level semantics for benchmark-driven tasks, it is less clear whether they can represent more nuanced dimensions such as speaker stance, social context, or pragmatic intent. In the following, we argue that implicit semantics remains a significantly underexplored dimension in the training and evaluation of text embeddings.

Takeaway: Landscape of Text Embedding Models

Research on text embedding models spans architectures, multilinguality, interpretability, and efficiency. Yet, the capacity to capture implicit semantics, which is central to real-world meaning, remains a significantly underexplored frontier.

4 Training Processes Fail to Capture Implicit Semantics

Despite significant progress in text embedding models, most training methods remain limited to capture implicit meaning. This section examines the two dominant paradigms: self-supervised and supervised learning, highlighting how both rely on datasets and objectives that prioritize surface-level semantics, leaving deeper contextual and social meanings underrepresented.

Research Question: Training Gap

How are text embedding models trained, and why does this fail to capture implicit meaning?

4.1 Self-Supervised Learning

Self-supervised learning trains on unlabeled text by extracting signals through augmentation or structural cues, without requiring manual labels. Techniques like SimCSE [38] uses dropout noise to create sentence pairs, while DenoSent [159] applies a denoising objective. Other approaches explore novel formulations, such as angle-based learning [52], dimension-wise contrastive loss [110], and similarity-weighted negative sampling [183]. Although these methods avoid costly annotations, they generally underperform compared to supervised approaches. Consequently, many embedding models adopt a two-stage pipeline: self-supervised pre-training followed by supervised fine-tuning [38].

4.2 Supervised Learning

Supervised training typically builds on pre-trained language models, and applies contrastive learning with losses such as Triplet Loss [123], SimCSE Loss [38], and Angle Loss [82]. These methods require labeled positive and negative pairs, which are absent in general-purpose corpora like C4 [120], leading researchers to rely on task-specific datasets such as Semantic Textual Similarity (STS), Natural Language Inference (NLI), and Information Retrieval (IR), or multi-task combinations.

Semantic Textual Similarity (STS) Datasets like STS17 [19] offer fine-grained similarity signals and have been widely used in models like Sentence-BERT [123] and TSDAE [153]. However, their small scale and narrow domain coverage often lead to overfitting and poor generalization [102].

Natural Language Inference (NLI) Datasets such as SNLI [16], MultiNLI [162], ANLI [104], and QA-NLI [28] annotate sentence pairs with *entailment*, *contradiction*, or *neutral*. These are widely used in models like Sentence-BERT [123], TSDAE [153], E5 [154], UDEVER [177], Angle [82], and GritLM [177]. While these datasets offer greater scale and domain diversity, the semantic signals often reflect shallow equivalence. For instance, SNLI pairs like "*A boy is jumping on a skateboard*" and "*The boy does a skateboarding trick*" fail to probe deeper pragmatic intent [16].

Information Retrieval (IR) Datasets, notably MS MARCO [9] and Natural Questions (NQ) [70], dominate retrieval-based training. Models like GTR [103], E5 [154], UDEVER [177], GritLM [101], and NV-Retriever [99] incorporate these datasets. The multilingual model mGTE [178] further draws

from sources including HotpotQA [170] and MIRACL [180]. Though useful for modeling lexical relevance, they reward literal matching and overlook implicit cues like stance or ideology.

Multi-Task Learning To improve generalization, models like mGTE [178] and the Jina Embeddings series [42, 43, 137] leverage multi-task corpora like C-MTP [166] and MEDI2 [101], which integrate STS, NLI, IR, QA, and other relevant data in pair or triplet format. While these datasets broaden coverage across tasks and domains, they still largely omit examples involving pragmatic inference, speaker stance, or sociocultural context, which are the key elements of implicit meaning.

Takeaway: Training Data Emphasize Surface Semantics

Despite innovations in architecture and supervision, current training pipelines remain anchored in datasets that prioritize surface-level similarity. While multi-task learning expands coverage, implicit semantics, such as pragmatics, stance, and social context, remain largely absent from the training signal.

5 Benchmarks Do Not Evaluate Implicit Semantics

Despite the growth of large-scale benchmark suites ranging from semantic similarity and retrieval to multi-task generalization, most evaluations remain focused on surface-level semantics. This section surveys widely used benchmarks, including STS datasets, retrieval-centric benchmarks like BEIR, comprehensive multi-task suites such as MTEB, and emerging Retrieval-Augmented Generation (RAG) evaluations. While these resources provide broad coverage across tasks, domains, and languages, they rarely assess how well models capture implicit, contextual, or socially situated meaning, leaving a critical gap in current evaluation practices.

Research Question: Evaluation Gap

How are current text embedding models evaluated, and why do existing benchmarks fall short in capturing implicit meaning?

Semantic Textual Similarity Benchmarks STS tasks measure alignment between model-predicted similarities and human-annotated semantic similarity scores, using metrics like Spearman correlation. Popular datasets include STS12–17 [4, 5, 2, 1, 3, 19], STS-B [94], and SICK-R [93], all included in the MTEB benchmark [102]. Related binary classification tasks include MRPC [31], QQP [130], and GIS [82]. Although STS tasks are theoretically capable of evaluating deeper meaning, most focus on lexical variation and syntactic paraphrasing. Their scope is limited by construction methods and annotator biases, and they were developed largely before the rise of LLMs. As a result, they fail to probe pragmatic, attitudinal, or culturally embedded semantics.

Information Retrieval Benchmarks IR benchmarks assess how well models retrieve relevant documents using embedding similarity. Performance is measured using ranking metrics such as MRR, nDCG@k, and Recall@k [160, 148]. Datasets like MS MARCO [9], NQ [70], SQuAD [121], Mr. TyDi [179], and MIRACL [180] are commonly used, with BEIR [148] aggregating 18 such datasets across diverse retrieval scenarios. Newer domain- and language-specific benchmarks include BRIGHT [139], LitSearch [6], COIR [81], the Scandinavian Benchmark [35], BEIR-NL [11], and RusBEIR [68]. Despite impressive coverage across domains and languages, IR tasks mostly evaluate surface-level relevance and do not test whether models capture deeper semantic alignment. Tasks like retrieving documents that match a speaker's stance or ideological framing remain underexplored.

Multi-Task Benchmarks MTEB [102] is a leading benchmark suite spanning 58 datasets and 8 task types. Variants such as C-MTEB [166], MMTEB [34], ArabicMTEB [14], PL-MTEB [118], and FaMTEB [186] expand coverage across languages, while domain-specific extensions like SciRepEval [133], FinMTEB [147], ChemTEB [58], and MIEB [164] target specific verticals. Challenging tasks like reasoning and instruction-following have been introduced in ATEB [44]. Crowdsourced platforms like MTEB Arena[1] and Search Arena[2] provide user-driven comparisons across tasks [129].

[1] `https://huggingface.co/spaces/mteb/arena`

[2] `https://blog.lmarena.ai/blog/2025/search-arena/`

Despite offering flexible, model-agnostic evaluation, these platforms still rely on traditional metrics and rarely test for implicit meaning. In practice, only a few MTEB datasets go beyond surface semantics, limiting their value for evaluating interpretive depth.

Retrieval-Augmented Generation (RAG) Benchmarks RAG benchmarks evaluate how well embeddings retrieve relevant content to support generative tasks. Benchmarks such as RGB [24], CRUD-RAG [90], DomainRAG [158], MultiHop-RAG [146], , LegalBench-RAG [116], MIRAGE [168] and CyberMetric [151], cover multilingual, domain-specific, and multi-hop scenarios. Other general-purpose tools include RAGBench [36], RAGTruth [105], UDA [51], and toolkits like BERGEN [122]. Recent proposals like eRAG [124] and another MIRAGE [111] support fine-grained retrieval evaluation. While RAG setups touch on reasoning and hallucination, they still prioritize factual retrieval. As a result, the underlying semantic evaluations resemble IR tasks, offering limited insight into how well embeddings reflect implicit intent, stance, or social meaning.

Takeaway: Current Evaluation Focuses Mostly on Surface Semantics

Current benchmarks provide broad task and domain coverage, but overwhelmingly emphasize surface-level similarity and relevance. They rarely assess a model's ability to capture implicit meaning, such as pragmatics, stance, or social context, leaving a critical gap in how we evaluate semantic understanding.

6 Empirical Evidences

To provide empirical evidence and motivate future research, we conduct a pilot study evaluating whether state-of-the-art embedding models can effectively capture implicit semantics.

Research Question: Empirical Gap

Do state-of-the-art embedding models effectively capture implicit meaning across utterance, speaker, and society levels?

Experimental Setup We evaluate embeddings on seven datasets spanning three tiers of implicit meaning: (1) Utterance level: Pragmatics Understanding Benchmark (PUB), including Implicature (**P-IMP**), Presupposition (**P-PRE**), and Reference & Deixis (**P-R&D**) [136, 89, 182, 87, 21, 53, 112]; (2) Speaker level: **P-Stance** dataset [84] for stance detection; and (3) Society level: the datasets of Implicit Hate Speech (**IHS**) [33], Social Bias Inference Corpus (**SBIC**) [126], and Political Bias (**Pol. Bias**) [10]. Together, these datasets provide a structured view of implicit meaning.

Since these datasets were not originally designed for embedding evaluation, we reformulate them into classification, pairwise classification (following the MTEB benchmark [102]), and zero-shot formats, where models select the label with the highest embedding similarity. We test models from four representative categories: encoder-only models, LLM-based models, multimodal encoder models, and proprietary embeddings (OpenAI). Bag-of-Tokens [45, 141] and random baselines are included for comparison. Implementation details are provided in Appendix A.1.

Results and Analysis As depicted in Table 1, encoder-only models often perform only marginally better than Bag-of-Tokens and random baselines. LLM-based models and OpenAI embeddings generally achieve stronger results. Although OpenAI models rank lower on MTEB, they perform well on these implicit semantics datasets, highlighting a potential disconnect between benchmark performance and deeper semantic competence.

Moreover, performance varies by semantic tier. As shown, **Linq-Mistral** excels in utterance-level tasks, **OpenAI-Large** leads in speaker and societal datasets, and **E5-Mistral** shows strength in political bias detection. These differences suggest that current models may specialize in different semantic dimensions, revealing fragmentation in their implicit meaning capabilities.

Overall, these observations affirm this paper's central claim: *state-of-the-art embedding models remain limited in capturing implicit semantics.* High MTEB scores do not translate to robustness on tasks involving pragmatic inference, stance, or social context. The fact that many models barely surpass Bag-of-Tokens underscores a fundamental evaluation gap.

Model	Utterance Level			Speaker Level	Society Level			Avg. Acc. ↑
	P-IMP	P-PRE	P-R&D	P-Stance	IHS	SBIC	Pol. Bias	
Random	48.5	54.1	38.8	51.3	27.5	59.2	34.5	44.8
Bag-of-Tokens [141]	56.5	75.3	48.2	73.4	59.6	80.7	41.6	62.2
Encoder Only Models								
S-BERT [123]	61.7	72.8	55.7	72.9	60.8	81.8	47.9	64.8
GIST-Small [134]	65.8	76.1	58.8	76.0	61.8	81.6	49.1	67.0
BGE-Base [167]	65.0	75.6	57.3	74.4	62.9	82.1	52.1	67.1
Angle [82]	69.4	78.8	57.2	76.4	59.5	83.7	50.4	67.9
BGE-Large [167]	68.3	75.5	58.1	76.0	63.5	83.4	51.5	68.0
MXBAI-Large [74]	69.8	78.2	59.4	75.6	60.1	83.6	50.5	68.2
GIST-Large [134]	68.8	76.6	62.9	76.7	64.2	83.4	52.5	69.3
Stella [175]	72.1	81.5	59.6	76.5	60.4	84.0	54.4	69.8
Large Language Models								
Linq-Mistral [65]	**80.3**	**87.7**	**70.4**	75.8	61.4	82.0	56.8	73.5
E5-Mistral [155, 154]	78.1	81.8	63.4	81.1	63.9	84.8	**71.5**	74.9
GTE-Qwen [85]	73.4	87.3	68.1	80.9	65.6	84.5	66.8	75.2
Multimodal Encoder Models								
Jasper [175]	73.3	80.1	63.0	80.1	65.7	84.2	63.9	72.9
Proprietary Models								
OpenAI-Small	71.3	78.1	64.3	80.0	66.2	83.9	56.6	71.5
OpenAI-Large	76.0	80.2	66.4	**83.7**	**67.1**	**85.4**	66.3	75.0

Table 1: Average accuracy (%) of embedding models across seven datasets representing three tiers of implicit semantics: utterance level (pragmatics), speaker level (stance), and societal level (social meaning). Results highlight differences in model capabilities across semantic levels and underscore the challenges of capturing implicit meaning.

Takeaway: Empirical Evidence Highlights the Implicit Semantics Gap

Despite strong performance on standard benchmarks, embedding models struggle with tasks involving implicature, stance, and social meaning. Their inconsistent performance across semantic tiers and the proximity to Bag-of-Tokens baselines underscores the need for new training and evaluation strategies that directly target implicit semantics.

7 Towards Embeddings that Capture Implicit Meaning

To address the implicit semantics gap, we propose three complementary directions: enriching training data, designing targeted benchmarks, and treating implicit meaning as a core modeling objective. Together, these steps can guide the development of embeddings that go beyond surface-level similarity.

Research Question: Research Agenda

What steps should be taken to enable text embeddings to capture implicit meaning?

7.1 Curating More Diverse Training Data

Training data fundamentally shape what embedding models learn. As the adage goes, "*garbage in, garbage out*"–surface-level inputs yield surface-level representations. To enable models to capture implicit meaning, we must expand beyond narrow datasets and embrace greater linguistic, cultural, and contextual diversity. Beyond manual curation, recent advances in LLM-based data generation offer promising directions. Prior work has used LLMs to synthesize training examples for embedding models [155, 22]; future efforts should guide this generation toward phenomena like implicature, presupposition, and stance.

Linguistic theory provides a rich foundation for this endeavor. Decades of research have outlined typologies of implicit meaning, which can inform the design of more semantically grounded training

signals. Aligning synthetic data with these frameworks can help embeddings internalize meanings rooted in pragmatics and social context–dimensions often absent from existing datasets.

7.2 Designing Benchmarks for Implicit Meaning

Benchmarks drive progress by defining what models are expected to learn. However, existing suites like MTEB primarily test surface similarity. Their open-source nature has also led to data leakage and leaderboard inflation, weakening their value as generalization tests. This shift toward leaderboard optimization deviates from the original goal of embeddings: producing general-purpose, transferable representations. Among MTEB's 58 tasks, only a few probe beyond surface meaning, and even recent additions like ATEB emphasize reasoning or safety over pragmatic and cultural nuance.

New benchmarks should be explicitly constructed to test underrepresented forms of meaning. Tasks should include inference from indirect cues, stance recognition, and sociolinguistic variation, reflecting the interpretive demands of real-world language understanding.

7.3 Framing Implicit Semantics as a Modeling Goal

A deeper challenge is that implicit meaning is rarely treated as a first-class modeling objective. While LLM research increasingly investigates contextual, attitudinal, and social understanding [79, 78, 59, 136, 173, 27, 142, 91], embedding models remain optimized for benchmarks that reward superficial similarity. This misalignment leads models to optimize for what is easy to measure over what is meaningful to understand. Without explicitly targeting implicit semantics, advances in architecture and supervision risk reinforcing shallow representations. Reframing modeling goals around deeper semantic dimensions can produce embeddings that more faithfully reflect human communication.

Takeaway: Future Opportunities for Text Embedding Research

To move forward, text embedding research must embrace implicit meaning as a central objective. This includes: (1) curating linguistically informed and culturally diverse training data, (2) designing benchmarks that evaluate pragmatic, attitudinal, and social understanding, and (3) reframing implicit semantics as a core modeling goal. Such a shift will lead to more robust, context-aware representations for real-world applications.

8 Alternative Views

While this paper advocates for embedding models to capture implicit semantics, alternative perspectives support maintaining the current focus on surface-level similarity. One argument is that for many practical tasks, such as search, recommendation, or clustering, surface semantics are often sufficient. Incorporating deeper meaning may add complexity without clear benefits.

Another view holds that pragmatic and socially grounded meaning is better handled by LLMs, which are explicitly designed for contextual reasoning and discourse-level understanding. In contrast, embeddings are valued for their efficiency and general-purpose utility. From this perspective, expecting embeddings to model implicit meaning may blur their role and dilute their purpose.

9 Conclusions

Despite significant progress in text embedding research, current models remain narrowly focused on surface-level semantics, failing to capture the implicit meanings that are central to human communication. This paper calls for a paradigm shift: embedding models must move beyond lexical similarity to explicitly model pragmatic, attitudinal, and sociocultural meaning. Drawing from linguistic theory, we propose a three-tier framework for implicit meaning and present empirical evidence that state-of-the-art models struggle with tasks requiring deeper interpretive reasoning. To advance the field, we advocate for semantically richer and more diverse training data, benchmarks that directly evaluate implicit understanding, and a reframing of implicit semantics as a core modeling objective. Embeddings that capture these deeper dimensions will enable more robust, context-aware systems aligned with the complexity of real-world language.

References

[1] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, 2015.

[2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, 2014.

[3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, 2016.

[4] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, 2012.

[5] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, June 2013.

[6] Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. Litsearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15068–15083, 2024.

[7] Haritha Ananthakrishnan, Julian Dolby, Harsha Kokel, Horst Samulowitz, and Kavitha Srinivas. Can cross encoders produce useful sentence embeddings? *arXiv preprint arXiv:2502.03552*, 2025.

[8] Dimo Angelov and Diana Inkpen. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, 2024.

[9] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[10] Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, 2020.

[11] Nikolay Banar, Ehsan Lotfi, and Walter Daelemans. Beir-nl: Zero-shot information retrieval benchmark for the dutch language. *arXiv preprint arXiv:2412.08329*, 2024.

[12] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

[13] Vinamra Benara, Chandan Singh, John Xavier Morris, Richard Antonello, Ion Stoica, Alexander Huth, and Jianfeng Gao. Crafting interpretable embeddings for language neuroscience by asking LLMs questions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, page 124137, 2024.

[14] Gagan Bhatia, El Moatez Billah Nagoudi, Abdellah El Mekki, Fakhraddin Alwajih, and Muhammad Abdul-Mageed. Swan and arabicmteb: Dialect-aware, arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. *arXiv preprint arXiv:2411.01192*, 2024.

[15] Pierre Bourdieu. Language and symbolic power. *Polity*, 1991.

[16] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642, 2015.

[17] Mary Bucholtz and Kira Hall. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7(4-5):585–614, 2005.

[18] Erik Cambria. Pragmatics processing. In *Understanding Natural Language Understanding*, pages 229–338. 2024.

[19] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

[20] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 169–174, 2018.

[21] Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. Flute: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7139–7159, 2022.

[22] Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. Little giants: Synthesizing high-quality embedding data at scale. *arXiv preprint arXiv:2410.18634*, 2024.

[23] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

[24] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 17754–17762, 2024.

[25] João Coelho, Bruno Martins, João Magalhães, Jamie Callan, and Chenyan Xiong. Dwell in the beginning: How language models embed long documents for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–377, 2024.

[26] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, 2017.

[27] Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy. Classist tools: Social class correlates with performance in nlp. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12643–12655, 2024.

[28] Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*, 2018.

[29] Jingcheng Deng, Zhongtao Jiang, Liang Pang, Liwei Chen, Kun Xu, Zihao Wei, Huawei Shen, and Xueqi Cheng. Following the autoregressive nature of llm embeddings via compression and alignment. *arXiv preprint arXiv:2502.11401*, 2025.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (NAACL-HLT)*, pages 4171–4186, 2019.

[31] Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*, 2005.

[32] John W Du Bois. The stance triangle. In *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, pages 139–182. 2008.

[33] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 345–363, 2021.

[34] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

[35] Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer L Nielbo. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:40336–40358, 2024.

[36] Robert Friel, Masha Belyi, and Atindriyo Sanyal. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*, 2024.

[37] Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, and Qing Gu. Token prepending: A training-free approach for eliciting better sentence embeddings from llms. *arXiv preprint arXiv:2412.11556*, 2024.

[38] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910, 2021.

[39] Waris Gill, Justin Cechmanek, Tyler Hutcherson, Srijith Rajamohan, Jen Agarwal, Muhammad Ali Gulzar, Manvinder Singh, and Benoit Dion. Advancing semantic caching for llms with domain-specific embeddings and synthetic data. *arXiv preprint arXiv:2504.02268*, 2025.

[40] Herbert P Grice. Logic and conversation. In *Speech Acts*, pages 41–58. Brill, 1975.

[41] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[42] Michael Günther, Louis Milliken, Jonathan Geuter, Georgios Mastrapas, Bo Wang, and Han Xiao. Jina embeddings: A novel set of high-performance sentence embedding models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, December 2023.

[43] Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, et al. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*, 2023.

[44] Simeng Han, Frank Palma Gomez, Tu Vu, Zefei Li, Daniel Cer, Hansi Zeng, Chris Tar, Arman Cohan, and Gustavo Hernandez Abrego. Ateb: Evaluating and improving advanced nlp tasks for text embedding models. *arXiv preprint arXiv:2502.16766*, 2025.

[45] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[46] Liyang He, Chenglong Liu, Rui Li, Zhenya Huang, Shulan Ruan, Jun Zhou, and Enhong Chen. Refining sentence embedding model through ranking sentences generation with large language models. *arXiv preprint arXiv:2502.13656*, 2025.

[47] Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 588–602, 2021.

[48] Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, 2023.

[49] James Huang, Wenlin Yao, Kaiqiang Song, Hongming Zhang, Muhao Chen, and Dong Yu. Bridging continuous and discrete spaces: Interpretable sentence representation learning via compositional operations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14584–14595, 2023.

[50] Yan Huang. Introduction: What is pragmatics? In Yan Huang, editor, *The Oxford Handbook of Pragmatics*, Oxford Handbooks. Oxford University Press, 2017.

[51] Yulong Hui, Yao Lu, and Huanchen Zhang. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. *arXiv preprint arXiv:2406.15187*, 2024.

[52] Yoo Hyun Jeong, Myeongsoo Han, and Dong-Kyu Chae. A simple angle-based approach for contrastive learning of unsupervised sentence representation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5553–5572, 2024.

[53] Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models imppressive? learning implicature and presupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4437–4452, 2020.

[54] Kishlay Jha, Yaqing Wang, Guangxu Xun, and Aidong Zhang. Interpretable word embeddings for medical domain. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1061–1066, 2018.

[55] Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrapas, Saba Sturua, Isabelle Mohr, Andreas Koukounas, Mohammad Kalim Akram, Nan Wang, and Han Xiao. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. *arXiv preprint arXiv:2408.16672*, 2024.

[56] Yifan Ji, Zhipeng Xu, Zhenghao Liu, Yukun Yan, Shi Yu, Yishan Li, Zhiyuan Liu, Yu Gu, Ge Yu, and Maosong Sun. Learning more effective representations for dense retrieval through deliberate thinking before search. *arXiv preprint arXiv:2502.12974*, 2025.

[57] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[58] Ali Shiraee Kasmaee, Mohammad Khodadad, Mohammad Arshi Saloot, Nicholas Sherck, Stephen Dokas, Hamidreza Mahyar, and Soheila Samiee. Chemteb: Chemical text embedding benchmark, an overview of embedding models performance & efficiency on a specific domain. *arXiv preprint arXiv:2412.00532*, 2024.

[59] Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:39052–39074, 2023.

[60] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*, pages 39–48, 2020.

[61] Scott F Kiesling. Dude. *American Speech*, 79(3):281–305, 2004.

[62] Scott F. Kiesling. Style as stance: Stance as the explanation for patterns of sociolinguistic variation. In *Stance: Sociolinguistic Perspectives*, pages 171–194. 2009.

[63] Scott F Kiesling. Stance and stancetaking. *Annual Review of Linguistics*, 8(1):409–426, 2022.

[64] Scott F Kiesling, Umashanthi Pavalanathan, Jim Fitzpatrick, Xiaochuang Han, and Jacob Eisenstein. Interactional stancetaking in online forums. *Computational Linguistics*, 44(4):683–718, 2018.

[65] Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy yong Sohn, and Chanyeol Choi. Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog, 2024.

[66] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS)*, pages 3294–3302, 2015.

[67] Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, et al. Jina clip: Your clip model is also your text retriever. *arXiv preprint arXiv:2405.20204*, 2024.

[68] Grigory Kovalev, Mikhail Tikhomirov, Evgeny Kozhevnikov, Max Kornilov, and Natalia Loukachevitch. Building russian benchmark for evaluation of information retrieval models. *arXiv preprint arXiv:2504.12879*, 2025.

[69] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:30233–30249, 2022.

[70] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics (TACL)*, 7:453–466, 2019.

[71] Peichao Lai, Zhengfeng Zhang, Wentao Zhang, Fangcheng Fu, and Bin Cui. Enhancing unsupervised sentence embeddings via knowledge-driven data augmentation and gaussian-decayed contrastive learning. *arXiv preprint arXiv:2409.12887*, 2024.

[72] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

[73] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*, 2024.

[74] Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread - new fluffy embeddings model, 2024.

[75] Michael Lempert. The poetics of stance: Text-metricality, epistemicity, interaction. *Language in Society*, 37(4):569–592, 2008.

[76] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 9459–9474, 2020.

[77] Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700*, 2024.

[78] Hengli Li, Song-Chun Zhu, and Zilong Zheng. Diplomat: a dialogue dataset for situated pragmatic reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:46856–46884, 2023.

[79] Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 2642–2652, 2020.

[80] Mingxin Li, Zhijie Nie, Yanzhao Zhang, Dingkun Long, Richong Zhang, and Pengjun Xie. Improving general text embedding model: Tackling task conflict and data imbalance through model merging. *arXiv preprint arXiv:2410.15035*, 2024.

[81] Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Hao Zhang, Xinyi Dai, Yasheng Wang, and Ruiming Tang. Coir: A comprehensive benchmark for code information retrieval models. *arXiv preprint arXiv:2407.02883*, 2024.

[82] Xianming Li and Jing Li. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1825–1839, 2024.

[83] Xianming Li and Jing Li. BeLLM: Backward dependency enhanced large language model for sentence embeddings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 792–804, 2024.

[84] Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL 2021*, pages 2355–2365, 2021.

[85] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

[86] Ziyue Li and Tianyi Zhou. Your mixture-of-experts llm is secretly an embedding model for free. *arXiv preprint arXiv:2410.10814*, 2024.

[87] Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4437–4452, 2022.

[88] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[89] Annie Louis, Dan Roth, and Filip Radlinski. "i'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, 2020.

[90] Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*, 43(2):1–32, 2025.

[91] Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *arXiv preprint arXiv:2502.12378*, 2025.

[92] Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, and Thien Huu Nguyen. Ullme: A unified framework for large language model embeddings with generation-augmented learning. *arXiv preprint arXiv:2408.03402*, 2024.

[93] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 216–223, 2014.

[94] Philip May. Machine translated multilingual sts benchmark dataset., 2021.

[95] Denis McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron Wallace. CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8477–8494, 2023.

[96] Zhongtao Miao, Qiyu Wu, Kaiyan Zhao, Zilong Wu, and Yoshimasa Tsuruoka. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3225–3236, 2024.

[97] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119, 2013.

[98] Isabelle Mohr, Markus Krimmel, Saba Sturua, Mohammad Kalim Akram, Andreas Koukounas, Michael Günther, Georgios Mastrapas, Vinit Ravishankar, Joan Fontanals Martínez, Feng Wang, et al. Multi-task contrastive learning for 8192-token bilingual text embeddings. *arXiv preprint arXiv:2402.17016*, 2024.

[99] Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 2024.

[100] Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*, 2022.

[101] Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024.

[102] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2014–2037, 2023.

[103] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9844–9855, 2022.

[104] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4885–4901, 2020.

[105] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10862–10878, 2024.

[106] Charles O'Neill, Christine Ye, Kartheik Iyer, and John F Wu. Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657*, 2024.

[107] Juri Opitz and Anette Frank. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, pages 625–638, 2022.

[108] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 528–540, 2018.

[109] Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. Word2Sense: Sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5692–5705, 2019.

[110] Duccio Pappadopulo and Marco Farina. Non-contrastive sentence representations via self-supervision. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4274–4284, 2024.

[111] Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heui-Seok Lim. Mirage: A metric-intensive benchmark for retrieval-augmented generation evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2883–2900, 2025.

[112] Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel Bowman, and Tal Linzen. Nope: A corpus of naturally-occurring presuppositions in english. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, pages 349–366, 2021.

[113] Letian Peng, Yuwei Zhang, Zilong Wang, Jayanth Srinivasa, Gaowen Liu, Zihan Wang, and Jingbo Shang. Answer is all you need: Instruction-following text embedding via answering the question. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 459–477, 2024.

[114] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[115] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2227–2237, 2018.

[116] Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024.

[117] Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Sarana Nutanong. Space decomposition for sentence embedding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11227–11239, 2024.

[118] Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. Pl-mteb: Polish massive text embedding benchmark. *arXiv preprint arXiv:2405.10138*, 2024.

[119] Christopher Potts. Presupposition and implicature. *The handbook of contemporary semantic theory*, pages 168–202, 2015.

[120] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

[121] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[122] David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. Bergen: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663, 2024.

[123] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, 2019.

[124] Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2395–2400, 2024.

[125] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3715–3734, 2022.

[126] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5477–5490, 2020.

[127] Soma Sato, Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. Improving sentence embeddings with automatic generation of training data using few-shot examples. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 519–530, 2024.

[128] Lutfi Kerem Senel, Ihsan Utlu, Veysel Yucesoy, Aykut Koc, and Tolga Cukur. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(10):1769–1779, 2018.

[129] Sahel Sharifymoghaddam, Shivani Upadhyay, Nandan Thakur, Ronak Pradeep, and Jimmy Lin. Chatbot arena meets nuggets: Towards explanations and diagnostics in the evaluation of llm responses. *arXiv preprint arXiv:2504.20006*, 2025.

[130] Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041*, 2019.

[131] Michael Silverstein. Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, 23(3-4):193–229, 2003.

[132] Adi Simhi and Shaul Markovitch. Interpreting embedding spaces by conceptualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1704–1719, 2023.

[133] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5548–5566, 2023.

[134] Aivin V Solatorio. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*, 2024.

[135] Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. Repetition improves language model embeddings. *arXiv preprint arXiv:2402.15449*, 2024.

[136] Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, 2024.

[137] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. Jina embeddings v3: Multilingual text encoder with low-rank adaptations. In *European Conference on Information Retrieval (ECIR)*, pages 123–129, 2025.

[138] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, July 2023.

[139] Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, et al. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*, 2024.

[140] Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. SPINE: SParse Interpretable Neural Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 4921–4928, 2018.

[141] Yiqun Sun, Qiang Huang, Yixuan Tang, Anthony K. H. Tung, and Jun Yu. A general framework for producing interpretable semantic text embeddings. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[142] Yiqun Sun, Qiang Huang, Yanhao Wang, and Anthony KH Tung. Diversinews: Enriching news consumption with relevant yet diverse news articles retrieval. *Proceedings of the VLDB Endowment*, 17(12):4277–4280, 2024.

[143] Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, and Jimmy Lin. Teaching dense retrieval models to specialize with listwise distillation and llm data augmentation. *arXiv preprint arXiv:2502.19712*, 2025.

[144] Manveer Singh Tamber, Jasper Xian, and Jimmy Lin. Can't hide behind the api: Stealing black-box commercial embedding models. *arXiv preprint arXiv:2406.09355*, 2024.

[145] Hongming Tan, Shaoxiong Zhan, Hai Lin, Hai-Tao Zheng, and Wai Kin Chan. QAEA-DR: A Unified Text Augmentation Framework for Dense Retrieval. *IEEE Transactions on Knowledge & Data Engineering (TKDE)*, 37(06):3669–3683, 2025.

[146] Yixuan Tang and Yi Yang. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*, 2024.

[147] Yixuan Tang and Yi Yang. Finmteb: Finance massive text embedding benchmark. *arXiv preprint arXiv:2502.10990*, 2025.

[148] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[149] Raghuveer Thirukovalluru and Bhuwan Dhingra. Geneol: Harnessing the generative power of llms for training-free sentence embeddings. *arXiv preprint arXiv:2410.14635*, 2024.

[150] Raghuveer Thirukovalluru, Xiaolan Wang, Jun Chen, Shuyang Li, Jie Lei, Rong Jin, and Bhuwan Dhingra. Sumcse: Summary as a transformation for contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3577–3588, 2024.

[151] Norbert Tihanyi, Mohamed Amine Ferrag, Ridhi Jain, Tamas Bisztray, and Merouane Debbah. Cybermetric: a benchmark dataset based on retrieval-augmented generation for evaluating llms in cybersecurity knowledge. In *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 296–302, 2024.

[152] Peter Trudgill. Sex, covert prestige and linguistic change in the urban british english of norwich. *Language in Society*, 1(2):179–195, 1972.

[153] Kexin Wang, Nils Reimers, and Iryna Gurevych. Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, 2021.

[154] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[155] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023.

[156] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

[157] Qitong Wang, Mohammed J Zaki, Georgios Kollias, and Vasileios Kalantzis. Multi-sense embeddings for language models and knowledge distillation. *arXiv preprint arXiv:2504.06036*, 2025.

[158] Shuting Wang, Jiongnan Liu, Shiren Song, Jiehan Cheng, Yuqi Fu, Peidong Guo, Kun Fang, Yutao Zhu, and Zhicheng Dou. Domainrag: A chinese benchmark for evaluating domain-specific retrieval-augmented generation. *arXiv preprint arXiv:2406.05654*, 2024.

[159] Xinghao Wang, Junliang He, Pengyu Wang, Yunhua Zhou, Tianxiang Sun, and Xipeng Qiu. Denosent: A denoising objective for self-supervised sentence representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 19180–19188, 2024.

[160] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of NDCG type ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 25–54, 2013.

[161] Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. Followir: Evaluating and teaching information retrieval models to follow instructions. *arXiv preprint arXiv:2403.15246*, 2024.

[162] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, pages 1112–1122, 2018.

[163] LI Xianming, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. Ese: Espresso sentence embeddings. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[164] Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. *arXiv preprint arXiv:2504.10471*, 2025.

[165] Chenghao Xiao, Zhuoxu Huang, Danlu Chen, G Thomas Hudson, Yizhi Li, Haoran Duan, Chenghua Lin, Jie Fu, Jungong Han, and Noura Al Moubayed. Pixel sentence representation learning. *arXiv preprint arXiv:2402.08183*, 2024.

[166] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 641–649, 2024.

[167] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval (SIGIR)*, pages 641–649, 2024.

[168] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, 2024.

[169] Kosuke Yamada and Peinan Zhang. Out-of-the-box conditional text embeddings from large language models. *arXiv preprint arXiv:2504.16411*, 2025.

[170] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[171] Young Yoo, Jii Cha, Changhyeon Kim, and Taeuk Kim. Hyper-cl: Conditioning sentence representations with hypernetworks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 700–711, 2024.

[172] Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. Arctic-embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*, 2024.

[173] Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. Do large language models understand conversational implicature–a case study with a chinese sitcom. In *China National Conference on Chinese Computational Linguistics*, pages 402–418, 2024.

[174] Bowen Zhang, Zixin Song, and Chunping Li. Cse-sfp: Enabling unsupervised sentence representation learning via a single forward pass. *arXiv preprint arXiv:2505.00389*, 2025.

[175] Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*, 2024.

[176] Junlei Zhang, Zhenzhong Lan, and Junxian He. Contrastive learning of sentence embeddings from scratch. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3916–3932, 2023.

[177] Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. Language models are universal embedders. *arXiv preprint arXiv:2310.08232*, 2023.

[178] Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*, 2024.

[179] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, 2021.

[180] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1114–1131, 2023.

[181] Kaiyan Zhao, Qiyu Wu, Zhongtao Miao, and Yoshimasa Tsuruoka. Prompt tuning can simply adapt large language models to text encoders. In *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pages 38–50, 2025.

[182] Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. Grice: A grammar-based dataset for recovering implicature and conversational reasoning. In *Findings of the Association for Computational Linguistics: ACL 2021*, pages 2074–2085, 2021.

[183] Haojie Zhuang, Wei Emma Zhang, Jian Yang, Weitong Chen, and Quan Z Sheng. Not all negatives are equally negative: Soft contrastive learning for unsupervised sentence representations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 3591–3601, 2024.

[184] Shengyao Zhuang, Shuai Wang, Bevan Koopman, and Guido Zuccon. Starbucks: Improved training for 2d matryoshka embeddings. *arXiv preprint arXiv:2410.13230*, 2024.

[185] Wenjie Zhuo, Yifan Sun, Xiaohan Wang, Linchao Zhu, and Yi Yang. WhitenedCSE: Whitening-based contrastive learning of sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 12135–12148, 2023.

[186] Erfan Zinvandi, Morteza Alikhani, Mehran Sarmadi, Zahra Pourbahman, Sepehr Arvin, Reza Kazemi, and Arash Amini. Famteb: Massive text embedding benchmark in persian language. *arXiv preprint arXiv:2502.11571*, 2025.

A Experiment Details

A.1 Implementation Details

Checkpoints Table 2 lists the model checkpoints used in the experiments presented in Section 6. We adopt the official checkpoints used by the MTEB benchmark and evaluate all models using the default settings from the Sentence Transformers library [123], without additional parameter tuning or prompting. For OpenAI's proprietary models, we obtain embeddings using OpenAI's official client library.[3] A random baseline is implemented by sampling predictions according to the label distribution of each dataset. For the baseline of Bag-of-Tokens [45, 141], we use the `google-bert/bert-base-uncased` tokenizer.

Model	Model Size	Checkpoint
S-BERT [123]	22.7M	`sentence-transformers\all-MiniLM-L6-v2`
GIST-Small [134]	33.4M	`avsolatorio\GIST-small-Embedding-v0`
BGE-Base [167]	109M	`BAAI\bge-base-en-v1.5`
Angle [82]	335M	`WhereIsAI\UAE-Large-V1`
BGE-Large [167]	335M	`BAAI\bge-large-en-v1.5`
MXBAI-Large [74]	335M	`mixedbread-ai\mxbai-embed-large-v1`
GIST-Large [134]	335M	`avsolatorio\GIST-large-Embedding-v0`
Stella [175]	435M	`NovaSearch\stella_en_400M_v5`
Linq-Mistral [65]	7.11B	`Linq-AI-Research\Linq-Embed-Mistral`
E5-Mistral [155, 154]	7.11B	`intfloat\e5-mistral-7b-instruct`
GTE-Qwen [85]	7.61B	`Alibaba-NLP\gte-Qwen2-7B-instruct`
Jasper [175]	1.99B	`NovaSearch\jasper_en_vision_language_v1`
OpenAI-Small	Unknown	`text-embedding-3-small`
OpenAI-Large	Unknown	`text-embedding-3-large`

Table 2: List of models and their corresponding checkpoints.

Tasks We evaluate a diverse set of tasks designed to capture different aspects of implicit semantics. Due to data format differences, we organize them into three evaluation settings: **classification**, **pair classification**, and **zero-shot classification**. Each setting includes the following datasets:

- **Classification:** From the **Pragmatics Understanding Benchmark (PUB)**, we include *Task 1 (Direct/Indirect Classification)*, *Task 2 (Response Classification without Implied Meaning)*, *Task 3 (with Implied Meaning)*, *Task 6 (Understanding Sarcasm)*, *Task 10 (Implicature NLI)*, *Task 11 (Presupposition NLI)*, *Task 12 (Presupposition over QA)*, and *Task 13 (Deictic QA)*. We also evaluate all three subsets of the **P-Stance** dataset–*Trump*, *Biden*, and *Bernie*–for stance classification. For the **Implicit Hate Speech (IHS)** dataset, we include *detection*, *categorization*, and *target identification* tasks. For the **Social Bias Inference Corpus (SBIC)**, we evaluate five binary classification tasks: *whoTarget* (whether the target is a group), *intentYN* (intent to offend), *sexYN* (presence of sexual content), *offensiveYN* (offensiveness), and *hasBiasedImplication* (biased implications). Lastly, we include the **Political Bias (Pol. Bias)** classification dataset.
- **Pair Classification:** We adapt *Task 5 (Agreement Detection)* from **PUB**.
- **Zero-shot Classification:** We include *Task 4 (Implicature Recovery)*, *Task 7 (Figurative Language Understanding — No Hint)*, *Task 8 (with Positive Hint)*, *Task 9 (with Contrastive Hint)*, and *Task 14 (Reference via Metonymy)* from **PUB**.

Evaluation Protocols For classification and pair classification tasks, we follow the standard protocol from the MTEB benchmark [102]. For zero-shot classification, we adopt the embedding-based approach described in OpenAI's documentation,[4] where both the input question and text are embedded together, and each answer option is embedded separately. The answer option with the highest similarity to the input is selected as the prediction.

[3] `https://platform.openai.com/docs/api-reference/embeddings`

[4] `https://platform.openai.com/docs/guides/embeddings#use-cases`

Model	PUB													
	Implicature										Presupposition		Ref. & Deixis	
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
Random	48.0	49.2	52.9	24.9	50.0	51.0	49.2	47.1	49.7	62.9	36.4	71.7	56.0	21.7
Bag-of-Tokens [141]	81.2	69.4	78.2	29.3	51.0	12.5	50.4	74.7	32.7	86.0	66.9	83.6	64.5	31.9
Encoder Only Models														
S-BERT [123]	82.0	72.4	83.1	35.4	53.5	34.0	59.9	78.6	39.0	79.0	60.0	**85.6**	68.5	42.9
GIST-Small [134]	74.8	74.2	83.3	44.5	56.1	40.2	67.4	86.8	52.4	77.9	66.9	85.2	68.5	49.0
BGE-Base [167]	83.0	71.5	79.6	33.2	56.0	42.8	69.8	80.9	55.6	77.6	66.1	85.2	68.0	46.6
Angle [82]	97.2	77.3	81.2	41.5	58.1	32.8	75.1	82.8	60.7	87.6	74.2	83.4	66.0	48.4
BGE-Large [167]	87.0	76.8	79.8	43.0	57.7	41.8	75.3	84.0	59.6	77.9	65.8	85.2	68.0	48.1
MXBAI-Large [74]	97.2	78.4	81.0	41.2	58.5	34.0	75.5	83.3	61.2	87.4	73.6	82.8	67.5	51.2
GIST-Large [134]	79.6	76.8	83.1	46.9	57.9	37.8	78.0	88.5	61.5	78.3	68.1	85.2	71.5	54.3
Stella [175]	95.8	79.8	81.9	51.1	60.0	35.5	76.2	87.7	60.8	91.7	83.6	79.3	66.0	53.2
Large Language Models														
Linq-Mistral [65]	99.6	**89.3**	**88.6**	47.5	**70.0**	67.0	**88.6**	**94.5**	61.4	**96.2**	**91.4**	84.0	74.0	**66.7**
E5-Mistral [155, 154]	97.2	87.2	85.8	43.8	69.5	**69.0**	87.7	92.7	**62.9**	85.0	78.3	85.2	68.0	58.9
GTE-Qwen [85]	**100.0**	87.7	87.5	43.6	61.8	63.0	69.0	82.7	48.5	89.8	89.4	85.2	**78.0**	58.2
Multimodal Encoder Models														
Jasper [175]	97.8	84.2	85.4	50.6	61.6	49.8	78.0	88.4	55.4	81.9	75.0	85.2	70.5	55.6
Proprietary Models														
OpenAI-Small	98.8	79.4	84.9	**56.0**	56.7	35.5	79.3	89.9	55.0	78.1	71.1	85.2	73.0	55.6
OpenAI-Large	99.6	87.7	87.7	50.8	61.9	55.5	83.9	91.8	58.4	83.1	75.3	85.2	73.5	59.3

Table 3: The accuracy (%) of embedding models on the Pragmatics Understanding Benchmark (PUB) tasks. Each task is labeled as T1–T14, corresponding to the 14 tasks in the PUB benchmark.

A.2 Additional Results

The complete results, including the accuracy (%) for individual task, are presented in Tables 3 and 4. The values reported in Table 1 are computed by averaging across tasks within each dataset.

Widespread Variance Across Models The results reveal inconsistent performance across embedding models. For example, many models achieve near-perfect accuracy on *Task 1 (Direct/Indirect Classification)* from PUB, while models such as **GIST-Small**, **S-BERT**, and **BGE-Base** perform only marginally better or even worse than the Bag-of-Tokens baseline. Similarly, on *Task 10 (Implicature NLI)*, several models, including OpenAI's proprietary models and the LLM-based **E5-Mistral** and **Jasper**, underperform the Bag-of-Tokens baseline. These findings demonstrate that strong performance on surface-level benchmarks does not reliably transfer to tasks requiring deeper semantic understanding.

Strengths of Large and Multimodal Models Large and multimodal models tend to lead in overall performance. **Jasper**, for example, ranks among the top across a wide range of tasks, particularly within **IHS** and **SBIC**. Similarly, large-scale models such as **E5-Mistral** and **OpenAI-Large** perform well across domains, excelling in social bias classification and pragmatics reasoning. These results suggest that increased model size contributes positively to handling complex semantic phenomena.

Persistent Challenges in Implicature and Reference Tasks Despite their strengths, even the largest models struggle with specific pragmatic tasks. Notably, **Task 4 (Implicature Recovery)** remains difficult across all models, with scores rarely exceeding 50%. Even top-tier models like **GTE-Qwen** and **OpenAI-Large** achieve only modest gains over Bag-of-Tokens. These findings point to a fundamental limitation in how current training pipelines address implicit meaning.

Implications for Benchmark and Model Design In summary, these results reveal persistent blind spots in current embedding models, particularly for tasks involving implicature, figurative language, presupposition, and social inference. Addressing these challenges will require more linguistically grounded training strategies and benchmark datasets that explicitly target underexplored aspects of implicit meaning.

Model	P-Stance			IHS			SBIC					Pol. Bias
	Trump	Biden	Bernie	Det.	Cat.	Tar.	Tar.	Intent	Sex	Off.	Bias	
Random	51.7	50.9	51.2	52.5	16.6	13.4	48.7	58.2	76.6	62.0	50.5	34.5
Bag-of-Tokens [141]	74.6	75.4	70.1	74.5	55.0	49.3	77.7	76.1	92.0	79.6	78.1	41.6
Encoder Only Models												
S-BERT [123]	72.1	77.4	69.3	73.2	58.0	51.2	78.8	78.0	91.9	81.2	79.1	47.9
GIST-Small [134]	76.6	78.4	72.9	74.0	59.5	51.9	77.9	77.7	93.1	81.3	78.0	49.1
BGE-Base [167]	74.5	78.8	69.9	74.2	60.7	53.9	78.7	78.5	92.6	82.0	78.6	52.1
Angle [82]	77.2	79.9	72.1	75.9	56.0	46.6	80.4	80.4	93.3	83.7	80.5	50.4
BGE-Large [167]	75.8	80.0	72.1	75.1	61.4	53.9	80.3	79.9	93.3	83.0	80.6	51.5
MXBAI-Large [74]	76.4	78.9	71.5	76.4	56.5	47.3	80.4	80.3	93.1	83.6	80.4	50.5
GIST-Large [134]	76.8	80.1	73.2	75.0	63.1	54.4	80.5	79.3	93.3	82.9	80.9	52.5
Stella [175]	79.2	80.0	70.2	76.9	56.7	47.6	81.2	80.9	93.4	83.2	81.5	54.4
Large Language Models												
Linq-Mistral [65]	79.4	78.8	69.3	75.1	57.8	51.2	79.7	79.1	89.8	81.9	79.8	56.8
E5-Mistral [155, 154]	84.8	82.3	76.1	79.2	61.7	50.9	82.0	**82.5**	93.3	83.9	82.1	**71.5**
GTE-Qwen [85]	83.8	82.0	76.9	79.1	63.2	54.5	82.1	80.6	94.0	83.7	82.0	66.8
Multimodal Encoder Models												
Jasper [175]	81.6	82.6	76.2	78.4	64.6	54.1	81.4	81.2	93.9	83.1	81.5	63.9
Proprietary Models												
OpenAI-Small	82.4	81.1	76.7	78.5	64.7	**55.2**	80.7	81.1	93.5	83.3	80.8	56.6
OpenAI-Large	**87.5**	**83.8**	**79.7**	**80.2**	**67.3**	53.7	**82.9**	82.3	**94.2**	**84.7**	**83.1**	66.3

Table 4: The accuracy (%) of embedding models on additional implicit meaning benchmarks. **P-Stance** includes stance detection tasks for Trump, Biden, and Bernie. **Implicit Hate Speech (IHS)** comprises detection (Det.), categorization (Cat.), and target identification (Tar.) tasks. The **Social Bias Inference Corpus (SBIC)** includes target (Tar.), intent (Int.), sexism (Sex.), offensiveness (Off.), and bias detection tasks. **Political Bias (Pol. Bias)** refers to the political ideology classification task.