Position: Text Embeddings Should Capture Implicit Semantics, Not Just Surface Meaning

Yiqun Sun [1] **Qiang Huang** [2] **Anthony K. H. Tung** [1] **Jun Yu** [2]

Abstract

This position paper argues that text embedding research should move beyond surface meaning and embrace implicit semantics as a central modeling objective. Text embeddings are a foundational component of modern NLP, underpinning a wide range of applications and driving sustained research progress. Despite rapid progress, most embedding models remain narrowly focused on surface-level semantics, whereas linguistic theory emphasizes that much of human meaning is implicit, shaped by pragmatics, speaker intent, and sociocultural context. Current models are typically trained on datasets that lack such depth and evaluated using benchmarks that reward surface similarity. As a result, they struggle with tasks that require interpretive reasoning, stance recognition, or socially grounded understanding. Our pilot study makes this limitation explicit, showing that even state-of-the-art embeddings achieve only marginal improvements over simple lexical baselines on tasks probing implicit semantics. We therefore call for a paradigm shift: embedding research should prioritize linguistically grounded and diverse training data, develop benchmarks that probe deeper semantic understanding, and treat implicit meaning as a core modeling objective to better align embeddings with real-world language complexity. The code is available at `http://github.com/dukesun99/Implicit-Embeddings`.

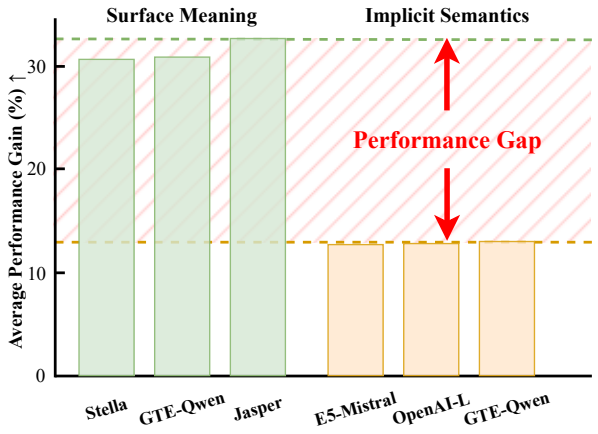

Figure 1. **Average gains over a lexical baseline (Bag-of-Tokens) for SOTA embedding models on Surface Meaning versus Implicit Semantics.** Surface performance is averaged over MTEB classification tasks (Muennighoff et al., 2023), while implicit-semantics performance is averaged over seven datasets probing pragmatic, attitudinal, and social meaning (Table 1). The comparison illustrates a performance divide: modern embeddings show large gains on surface-semantic benchmarks but much smaller gains on tasks requiring implicit interpretation.

1. Introduction

Text embedding models map sentences, passages, or documents into dense vectors whose proximity reflects semantic similarity (Reimers & Gurevych, 2019; Muennighoff et al., 2023; Sun et al., 2025d). They underpin much of modern NLP and Information Retrieval, serving as foundational components in Clustering (Grootendorst, 2022; Huang et al., 2023b; Angelov & Inkpen, 2024; Li et al., 2026), Classification (Muennighoff et al., 2023; Enevoldsen et al., 2025), Dense Retrieval (Thakur et al., 2021; Karpukhin et al., 2020; Sun et al., 2024; Huang et al., 2024; Tang et al., 2025), and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; You et al., 2026a;b; Dai et al., 2026; Sun et al., 2026). As a result, embedding models are widely deployed in a pre-trained, off-the-shelf manner and treated as general-purpose semantic interfaces for downstream decision-making.

This central role has driven rapid progress across model architectures (Reimers & Gurevych, 2019; Li & Zhou, 2024; BehnamGhader et al., 2024), training objectives (Thirukovalluru et al., 2024; Li & Li, 2024a; Xianming et al., 2025), and large-scale evaluation benchmarks (Muennighoff et al., 2023; Han et al., 2025; Enevoldsen et al., 2025). By stan-

[1]National University of Singapore [2]Harbin Institute of Technology (Shenzhen). Correspondence to: Qiang Huang <huangqiang@hit.edu.cn>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

dard benchmarks, modern embedding models appear increasingly strong, robust, and general-purpose.

The Overlooked Dimension: Implicit Semantics Despite this progress, we argue that contemporary embedding research remains narrowly focused on surface meaning, such as semantic signals derived from lexical overlap, syntactic alternation, and topical similarity, while systematically underrepresenting the implicit semantics that are fundamental to human language understanding. Decades of linguistic research demonstrate that meaning is frequently conveyed indirectly, shaped by pragmatic inference, speaker intent, stance-taking, and sociocultural context rather than explicit propositional content (Huang, 2017; Ma et al., 2025; Kiesling, 2022; Silverstein, 2003; Bucholtz & Hall, 2005). Such implicit meanings determine not only what is said, but how and why it is understood in context.

Crucially, these dimensions of meaning are not peripheral edge cases. They govern everyday interpretation in domains such as persuasion, ideology, politeness, sarcasm, safety, and social signaling. Yet they remain largely invisible to embedding models optimized for surface-level similarity.

Why Embedding Models Miss Implicit Meaning This limitation is structural rather than incidental. First, dominant training corpora provide little supervision for implicit semantics. Most embedding models are trained on retrieval, entailment, or paraphrase datasets (Bajaj et al., 2016; Kwiatkowski et al., 2019), where success is defined by lexical relevance or literal semantic equivalence rather than contextual interpretation. Second, evaluation benchmarks overwhelmingly reward surface alignment. Widely used suites rarely test whether embeddings distinguish implied intent, speaker stance, or socially grounded meaning (Thakur et al., 2021; Muennighoff et al., 2023).

As a result, embedding models are optimized for what is easy to measure rather than what is linguistically consequential. Even advanced embeddings, despite high benchmark scores, are neither trained nor evaluated to capture the deeper layers of meaning that humans routinely infer.

A Performance Divide To examine this gap empirically, we conduct a pilot study spanning three tiers of implicit semantics: (1) utterance-level pragmatic inference, (2) speaker-level stance, and (3) society-level political and social meaning. Across a suite of linguistically motivated datasets, we find that leading embedding models perform only marginally better than sparse lexical baselines when implicit understanding is required.

Figure 1 summarizes this contrast. While modern embeddings achieve substantial gains over Bag-of-Tokens representations on surface-meaning benchmarks, their improvements nearly collapse on implicit semantics tasks. This performance divide highlights a fundamental mismatch between benchmark success and interpretive competence.

Our Position **We argue that text embedding research must move beyond surface-level semantics and treat implicit meaning as a first-class modeling objective.** This shift is essential for applications where semantic similarity alone is insufficient, such as stance-aware retrieval, ideological differentiation, or safety-critical filtering of content that appears benign on the surface but is implicitly harmful.

Our position is not to replace existing paradigms, but to broaden the community's understanding of what embeddings should represent. Aligning modeling objectives with linguistic theory and real-world interpretive demands can better capture the complexity of human communication.

2. Linguistic Foundations of Implicit Meaning

To clarify what we mean by *implicit meaning*, we draw on linguistic theory and organize it into a three-tier framework spanning the utterance, speaker, and society levels. We emphasize that this framework is intended as an analytical lens for organizing NLP-relevant phenomena, rather than a strict ontology or exhaustive linguistic definition of implicit meaning. This framework highlights how meaning systematically extends beyond literal content, emerging from pragmatic inference, speaker positioning, and sociocultural context.

QUESTION How do linguistic theories explain meaning that is conveyed but not explicitly stated?

Utterance Level: Pragmatic Sources of Implicit Meaning Pragmatics investigates how utterances derive meaning from context, bridging the gap between literal surface semantics and a speaker's intended message (Grice, 1975; Huang, 2017; Ma et al., 2025). Rather than focusing solely on what is explicitly said, pragmatics emphasizes what is left unsaid yet reliably understood, revealing interpretive layers that surface-level semantic analysis cannot fully capture. This view has long influenced NLP research on tasks that require contextual reasoning and inference (Hovy & Yang, 2021; Cambria, 2024).

A central insight of pragmatics is that meaning emerges from shared background knowledge, social norms, and situational context, which jointly constrain interpretation (Huang, 2017; Ma et al., 2025). Within this framework, speakers frequently rely on *implicature*–meanings inferred indirectly rather than stated outright (Grice, 1975; Potts, 2015; Hoyle et al., 2023; Ma et al., 2025). For instance, the sentence "*Bart managed to pass the test*" implies that his success was unexpected, even though this is not logically entailed.

Another key mechanism is *presupposition*, where utterances embed background assumptions required for comprehension (Potts, 2015; Ma et al., 2025). A statement like "*Sam quit smoking*" presupposes that Sam previously smoked, an

assumption that persists even under negation or questioning. Together, these phenomena illustrate how meaning depends not only on explicit content but also on what listeners infer or take for granted, posing a fundamental challenge for text embeddings that aim to model such nuance.

Speaker Level: Stance and Implicit Meaning While pragmatics focuses on utterances in context, *stance* focuses on the speaker's internal positioning: attitudes, evaluations, and degrees of alignment or commitment (Kiesling, 2022). Stance-taking is central to implicit meaning because it conveys emotional and social orientation through subtle linguistic cues rather than explicit statements.

Linguistic research characterizes stance along three dimensions: evaluation (positive or negative appraisal), alignment (social positioning relative to others), and investment (the strength of speaker commitment) (Du Bois, 2008; Lempert, 2008). These dimensions are often expressed through sociolinguistic variation. For instance, forms such as "*-in'*" versus "*-ing*" function not merely as dialectal variants, but as signals of informality, toughness, or solidarity (Kiesling, 2009; Trudgill, 1972). Over time, such forms may become detached from specific groups and reused more broadly to index stance. The evolution of the word "*dude*," from a gendered term to a marker of casual camaraderie, exemplifies this process (Kiesling, 2004).

Importantly, stance is dynamic. Quantitative studies show that speakers shift stance across discourse as they adjust intent and alignment (Kiesling et al., 2018). This introduces a relational and affective layer of meaning that complements pragmatics but remains difficult for embedding models to capture, as it depends on speaker intent rather than propositional content alone.

Society Level: Sociocultural Foundations of Implicit Meaning Beyond individual utterances and speakers, sociolinguistics examines 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–functions as a social index, signaling class, group membership, or regional identity (Silverstein, 2003). These signals are culturally contingent: the same linguistic form may index friendliness in one context and stigma in another.

Language ideologies further shape implicit meaning by privileging certain varieties while marginalizing others (Bourdieu, 1991). Because high-status registers dominate most pretraining corpora, embedding models risk encoding and amplifying existing social hierarchies–for example, by systematically disadvantaging African-American Vernacular English relative to Standard English. Speakers also engage in style-shifting, alternating registers, dialects, or slang to negotiate identity and social relationships (Bucholtz & Hall, 2005). These shifts carry implicit meaning, signaling authority, solidarity, or deference.

Static embeddings, which average across usage, struggle to capture such rapid shifts of meaning. Capturing the societal dimension of language, therefore, requires sensitivity to culturally embedded cues beyond surface form.

TAKEAWAY Implicit meaning unfolds across three interconnected layers: (1) **pragmatics** captures what is implied but unsaid at the utterance level; (2) **stance-taking** reveals the speaker's evaluative and relational positioning; and (3) **sociolinguistics** exposes how language encodes identity, culture, and power. Together, these layers underscore that meaning is deeply contextual, socially embedded, and dynamically constructed, posing a significant challenge for text embeddings still anchored in surface-level representations.

3. Text Embedding Models: Landscape and Limitations

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 reviews the evolution of embedding models, summarizes active research directions, and critically examines the field's current limitations. Figure 2 provides a high-level overview of major model classes and emerging trends that shape today's embedding landscape.

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

Early Models Early embedding approaches relied on static word vectors, such as Word2Vec and GloVe (Mikolov et al., 2013; Pennington et al., 2014), pooled into sentence-level representations. Subsequent models, including Skip-Thought (Kiros et al., 2015), InferSent (Conneau et al., 2017), Sent2Vec (Pagliardini et al., 2018), and the Universal Sentence Encoder (Cer et al., 2018), explore recurrent, transformer, or bilinear architectures to better capture sentence-level semantics. ELMo (Peters et al., 2018) marks a shift toward contextualized embeddings, dynamically encoding word meaning based on surrounding context.

Encoder-Only Models Pretrained encoder-only Transformers, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), further advanced sentence embedding by enabling context-aware representations via pooled token embeddings. These models are typically optimized using contrastive or denoising objectives (Zhang et al., 2025c). Building on this foundation, methods such as Sentence-BERT (Reimers & Gurevych, 2019), SimCSE (Gao et al., 2021), TSDAE (Wang et al., 2021), and E5 (Wang et al., 2022) substantially improved embedding quality through refined training strategies, better negative sampling, and

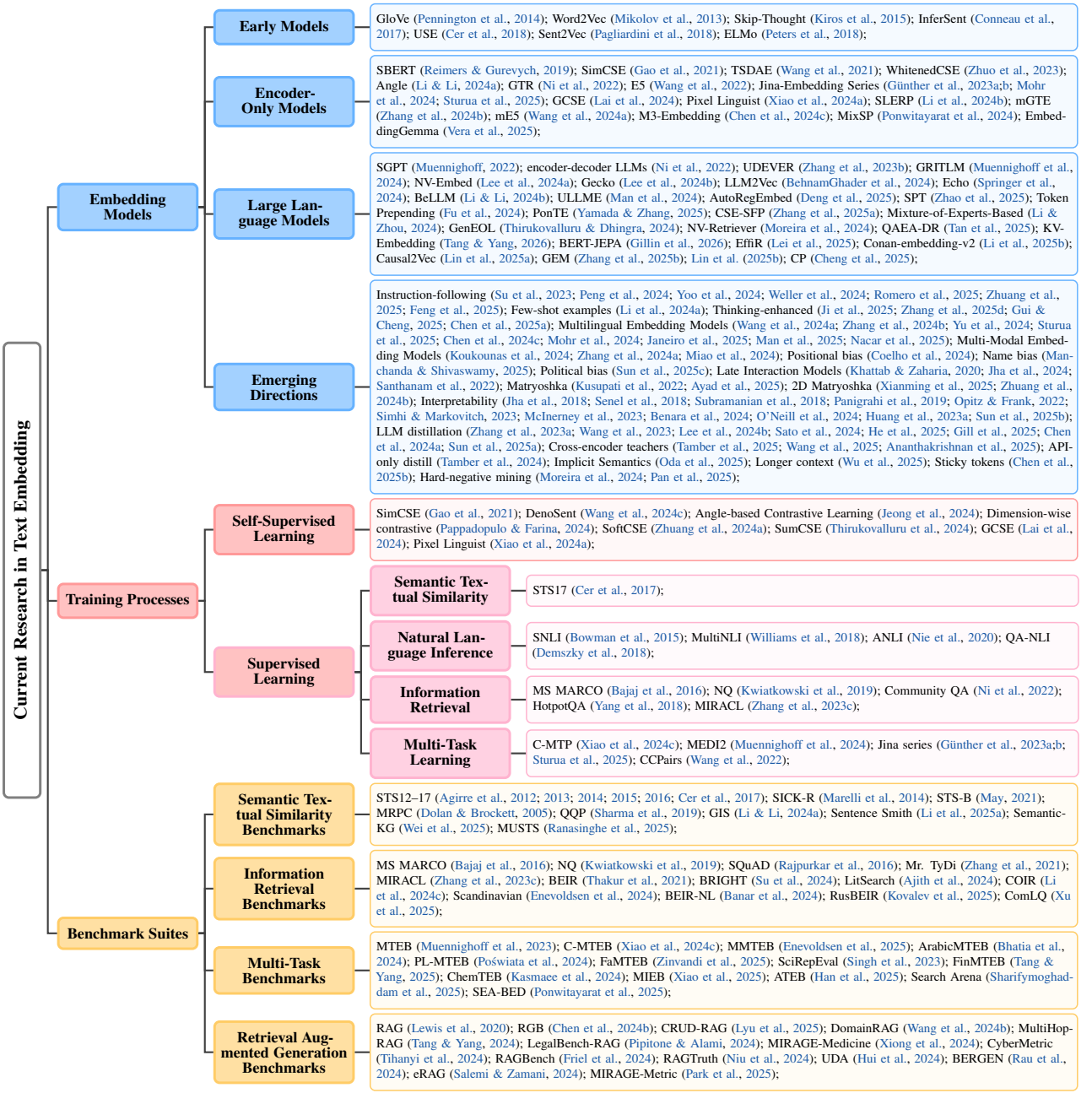

Figure 2. **A taxonomy of text embedding research:** tracing the evolution from early static models to recent trends in encoder-based architectures, LLM adaptations, multilingual embeddings, interpretability, and LLM distillation. While research has diversified, the modeling of implicit semantics remains underexplored.

architectural adjustments. As a result, encoder-only models remain the dominant choice in many retrieval and similarity-based applications due to their efficiency and strong benchmark performance (Zhuo et al., 2023; Koukounas et al., 2024; Lai et al., 2024; Xiao et al., 2024a; Li et al., 2024b; Ponwitayarat et al., 2024; Sturua et al., 2025; Chen et al., 2024c; Mohr et al., 2024; Günther et al., 2023a;b).

LLMs as Embedders More recently, LLMs have been adapted for embedding tasks using both decoder-only and encoder-decoder designs. Research has explored fine-tuning, prompting, and hybrid objectives to repurpose general-purpose LLMs for dense semantic representations. These efforts include adapting decoder-only models (Muennighoff, 2022; Cheng et al., 2025; Lin et al., 2025b), leveraging encoder-decoder architectures (Ni et al., 2022), or and building embedding-specific LLM variants (Muennighoff et al., 2024; Lee et al., 2024a;b; BehnamGhader et al., 2024;

Springer et al., 2024; Li & Li, 2024b; Man et al., 2024; Deng et al., 2025; Zhao et al., 2025; Fu et al., 2024; Yamada & Zhang, 2025; Zhang et al., 2025a; Li & Zhou, 2024; Zhang et al., 2025b). While such models often demonstrate strong semantic capabilities, converting generative LLMs into embedders typically requires adaptation with contrastive, retrieval, or similarity-based objectives, which may improve surface-semantic alignment without fully preserving the reasoning-relevant or implicit-semantic information available in the underlying model; their large size and high inference cost also sustain demand for lighter, encoder-based alternatives.

Emerging Research Directions Several trends characterize recent progress in text embeddings. Instruction-following (Su et al., 2023; Peng et al., 2024; Yoo et al., 2024; Weller et al., 2024; Feng et al., 2025; Zhuang et al., 2025; Romero et al., 2025), few-shot embedding (Li et al., 2024a), and "thinking-enhanced" representations (Ji et al., 2025) aim to improve adaptability across tasks. Multilingual and cross-lingual models (Wang et al., 2024a; Zhang et al., 2024b; Yu et al., 2024; Sturua et al., 2025; Chen et al., 2024c; Mohr et al., 2024) extend embedding utility beyond English, while new architectural designs address efficiency, political bias, and late interaction (Yoo et al., 2024; Coelho et al., 2024; Sun et al., 2025c; Khattab & Zaharia, 2020; Santhanam et al., 2022; Jha et al., 2024; Kusupati et al., 2022; Xianming et al., 2025; Zhuang et al., 2024b). Interpretability has also emerged as an important concern, with growing efforts to produce embeddings that align more closely with human-understandable concepts (Jha et al., 2018; Senel et al., 2018; Subramanian et al., 2018; Panigrahi et al., 2019; Opitz & Frank, 2022; Simhi & Markovitch, 2023; McInerney et al., 2023; Benara et al., 2024; O'Neill et al., 2024; Huang et al., 2023a; Sun et al., 2025b; Zhao et al., 2026).

Another prominent direction involves distilling knowledge from LLMs into lightweight embedding models. This includes mining hard negatives (Moreira et al., 2024; Pan et al., 2025), generating synthetic training data (Zhang et al., 2023a; Wang et al., 2023; Lee et al., 2024b; Sato et al., 2024; He et al., 2025; Gill et al., 2025), and transferring supervision from more accurate but slower teacher models (Tamber et al., 2025; Wang et al., 2025; Ananthakrishnan et al., 2025; Thirukovalluru et al., 2024; Tamber et al., 2024). While these techniques expand supervision and improve performance, they largely reinforce surface-level semantic alignment, with limited attention to implicit meaning.

TAKEAWAY Research on text embeddings spans architectures, training paradigms, multilinguality, interpretability, and efficiency. Yet, despite this rapid progress, the ability to capture implicit semantics–central to real-world language understanding–remains significantly underexplored. This gap motivates our subsequent analysis of training signals and evaluation practices, our pilot empirical study, and the research agenda outlined in the following sections.

4. Training Processes Fail to Capture Implicit Semantics

Despite architectural advances, current training signals for text embeddings overwhelmingly supervise surface-level similarity rather than implicit meaning. This section examines the two dominant paradigms: self-supervised and supervised learning, and shows how both rely on datasets and objectives that privilege lexical overlap, paraphrastic equivalence, or relevance matching, leaving deeper pragmatic, attitudinal, and social meanings largely unmodeled.

QUESTION How are text embedding models trained, and why do these methods fail to capture implicit meaning?

4.1. Self-Supervised Learning

Self-supervised learning extracts training signals directly from unlabeled text using augmentation or structural cues, without requiring manual labels. Representative approaches include SimCSE (Gao et al., 2021), which creates positive pairs via dropout noise, and denoising-based objectives such as DenoSent (Wang et al., 2024c). More recent variants explore alternative formulations, including angle-based objectives (Jeong et al., 2024), dimension-wise contrastive loss (Pappadopulo & Farina, 2024), and similarity-weighted negative sampling (Zhuang et al., 2024a).

While self-supervised methods are attractive due to their scalability and low annotation cost, they consistently underperform supervised approaches in semantic benchmarks. As a result, most practical embedding pipelines adopt a two-stage strategy: self-supervised pretraining followed by supervised fine-tuning (Gao et al., 2021). Crucially, because the self-supervised signal is derived from surface perturbations of the same text, it primarily reinforces invariance to form rather than sensitivity to implicit intent or context.

4.2. Supervised Learning

Supervised embedding models use contrastive objectives such as triplet, SimCSE, or angle-based losses (Reimers & Gurevych, 2019; Gao et al., 2021; Li & Li, 2024a) on pretrained language models, but rely on task-specific datasets, primarily Semantic Textual Similarity (STS), Natural Language Inference (NLI), and Information Retrieval (IR), due to the scarcity of labeled pairs in general-purpose corpora (Raffel et al., 2020).

Semantic Textual Similarity (STS) STS datasets provide fine-grained similarity scores and have been widely used in early sentence embedding models (Reimers & Gurevych, 2019; Wang et al., 2021). However, their limited scale and

narrow domain coverage often lead to overfitting and weak generalization (Muennighoff et al., 2023). More importantly, STS annotations primarily reflect surface-level paraphrasing rather than deeper interpretive meaning.

Natural Language Inference (NLI) NLI datasets label sentence pairs as entailment, contradiction, or neutral and offer greater scale and diversity (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020; Demszky et al., 2018). They are widely used in modern embedding models (Reimers & Gurevych, 2019; Wang et al., 2021; 2022; Zhang et al., 2023b; Li & Li, 2024a; Zhang et al., 2023b). However, the semantic signal is often shallow: many entailment pairs differ only lexically or syntactically. For instance, pairs like "*A boy is jumping on a skateboard*" and "*The boy does a skateboarding trick*" test literal equivalence rather than pragmatic intent (Bowman et al., 2015).

Information Retrieval (IR) Retrieval datasets such as MS MARCO (Bajaj et al., 2016) and Natural Questions (NQ) (Kwiatkowski et al., 2019) dominate large-scale embedding training (Ni et al., 2022; Wang et al., 2022; Muennighoff et al., 2024; Moreira et al., 2024; Zhang et al., 2024b). These datasets are effective for learning lexical relevance and topic matching, but they reward literal overlap between queries and documents. As a result, embeddings trained on IR objectives struggle to encode implicit cues such as stance, ideology, sarcasm, or social framing.

Multi-Task Learning To improve robustness, recent models such as mGTE (Zhang et al., 2024b) and Jina (Günther et al., 2023a;b; Sturua et al., 2025) adopt multi-task training that combines STS, NLI, IR, QA, and related datasets (Xiao et al., 2024c; Muennighoff et al., 2024). While this broadens task and domain coverage, supervision remains largely surface-oriented, with few examples probing pragmatic inference, speaker stance, or sociocultural meaning.

TAKEAWAY Across both self-supervised and supervised paradigms, current embedding training pipelines are anchored in datasets that prioritize surface-level similarity. Although multi-task learning increases diversity, it does not fundamentally change what models are optimized to learn. As a result, core components of implicit meaning, i.e., pragmatics, stance, and social context, remain weakly represented or entirely missing from training objectives.

5. Benchmarks Do Not Evaluate Implicit Semantics

Despite the rapid expansion of large-scale benchmark suites–from semantic similarity and retrieval to multi-task generalization–most evaluations remain focused on surface-level semantics. This section surveys widely used STS datasets, retrieval, multi-task, and RAG benchmarks, highlighting a persistent gap: the limited evaluation of implicit, contextual, and socially situated meaning.

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

STS Benchmarks evaluate how well model-predicted similarities align with human judgments, typically using correlation-based metrics. Popular datasets include STS12–17 (Agirre et al., 2012; 2013; 2014; 2015; 2016; Cer et al., 2017), STS-B (May, 2021), and SICK-R (Marelli et al., 2014), along with related paraphrase classification tasks such as MRPC (Dolan & Brockett, 2005), QQP (Sharma et al., 2019), and GIS (Li & Li, 2024a). While STS tasks can probe deeper meaning, they are largely constrained to lexical and syntactic variation, rarely testing pragmatic, attitudinal, or cultural interpretation.

IR Benchmarks IR benchmarks evaluate how effectively embeddings retrieve relevant documents using ranking metrics such as MRR, nDCG, and recall@k (Wang et al., 2013; Thakur et al., 2021). Datasets like MS MARCO (Bajaj et al., 2016), Natural Questions (Kwiatkowski et al., 2019), SQuAD (Rajpurkar et al., 2016), Mr. TyDi (Zhang et al., 2021), and MIRACL (Zhang et al., 2023c) are commonly used, with BEIR (Thakur et al., 2021) aggregating many such tasks across diverse retrieval settings. More recent benchmarks extend coverage to new domains and languages (Su et al., 2024; Ajith et al., 2024; Li et al., 2024c; Enevoldsen et al., 2024; Banar et al., 2024; Kovalev et al., 2025). Despite this breadth, IR benchmarks focus on surface relevance and seldom evaluate alignment with implicit criteria such as stance or ideology. Retrieval based on speaker stance or ideological framing remains largely unexplored.

Multi-Task Benchmarks MTEB (Muennighoff et al., 2023) seldom test whether retrieved documents align with implicit criteria such as stance, ideological framing, or communicative intent (Xiao et al., 2024c; Enevoldsen et al., 2025; Bhatia et al., 2024; Poświata et al., 2024; Zinvandi et al., 2025; Singh et al., 2023; Tang & Yang, 2025; Kasmaee et al., 2024; Xiao et al., 2025). More challenging variants introduce reasoning or instruction-following tasks (Han et al., 2025), and crowdsourced platforms like MTEB Arena[1] and Search Arena[2] provide user-driven comparisons (Sharifymoghaddam et al., 2025). Despite their scale and flexibility, these benchmarks rely largely on surface-level metrics and datasets. Only a small subset probes beyond lexical meaning, so strong MTEB performance often reflects surface alignment rather than sensitivity to implicit meaning.

RAG Benchmarks RAG benchmarks assess how well embeddings support retrieval for downstream generation.

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

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

Existing benchmarks cover multilingual, domain-specific, and multi-hop scenarios (Chen et al., 2024b; Lyu et al., 2025; Wang et al., 2024b; Tang & Yang, 2024; Pipitone & Alami, 2024; Xiong et al., 2024; Tihanyi et al., 2024), and include tools for evaluating retrieval quality and hallucination (Friel et al., 2024; Niu et al., 2024; Hui et al., 2024; Rau et al., 2024; Salemi & Zamani, 2024; Park et al., 2025). While these setups introduce more complex pipelines, the underlying retrieval objectives remain largely factual. 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 benchmarks provide extensive 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 ground our argument empirically and motivate future research, we conduct a pilot study examining whether state-of-the-art embedding models capture implicit meaning across three linguistic tiers: utterance, speaker, and society.

Experimental Setup We evaluate embeddings on seven datasets spanning three levels of implicit semantics: (1) Utterance level: Pragmatics Understanding Benchmark (PUB), including Implicature (**P-IMP**), Presupposition (**P-PRE**), and Reference & Deixis (**P-R&D**) (Sravanthi et al., 2024; Louis et al., 2020; Zheng et al., 2021; Liu et al., 2022; Chakrabarty et al., 2022; Jeretic et al., 2020; Parrish et al., 2021); (2) Speaker level: **P-Stance** dataset (Li et al., 2021) for stance detection; and (3) Society level: the datasets of Implicit Hate Speech (**IHS**) (ElSherief et al., 2021), Social Bias Inference Corpus (**SBIC**) (Sap et al., 2020), and Political Bias (**Pol. Bias**) (Baly et al., 2020). To provide a structured comparison, we report mean performance across subtasks within each semantic tier.

Since these datasets were not originally designed for embedding evaluation, we reformulate them into classification, pairwise classification (following the MTEB protocol (Muennighoff et al., 2023)), and zero-shot similarity settings, where models select labels based on embedding similarity. We evaluate four representative model families: (1) encoder-only models, (2) LLM-based embeddings, (3) multimodal encoders, and (4) proprietary embeddings (OpenAI). As baselines, we include a Bag-of-Tokens lexical model (Harris, 1954; Sun et al., 2025b) and a random predictor (Random) as a hardness baseline. Full implementation details are provided in Appendix A.

Results and Analysis As depicted in Table 1, the results reveal a consistent and striking pattern. Encoder-only models typically perform only marginally better than the Bag-of-Tokens baseline and, in some cases, approach random performance. In contrast, LLM-based models and proprietary embeddings achieve stronger overall results. Notably, although OpenAI embeddings rank lower on standard benchmarks such as MTEB, they perform comparatively well on implicit semantics tasks, suggesting a disconnect between benchmark success and deeper semantic competence.

Performance also varies substantially across semantic tiers. Linq-Mistral excels at utterance-level pragmatic tasks, OpenAI-Large performs best on speaker- and society-level datasets, and E5-Mistral shows particular strength in political bias detection. These differences suggest that current models specialize unevenly across dimensions of implicit meaning, rather than learning a unified representation.

These positive results also clarify the nature of the limitation. Current embedding models are not uniformly poor on all implicit semantics tasks; rather, they perform better when the relevant implicit meaning is strongly associated with local lexical, syntactic, discourse, or label-specific cues. This is especially plausible for relatively conventionalized pragmatic phenomena, where large-scale pretraining and instruction tuning may help models recover recurring surface patterns correlated with the intended label. Thus, the issue is not complete failure, but uneven generalization: current embeddings appear to capture some shallow or partially lexicalized forms of implicit meaning, while remaining less robust on cases that require richer contextual grounding, speaker modeling, or socially situated interpretation.

These results support our central claim: **state-of-the-art embedding models remain limited in their ability to capture implicit semantics.** High MTEB scores do not translate into robustness on tasks involving pragmatic inference, stance recognition, or social meaning. The shrinking gains over Bag-of-Tokens, especially on the hardest pragmatic tasks, underscore a fundamental gap between how embeddings are evaluated and how meaning is actually constructed in language. Detailed results are provided in Appendix B.

7. Towards Embeddings that Capture Implicit Meaning

Building on the empirical findings in Section 6, we outline three directions for advancing text embeddings beyond surface-level semantics: (1) linguistically and culturally diverse training data, (2) benchmarks that directly evaluate pragmatic, attitudinal, and social understanding, and (3) treating implicit semantics as a core modeling objective. Together, these directions realign embedding research with the realities of human language interpretation.

Table 1. **Average accuracy (%) of representative embedding models** on seven datasets spanning three tiers of implicit semantics: utterance-level pragmatics, speaker-level stance, and societal-level social meaning. Results reveal substantial variation across semantic tiers and demonstrate that strong performance on standard benchmarks does not translate to robust modeling of implicit meaning.

Model	Utterance Level			Speaker Level	Society Level			Average Accuracy ↑
	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 (Sun et al., 2025b)	56.5	75.3	48.2	73.4	59.6	80.7	41.6	62.2
Encoder-only Models								
S-BERT (Reimers & Gurevych, 2019)	61.7	72.8	55.7	72.9	60.8	81.8	47.9	64.8
GIST-Small (Solatorio, 2024)	65.8	76.1	58.8	76.0	61.8	81.6	49.1	67.0
BGE-Base (Xiao et al., 2024b)	65.0	75.6	57.3	74.4	62.9	82.1	52.1	67.1
Angle (Li & Li, 2024a)	69.4	78.8	57.2	76.4	59.5	83.7	50.4	67.9
BGE-Large (Xiao et al., 2024b)	68.3	75.5	58.1	76.0	63.5	83.4	51.5	68.0
MXBAI-Large (Lee et al., 2024c)	69.8	78.2	59.4	75.6	60.1	83.6	50.5	68.2
GIST-Large (Solatorio, 2024)	68.8	76.6	62.9	76.7	64.2	83.4	52.5	69.3
Stella (Zhang et al., 2024a)	72.1	81.5	59.6	76.5	60.4	84.0	54.4	69.8
LLM-based Embeddings								
Linq-Mistral (Kim et al., 2024)	**80.3**	**87.7**	**70.4**	75.8	61.4	82.0	56.8	73.5
E5-Mistral (Wang et al., 2023; 2022)	78.1	81.8	63.4	81.1	63.9	84.8	**71.5**	74.9
GTE-Qwen (Li et al., 2023b)	73.4	87.3	68.1	80.9	65.6	84.5	66.8	**75.2**
Multimodal Encoders								
Jasper (Zhang et al., 2024a)	73.3	80.1	63.0	80.1	65.7	84.2	63.9	72.9
Proprietary Embeddings								
OpenAI-Small (Neelakantan et al., 2022)	71.3	78.1	64.3	80.0	66.2	83.9	56.6	71.5
OpenAI-Large (Neelakantan et al., 2022)	76.0	80.2	66.4	**83.7**	**67.1**	**85.4**	66.3	75.0

7.1. Curating More Diverse Training Data

Training data ultimately determines what embedding models can learn. When supervision emphasizes surface-level signals, representations inevitably mirror surface meaning. Capturing implicit semantics, therefore, requires expanding beyond narrow, convenience-driven datasets toward linguistically richer and culturally diverse sources.

Importantly, by more diverse training data, we do not simply mean scaling raw web text. The key need is more diverse supervision for embedding learning. Unlike autoregressive LMs, embedding models are usually trained with pairwise, contrastive, retrieval-style, or classification-oriented signals. For implicit semantics, such supervision should include cases where surface meaning is similar but implied meaning differs, such as contrastive examples involving implicature, presupposition, stance, indirectness, sarcasm, social framing, dialectal variation, or ideology.

Recent progress in LLM-based data generation offers a practical path forward. Prior work shows that LLMs can synthesize effective supervision for embedding training, often by enforcing semantic invariance or contrast (Wang et al., 2023; Chen et al., 2024a; Lippmann & Yang, 2025; Truong et al., 2025). Future efforts should target implicit phenomena such as implicature, presupposition, stance, and indirect evaluation, rather than treating them as incidental variation. Such data can be curated from dialogue, indirect answers, stance-rich discourse, and socially situated language, or synthesized and relabeled using LLMs or stronger cross-encoder teachers guided by linguistic theory.

Linguistic theory provides a rich foundation for this shift. Decades of research have formalized typologies of implicit meaning across pragmatic and sociocultural dimensions. Aligning synthetic and curated data with these frameworks can help embedding models internalize forms of meaning that are largely absent from existing training corpora.

7.2. Designing Benchmarks for Implicit Meaning

Benchmarks shape research priorities by defining what success looks like. Existing suites, most notably MTEB, have played an important role in standardizing evaluation, but they overwhelmingly emphasize surface similarity. Their

open and leaderboard-driven nature has also led to data leakage and score inflation, weakening their ability to measure generalization (Chung et al., 2025; Sancheti et al., 2025).

Empirical studies show that strong MTEB results do not reliably translate to downstream robustness and may even correlate negatively with performance on certain tasks. This trend reflects a broader shift toward optimizing benchmark artifacts rather than transferable semantic understanding.

To address this gap, new benchmarks should explicitly target implicit meaning. This includes tasks that require inference from indirect cues, recognition of speaker stance, sensitivity to sociolinguistic variation, and interpretation of socially situated language. Without such benchmarks, progress on implicit semantics will remain invisible and undervalued.

7.3. Framing Implicit Semantics as a Modeling Goal

A deeper issue is that implicit meaning is rarely treated as a first-class objective in embedding research. While work on LLMs increasingly targets pragmatic reasoning, social understanding, and discourse-level interpretation (Li et al., 2020; 2023a; Kazemi et al., 2023; Sravanthi et al., 2024; Yue et al., 2024; Curry et al., 2024; Sun et al., 2024; 2025b; Ma et al., 2025), embedding models continue to be optimized for objectives that reward superficial alignment.

This misalignment pushes models to optimize what is easy to measure rather than what is meaningful to understand, reinforcing shallow representations in the absence of explicit objectives for implicit semantics. Reframing implicit semantics as a core modeling goal can better align embeddings with how humans interpret language in context.

Concretely, this goal can be operationalized through objectives that distinguish texts with similar explicit content but different implied meanings, multi-task supervision over pragmatics, stance, and social meaning, or distillation from models that explicitly reason over implicit interpretations. The goal is not to make embeddings preserve all information in the original text. Rather, it is to preserve task-relevant implicit signals that matter for downstream applications such as stance-aware retrieval, socially grounded classification, recommendation, clustering, filtering, and reranking.

Instruction-following retrieval is a particularly relevant step in this direction because it conditions embeddings on user intent and task framing rather than only surface similarity. Recent work on instruction-following embeddings and retrieval models (Su et al., 2023; Peng et al., 2024; Weller et al., 2024; Feng et al., 2025; Zhuang et al., 2025) shows that embedding research is already moving toward richer task-conditioned representations.

We view this line of work as a promising bridge between standard semantic embeddings and implicit-semantics-aware representations. Nevertheless, following retrieval instructions does not by itself guarantee sensitivity to broader implicit phenomena such as stance, presupposition, sociocultural meaning, or social indexicality.

8. Alternative Views

It is reasonable to argue that surface-level semantics suffice for many practical applications, such as search, recommendation, or clustering. In these settings, modeling deeper meaning may introduce unnecessary complexity and additional training or evaluation costs without clear returns. Another perspective holds that implicit, pragmatic, and socially grounded meaning is better handled by LLMs designed for contextual reasoning, while embeddings should prioritize efficiency and general-purpose utility. From this perspective, expanding the embedding scope risks blurring their role.

We acknowledge these positions. Our argument is not that embeddings should replace decoder-only LMs for complex reasoning, nor that all implicit understanding must be solved purely in vector space. Rather, embeddings often serve as first-stage representations in practical systems, where they determine which examples are retrieved, clustered, recommended, filtered, or passed to downstream models. In such settings, if important signals such as stance, intent, or social framing are not preserved in the embedding space, later components may never receive the relevant evidence.

Thus, as embeddings increasingly serve as semantic interfaces for downstream decision-making, their limitations in capturing implicit meaning become harder to ignore, especially in applications where stance, intent, or social interpretation are central.

9. Conclusions

Despite rapid progress, contemporary text embedding models are predominantly optimized for surface-level semantics, exhibiting a fundamental limitation in capturing the implicit meanings that are essential to nuanced human communication. Grounded in a three-tier linguistic framework: spanning utterance-level pragmatics, speaker-level stance, and society-level sociocultural context, our empirical analysis reveals that even state-of-the-art models show only marginal gains over simple lexical baselines on tasks probing these deeper layers of interpretation. These findings expose a structural mismatch between training, evaluation, and real-world language use. To bridge this gap, we advocate for semantically richer data, benchmarks that explicitly test implicit understanding, and modeling objectives that treat implicit semantics as a first-class goal. Aligning embedding research with these dimensions of real-world language use is crucial for developing robust, context-aware models that power the next generation of language-aware applications.

Acknowledgements

Qiang Huang was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123302) and the National Natural Science Foundation of China (NSFC) under Grant No. U25B6003. Jun Yu was supported by the NSFC under Grant Nos. 62125201 and U24B20174. Yiqun Sun and Anthony T. H. Tung were supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 1 Grant (T1 251RES2517) and the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-029). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore, or the National Research Foundation, Singapore.

References

Agirre, E., Cer, D., Diab, M., and Gonzalez-Agirre, A. 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)*, pp. 385–393, 2012. URL https://aclanthology.org/S12-1051/.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. *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.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 81–91, 2014.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. 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)*, pp. 252–263, 2015.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 497–511, 2016.

Ajith, A., Xia, M., Chevalier, A., Goyal, T., Chen, D., and Gao, T. Litsearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 15068–15083, 2024.

Ananthakrishnan, H., Dolby, J., Kokel, H., Samulowitz, H., and Srinivas, K. Can cross encoders produce useful sentence embeddings? *arXiv preprint arXiv:2502.03552*, 2025.

Angelov, D. and Inkpen, D. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13528–13539, 2024.

Ayad, M. A. B., Dinzinger, M., Dastidar, K. G., Mitrovic, J., and Granitzer, M. Compressed concatenation of small embedding models. *arXiv preprint arXiv:2510.04626*, 2025.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., and Wang, T. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Baly, R., Da San Martino, G., Glass, J., and Nakov, P. 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)*, pp. 4982–4991, 2020.

Banar, N., Lotfi, E., and Daelemans, W. Beir-nl: Zero-shot information retrieval benchmark for the dutch language. *arXiv preprint arXiv:2412.08329*, 2024.

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., and Reddy, S. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

Benara, V., Singh, C., Morris, J. X., Antonello, R., Stoica, I., Huth, A., and Gao, J. Crafting interpretable embeddings for language neuroscience by asking LLMs questions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 124137, 2024.

Bhatia, G., Nagoudi, E. M. B., Mekki, A. E., Alwajih, F., and Abdul-Mageed, M. Swan and arabicmteb: Dialect-aware, arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks. *arXiv preprint arXiv:2411.01192*, 2024.

Bourdieu, P. Language and symbolic power. *Polity*, 1991.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 632–642, 2015.

Bucholtz, M. and Hall, K. Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, 7 (4-5):585–614, 2005.

Cambria, E. Pragmatics processing. In *Understanding Natural Language Understanding*, pp. 229–338. 2024.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pp. 169–174, 2018.

Chakrabarty, T., Saakyan, A., Ghosh, D., and Muresan, S. Flute: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7139–7159, 2022.

Chen, H., Wang, L., Yang, N., Zhu, Y., Zhao, Z., Wei, F., and Dou, Z. Little giants: Synthesizing high-quality embedding data at scale. *arXiv preprint arXiv:2410.18634*, 2024a.

Chen, J., Lin, H., Han, X., and Sun, L. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 17754–17762, 2024b.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024c.

Chen, J., Lan, J., Li, C., Lian, D., and Liu, Z. Reasonembed: Enhanced text embeddings for reasoning-intensive document retrieval. *arXiv preprint arXiv:2510.08252*, 2025a.

Chen, K., Wang, D., Liu, Y., Zhang, H., and Wang, W. Sticking to the mean: Detecting sticky tokens in text embedding models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28660–28681, 2025b.

Cheng, Z., Wang, Z., Fu, Y., Jiang, Z., Yin, Y., Wang, C., and Gu, Q. Contrastive prompting enhances sentence embeddings in llms through inference-time steering. *arXiv preprint arXiv:2505.12831*, 2025.

Chung, I., Kerboua, I., Kardos, M., Solomatin, R., and Enevoldsen, K. Maintaining mteb: Towards long term usability and reproducibility of embedding benchmarks. *arXiv preprint arXiv:2506.21182*, 2025.

Coelho, J., Martins, B., Magalhães, J., Callan, J., and Xiong, C. 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)*, pp. 370–377, 2024.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. 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)*, pp. 670–680, 2017.

Curry, A. C., Attanasio, G., Talat, Z., and Hovy, D. Classist tools: Social class correlates with performance in nlp. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12643–12655, 2024.

Dai, S., Huang, Q., You, X., and Yu, J. MG^2-RAG: Multi-Granularity Graph for Multimodal Retrieval-Augmented Generation. *arXiv preprint arXiv:2604.04969*, 2026.

Demszky, D., Guu, K., and Liang, P. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*, 2018.

Deng, J., Jiang, Z., Pang, L., Chen, L., Xu, K., Wei, Z., Shen, H., and Cheng, X. Following the autoregressive nature of llm embeddings via compression and alignment. *arXiv preprint arXiv:2502.11401*, 2025.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.

Dolan, B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*, 2005.

Du Bois, J. W. The stance triangle. In *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, pp. 139–182. 2008.

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., and Yang, D. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 345–363, 2021.

Enevoldsen, K., Kardos, M., Muennighoff, N., and Nielbo, K. L. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:40336–40358, 2024.

Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemiński, D., Winata, G. I., et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

Feng, Y., Sun, Y., Sun, Y., Zhu, M., Huang, Q., Tung, A. K. H., and Chen, W. Don't reinvent the wheel: Efficient instruction-following text embedding based on guided space transformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24511–24525, July 2025. doi: 10.18653/v1/2025.acl-long.1196.

Friel, R., Belyi, M., and Sanyal, A. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*, 2024.

Fu, Y., Cheng, Z., Jiang, Z., Wang, Z., Yin, Y., Li, Z., and Gu, Q. Token prepending: A training-free approach for eliciting better sentence embeddings from llms. *arXiv preprint arXiv:2412.11556*, 2024.

Gao, T., Yao, X., and Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6894–6910, 2021.

Gill, W., Cechmanek, J., Hutcherson, T., Rajamohan, S., Agarwal, J., Gulzar, M. A., Singh, M., and Dion, B. Advancing semantic caching for llms with domain-specific embeddings and synthetic data. *arXiv preprint arXiv:2504.02268*, 2025.

Gillin, T., Lalani, A., Zhang, K., and Salles, M. M. Bert-jepa: Reorganizing cls embeddings for language-invariant semantics. *arXiv preprint arXiv:2601.00366*, 2026.

Grice, H. P. Logic and conversation. In *Speech Acts*, pp. 41–58. Brill, 1975.

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

Gui, Y. and Cheng, J. Search-r3: Unifying reasoning and embedding generation in large language models. *arXiv preprint arXiv:2510.07048*, 2025.

Günther, M., Milliken, L., Geuter, J., Mastrapas, G., Wang, B., and Xiao, H. 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 2023a.

Günther, M., Ong, J., Mohr, I., Abdessalem, A., Abel, T., Akram, M. K., Guzman, S., Mastrapas, G., Sturua, S., Wang, B., et al. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*, 2023b.

Han, S., Gomez, F. P., Vu, T., Li, Z., Cer, D., Zeng, H., Tar, C., Cohan, A., and Abrego, G. H. Ateb: Evaluating and improving advanced nlp tasks for text embedding models. *arXiv preprint arXiv:2502.16766*, 2025.

Harris, Z. S. Distributional structure. *Word*, 10(2-3):146–162, 1954.

He, L., Liu, C., Li, R., Huang, Z., Ruan, S., Zhou, J., and Chen, E. Refining sentence embedding model through ranking sentences generation with large language models. *arXiv preprint arXiv:2502.13656*, 2025.

Hovy, D. and Yang, D. 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)*, pp. 588–602, 2021.

Hoyle, A., Sarkar, R., Goel, P., and Resnik, P. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13188–13214, 2023.

Huang, J., Yao, W., Song, K., Zhang, H., Chen, M., and Yu, D. 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)*, pp. 14584–14595, 2023a.

Huang, Q., Luo, P., and Tung, A. K. A new sparse data clustering method based on frequent items. *Proceedings of the ACM on Management of Data*, 1(1):1–28, 2023b.

Huang, Q., Wang, Y., Sun, Y., and Tung, A. K. Diversity-aware k-maximum inner product search revisited. *arXiv preprint arXiv:2402.13858*, 2024.

Huang, Y. Introduction: What is pragmatics? In Huang, Y. (ed.), *The Oxford Handbook of Pragmatics*, Oxford Handbooks. Oxford University Press, 2017. URL https://doi.org/10.1093/oxfordhb/9780199697960.013.33.

Hui, Y., Lu, Y., and Zhang, H. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. *arXiv preprint arXiv:2406.15187*, 2024.

Janeiro, J. M., Alastruey, B., Massa, F., Elbayad, M., Piwowarski, B., Gallinari, P., and Barrault, L. Mixture of languages: Improved multilingual encoders through language grouping. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29695–29710, 2025.

Jeong, Y. H., Han, M., and Chae, D.-K. A simple angle-based approach for contrastive learning of unsupervised sentence representation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5553–5572, 2024.

Jeretic, P., Warstadt, A., Bhooshan, S., and Williams, A. Are natural language inference models imppressive? learning implicature and presupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4437–4452, 2020.

Jha, K., Wang, Y., Xun, G., and Zhang, A. Interpretable word embeddings for medical domain. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 1061–1066, 2018.

Jha, R., Wang, B., Günther, M., Mastrapas, G., Sturua, S., Mohr, I., Koukounas, A., Akram, M. K., Wang, N., and Xiao, H. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. *arXiv preprint arXiv:2408.16672*, 2024.

Ji, Y., Xu, Z., Liu, Z., Yan, Y., Yu, S., Li, Y., Liu, Z., Gu, Y., Yu, G., and Sun, M. Learning more effective representations for dense retrieval through deliberate thinking before search. *arXiv preprint arXiv:2502.12974*, 2025.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

Kasmaee, A. S., Khodadad, M., Saloot, M. A., Sherck, N., Dokas, S., Mahyar, H., and Samiee, S. Chemteb: Chemical text embedding benchmark, an overview of embedding models performance & efficiency on a specific domain. *arXiv preprint arXiv:2412.00532*, 2024.

Kazemi, M., Yuan, Q., Bhatia, D., Kim, N., Xu, X., Imbrasaite, V., and Ramachandran, D. Boardgameqa: A dataset for natural language reasoning with contradictory information. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:39052–39074, 2023.

Khattab, O. and Zaharia, M. 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)*, pp. 39–48, 2020.

Kiesling, S. F. Dude. *American Speech*, 79(3):281–305, 2004.

Kiesling, S. F. Style as stance: Stance as the explanation for patterns of sociolinguistic variation. In *Stance: Sociolinguistic Perspectives*, pp. 171–194. 2009.

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

Kiesling, S. F., Pavalanathan, U., Fitzpatrick, J., Han, X., and Eisenstein, J. Interactional stancetaking in online forums. *Computational Linguistics*, 44(4):683–718, 2018.

Kim, J., Lee, S., Kwon, J., Gu, S., Kim, Y., Cho, M., yong Sohn, J., and Choi, C. Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog, 2024. URL https://getlinq.com/blog/linq-embed-mistral/.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. Skip-thought vectors. In *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS)*, pp. 3294–3302, 2015.

Koukounas, A., Mastrapas, G., Günther, M., Wang, B., Martens, S., Mohr, I., Sturua, S., Akram, M. K., Martínez, J. F., Ognawala, S., et al. Jina clip: Your clip model is also your text retriever. *arXiv preprint arXiv:2405.20204*, 2024.

Kovalev, G., Tikhomirov, M., Kozhevnikov, E., Kornilov, M., and Loukachevitch, N. Building russian benchmark for evaluation of information retrieval models. *arXiv preprint arXiv:2504.12879*, 2025.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:30233–30249, 2022.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics (TACL)*, 7:453–466, 2019.

Lai, P., Zhang, Z., Zhang, W., Fu, F., and Cui, B. Enhancing unsupervised sentence embeddings via knowledge-driven data augmentation and gaussian-decayed contrastive learning. *arXiv preprint arXiv:2409.12887*, 2024.

Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., and Ping, W. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024a.

Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J. R., Hui, K., Boratko, M., Kapadia, R., Ding, W., et al. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*, 2024b.

Lee, S., Shakir, A., Koenig, D., and Lipp, J. Open source strikes bread - new fluffy embeddings model, 2024c. URL https://www.mixedbread.ai/blog/mxbai-embed-large-v1.

Lei, Y., He, S., Li, A., and Yates, A. Making large language models efficient dense retrievers. *arXiv preprint arXiv:2512.20612*, 2025.

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

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 9459–9474, 2020.

Li, C., Qin, M., Xiao, S., Chen, J., Luo, K., Shao, Y., Lian, D., and Liu, Z. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700*, 2024a.

Li, H., Zhu, S.-C., and Zheng, Z. Diplomat: a dialogue dataset for situated pragmatic reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 46856–46884, 2023a.

Li, H., Michail, A., Gubelmann, R., Clematide, S., and Opitz, J. Sentence smith: Controllable edits for evaluating text embeddings. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26439–26456, 2025a.

Li, J., Liu, M., Kan, M.-Y., Zheng, Z., Wang, Z., Lei, W., Liu, T., and Qin, B. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pp. 2642–2652, 2020.

Li, L., Huang, Q., Ang, Y., Low, B. K. H., Tung, A. K., and Xiao, X. Weight-informed self-explaining clustering for mixed-type tabular data. *arXiv preprint arXiv:2604.05857*, 2026.

Li, M., Nie, Z., Zhang, Y., Long, D., Zhang, R., and Xie, P. Improving general text embedding model: Tackling task conflict and data imbalance through model merging. *arXiv preprint arXiv:2410.15035*, 2024b.

Li, S., Tang, Y., Liu, R., Chen, S.-Z., and Chen, X. Conan-embedding-v2: Training an llm from scratch for text embeddings. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15011–15027, 2025b.

Li, X. and Li, J. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1825–1839, 2024a.

Li, X. and Li, J. 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)*, pp. 792–804, 2024b.

Li, X., Dong, K., Lee, Y. Q., Xia, W., Zhang, H., Dai, X., Wang, Y., and Tang, R. Coir: A comprehensive benchmark for code information retrieval models. *arXiv preprint arXiv:2407.02883*, 2024c.

Li, Y., Sosea, T., Sawant, A., Nair, A. J., Inkpen, D., and Caragea, C. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL 2021*, pp. 2355–2365, 2021.

Li, Z. and Zhou, T. Your mixture-of-experts llm is secretly an embedding model for free. *arXiv preprint arXiv:2410.10814*, 2024.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023b.

Lin, A., Li, Z., Funakoshi, K., and Okumura, M. Causal2vec: Improving decoder-only llms as versatile embedding models. *arXiv preprint arXiv:2507.23386*, 2025a.

Lin, Z., Wu, H., Wang, S., Tu, K., Zheng, Z., and Jia, Z. Look both ways and no sink: Converting LLMs into text encoders without training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22839–22853, July 2025b.

Lippmann, P. and Yang, J. Zero-shot contextual embeddings via offline synthetic corpus generation. *arXiv preprint arXiv:2506.23662*, 2025.

Liu, E., Cui, C., Zheng, K., and Neubig, G. 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)*, pp. 4437–4452, 2022.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Louis, A., Roth, D., and Radlinski, F. "i'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7411–7425, 2020.

Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T., and Chen, E. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*, 43(2):1–32, 2025.

Ma, B., Li, Y., Zhou, W., Gong, Z., Liu, Y. J., Jasinskaja, K., Friedrich, A., Hirschberg, J., Kreuter, F., and Plank, B. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *arXiv preprint arXiv:2502.12378*, 2025.

Man, H., Ngo, N. T., Dernoncourt, F., and Nguyen, T. H. Ullme: A unified framework for large language model embeddings with generation-augmented learning. *arXiv preprint arXiv:2408.03402*, 2024.

Man, H., Ngo, N. T., Dac Lai, V., Rossi, R. A., Dernoncourt, F., and Huu Nguyen, T. Lusifer: Language universal space integration for enhanced representation in multilingual text embedding models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1360–1370, 2025.

Manchanda, S. and Shivaswamy, P. What is in a name? mitigating name bias in text embedding similarity via anonymization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17759–17781, 2025.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pp. 216–223, 2014.

May, P. Machine translated multilingual sts benchmark dataset., 2021. URL `https://github.com/PhilipMay/stsb-multi-mt`.

McInerney, D., Young, G., van de Meent, J.-W., and Wallace, B. CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8477–8494, 2023.

Miao, Z., Wu, Q., Zhao, K., Wu, Z., and Tsuruoka, Y. Enhancing cross-lingual sentence embedding for low-resource languages with word alignment. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3225–3236, 2024.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*, pp. 3111–3119, 2013.

Mohr, I., Krimmel, M., Sturua, S., Akram, M. K., Koukounas, A., Günther, M., Mastrapas, G., Ravishankar, V., Martínez, J. F., Wang, F., et al. Multi-task contrastive learning for 8192-token bilingual text embeddings. *arXiv preprint arXiv:2402.17016*, 2024.

Moreira, G. d. S. P., Osmulski, R., Xu, M., Ak, R., Schifferer, B., and Oldridge, E. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 2024.

Muennighoff, N. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*, 2022.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 2014–2037, 2023.

Muennighoff, N., Hongjin, S., Wang, L., Yang, N., Wei, F., Yu, T., Singh, A., and Kiela, D. Generative representational instruction tuning. In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024.

Nacar, O., Koubaa, A., Sibaee, S., Al-Habashi, Y., Ammar, A., and Boulila, W. Gate: General arabic text embedding for enhanced semantic textual similarity with matryoshka representation learning and hybrid loss training. *arXiv preprint arXiv:2505.24581*, 2025.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

Ni, J., Qu, C., Lu, J., Dai, Z., Abrego, G. H., Ma, J., Zhao, V., Luan, Y., Hall, K., Chang, M.-W., et al. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9844–9855, 2022.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4885–4901, 2020.

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., and Zhang, T. 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)*, pp. 10862–10878, 2024.

Oda, K., Chuang, P.-M., Shirai, K., and Kertkeidkachorn, N. One sentence, two embeddings: Contrastive learning of explicit and implicit semantic representations. *arXiv preprint arXiv:2510.09293*, 2025.

O'Neill, C., Ye, C., Iyer, K., and Wu, J. F. Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657*, 2024.

Opitz, J. and Frank, A. 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)*, pp. 625–638, 2022.

Pagliardini, M., Gupta, P., and Jaggi, M. 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)*, pp. 528–540, 2018.

Pan, T., Duan, Z., Li, Z., Dong, B., Liu, N., Li, X., and Wang, J. Negative matters: Multi-granularity hard-negative synthesis and anchor-token-aware pooling for enhanced text embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31102–31118, 2025.

Panigrahi, A., Simhadri, H. V., and Bhattacharyya, C. Word2Sense: Sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5692–5705, 2019.

Pappadopulo, D. and Farina, M. Non-contrastive sentence representations via self-supervision. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4274–4284, 2024.

Park, C., Moon, H., Park, C., and Lim, H.-S. Mirage: A metric-intensive benchmark for retrieval-augmented generation evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2883–2900, 2025.

Parrish, A., Schuster, S., Warstadt, A., Agha, O., Lee, S.-H., Zhao, Z., Bowman, S., and Linzen, T. Nope: A corpus of naturally-occurring presuppositions in english. In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, pp. 349–366, 2021.

Peng, L., Zhang, Y., Wang, Z., Srinivasa, J., Liu, G., Wang, Z., and Shang, J. 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)*, pp. 459–477, 2024.

Pennington, J., Socher, R., and Manning, C. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. 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)*, pp. 2227–2237, 2018.

Pipitone, N. and Alami, G. H. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024.

Ponwitayarat, W., Limkonchotiwat, P., Chuangsuwanich, E., and Nutanong, S. Space decomposition for sentence embedding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11227–11239, 2024.

Ponwitayarat, W., Ng, R., Montalan, J. R., Aung, T., Ngui, J. G., Susanto, Y., Tjhi, W., Tasawong, P., Cambria, E., Chuangsuwanich, E., et al. Sea-bed: Southeast asia embedding benchmark. *arXiv preprint arXiv:2508.12243*, 2025.

Poświata, R., Dadas, S., and Perełkiewicz, M. Pl-mteb: Polish massive text embedding benchmark. *arXiv preprint arXiv:2405.10138*, 2024.

Potts, C. Presupposition and implicature. *The handbook of contemporary semantic theory*, pp. 168–202, 2015.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Ranasinghe, T., Hettiarachchi, H., Orasan, C., and Mitkov, R. Musts: Multilingual semantic textual similarity benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 331–353, 2025.

Rau, D., Déjean, H., Chirkova, N., Formal, T., Wang, S., Clinchant, S., and Nikoulina, V. Bergen: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7640–7663, 2024.

Reimers, N. and Gurevych, I. 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)*, pp. 3982–3992, 2019.

Romero, M., Ding, S., Barret, C. D., Dinu, G., and Karypis, G. Beyond instruction-conditioning, mote: Mixture of task experts for multi-task embedding models. *arXiv preprint arXiv:2506.17781*, 2025.

Salemi, A. and Zamani, H. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 2395–2400, 2024.

Sancheti, A., Dale, D., Kozhevnikov, A., and Elbayad, M. Less mature is more adaptable for sentence-level language modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11680–11695, 2025.

Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., and Zaharia, M. 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)*, pp. 3715–3734, 2022.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. 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)*, pp. 5477–5490, 2020.

Sato, S., Tsukagoshi, H., Sasano, R., and Takeda, K. 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)*, pp. 519–530, 2024.

Senel, L. K., Utlu, I., Yucesoy, V., Koc, A., and Cukur, T. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(10):1769–1779, 2018.

Sharifymoghaddam, S., Upadhyay, S., Thakur, N., Pradeep, R., and Lin, J. Chatbot arena meets nuggets: Towards explanations and diagnostics in the evaluation of llm responses. *arXiv preprint arXiv:2504.20006*, 2025.

Sharma, L., Graesser, L., Nangia, N., and Evci, U. Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041*, 2019.

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

Simhi, A. and Markovitch, S. Interpreting embedding spaces by conceptualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1704–1719, 2023.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5548–5566, 2023.

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

Springer, J. M., Kotha, S., Fried, D., Neubig, G., and Raghunathan, A. Repetition improves language model embeddings. *arXiv preprint arXiv:2402.15449*, 2024.

Sravanthi, S., Doshi, M., Tankala, P., Murthy, R., Dabre, R., and Bhattacharyya, P. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12075–12097, 2024.

Sturua, S., Mohr, I., Kalim Akram, M., Günther, M., Wang, B., Krimmel, M., Wang, F., Mastrapas, G., Koukounas,

A., Wang, N., et al. Jina embeddings v3: Multilingual text encoder with low-rank adaptations. In *European Conference on Information Retrieval (ECIR)*, pp. 123–129, 2025.

Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N. A., Zettlemoyer, L., and Yu, T. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1102–1121, July 2023.

Su, H., Yen, H., Xia, M., Shi, W., Muennighoff, N., Wang, H.-y., Liu, H., Shi, Q., Siegel, Z. S., Tang, M., et al. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*, 2024.

Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., and Hovy, E. SPINE: SParse Interpretable Neural Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4921–4928, 2018.

Sun, J., Liu, S., Su, Z., Zhong, X., Jiang, P., Jin, B., Li, P., Shi, W., and Han, J. Grace: Generative representation learning via contrastive policy optimization. *arXiv preprint arXiv:2510.04506*, 2025a.

Sun, Y., Huang, Q., Wang, Y., and Tung, A. K. Diversinews: Enriching news consumption with relevant yet diverse news articles retrieval. *Proceedings of the VLDB Endowment*, 17(12):4277–4280, 2024.

Sun, Y., Huang, Q., Tang, Y., Tung, A. K. H., and Yu, J. A general framework for producing interpretable semantic text embeddings. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025b.

Sun, Y., Huang, Q., Tung, A. K. H., and Yu, J. PRISM: A framework for producing interpretable political bias embeddings with political-aware cross-encoder. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27719–27733, July 2025c.

Sun, Y., Huang, Q., Xu, Z., Sun, Y., Tang, Y., and Tung, A. K. One swallow does not make a summer: Understanding semantic structures in embedding spaces. *arXiv preprint arXiv:2512.00852*, 2025d.

Sun, Y., Wei, P., and Hsieh, L. B. Don't retrieve, navigate: Distilling enterprise knowledge into navigable agent skills for qa and rag. *arXiv preprint arXiv:2604.14572*, 2026.

Tamber, M. S., Xian, J., and Lin, J. Can't hide behind the api: Stealing black-box commercial embedding models. *arXiv preprint arXiv:2406.09355*, 2024.

Tamber, M. S., Kazi, S., Sourabh, V., and Lin, J. Teaching dense retrieval models to specialize with listwise distillation and llm data augmentation. *arXiv preprint arXiv:2502.19712*, 2025.

Tan, H., Zhan, S., Lin, H., Zheng, H.-T., and Chan, W. K. QAEA-DR: A Unified Text Augmentation Framework for Dense Retrieval. *IEEE Transactions on Knowledge & Data Engineering (TKDE)*, 37(06):3669–3683, 2025. ISSN 1558-2191. doi: 10.1109/TKDE.2025.3543203.

Tang, Y. and Yang, Y. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*, 2024.

Tang, Y. and Yang, Y. Finmteb: Finance massive text embedding benchmark. *arXiv preprint arXiv:2502.10990*, 2025.

Tang, Y. and Yang, Y. Kv-embedding: Training-free text embedding via internal kv re-routing in decoder-only llms. *arXiv preprint arXiv:2601.01046*, 2026.

Tang, Y., Shi, Y., Sun, Y., and Tung, A. K. H. Uncovering the bigger picture: Comprehensive event understanding via diverse news retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33927–33945, 2025.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. 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.

Thirukovalluru, R. and Dhingra, B. Geneol: Harnessing the generative power of llms for training-free sentence embeddings. *arXiv preprint arXiv:2410.14635*, 2024.

Thirukovalluru, R., Wang, X., Chen, J., Li, S., Lei, J., Jin, R., and Dhingra, B. Sumcse: Summary as a transformation for contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3577–3588, 2024.

Tihanyi, N., Ferrag, M. A., Jain, R., Bisztray, T., and Debbah, M. 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)*, pp. 296–302, 2024.

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

Truong, T. H., Verspoor, K., Cohn, T., and Baldwin, T. Learning robust negation text representations. *arXiv preprint arXiv:2507.12782*, 2025.

Vera, H. S., Dua, S., Zhang, B., Salz, D., Mullins, R., Panyam, S. R., Smoot, S., Naim, I., Zou, J., Chen, F., et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

Wang, K., Reimers, N., and Gurevych, I. Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 671–688, 2021.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024a.

Wang, Q., Zaki, M. J., Kollias, G., and Kalantzis, V. Multi-sense embeddings for language models and knowledge distillation. *arXiv preprint arXiv:2504.06036*, 2025.

Wang, S., Liu, J., Song, S., Cheng, J., Fu, Y., Guo, P., Fang, K., Zhu, Y., and Dou, Z. Domainrag: A chinese benchmark for evaluating domain-specific retrieval-augmented generation. *arXiv preprint arXiv:2406.05654*, 2024b.

Wang, X., He, J., Wang, P., Zhou, Y., Sun, T., and Qiu, X. Denosent: A denoising objective for self-supervised sentence representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 19180–19188, 2024c.

Wang, Y., Wang, L., Li, Y., He, D., and Liu, T.-Y. A theoretical analysis of NDCG type ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pp. 25–54, 2013.

Wei, Q., Morrell, E., Goetz, L., and van der Schaar, M. Semantic-kg: Using knowledge graphs to construct benchmarks for measuring semantic similarity. *arXiv preprint arXiv:2511.19925*, 2025.

Weller, O., Chang, B., MacAvaney, S., Lo, K., Cohan, A., Van Durme, B., Lawrie, D., and Soldaini, L. Followir: Evaluating and teaching information retrieval models to follow instructions. *arXiv preprint arXiv:2403.15246*, 2024.

Williams, A., Nangia, N., and Bowman, S. R. 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*, pp. 1112–1122, 2018.

Wu, J., Li, J., Li, Y., Liu, L., Xu, L., Li, J., Yeung, D.-Y., Zhou, J., and Yu, M. Sitemb-v1. 5: Improved context-aware dense retrieval for semantic association and long story comprehension. *arXiv preprint arXiv:2508.01959*, 2025.

Xianming, L., Li, Z., Li, J., Xie, H., and Li, Q. Ese: Espresso sentence embeddings. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Xiao, C., Huang, Z., Chen, D., Hudson, G. T., Li, Y., Duan, H., Lin, C., Fu, J., Han, J., and Moubayed, N. A. Pixel sentence representation learning. *arXiv preprint arXiv:2402.08183*, 2024a.

Xiao, C., Chung, I., Kerboua, I., Stirling, J., Zhang, X., Kardos, M., Solomatin, R., Moubayed, N. A., Enevoldsen, K., and Muennighoff, N. Mieb: Massive image embedding benchmark. *arXiv preprint arXiv:2504.10471*, 2025.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., and Nie, J.-Y. 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)*, pp. 641–649, 2024b.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., and Nie, J.-Y. 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)*, pp. 641–649, 2024c.

Xiong, G., Jin, Q., Lu, Z., and Zhang, A. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024.

Xu, G., Yin, Z., Zhang, L., Liang, J., Lu, W., Zhang, X., Yang, Z., Jiang, S., and Yang, D. Comlq: Benchmarking complex logical queries in information retrieval. *arXiv preprint arXiv:2511.12004*, 2025.

Yamada, K. and Zhang, P. Out-of-the-box conditional text embeddings from large language models. *arXiv preprint arXiv:2504.16411*, 2025.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

Yoo, Y., Cha, J., Kim, C., and Kim, T. Hyper-cl: Conditioning sentence representations with hypernetworks. In

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 700–711, 2024.

You, X., Huang, Q., Li, L., Chang, X., and Yu, J. Cut to the Chase: Training-free Multimodal Summarization via Chain-of-Events. *arXiv preprint arXiv:2603.06213*, 2026a.

You, X., Huang, Q., Li, L., Zhang, C., Liu, X., Zhang, M., and Yu, J. Knowledge completes the vision: A multimodal entity-aware retrieval-augmented generation framework for news image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12108–12116, 2026b. URL https://ojs.aaai.org/index.php/AAAI/article/view/38200.

Yu, P., Merrick, L., Nuti, G., and Campos, D. Arctic-embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*, 2024.

Yue, S., Song, S., Cheng, X., and Hu, H. Do large language models understand conversational implicature–a case study with a chinese sitcom. In *China National Conference on Chinese Computational Linguistics*, pp. 402–418, 2024.

Zhang, B., Song, Z., and Li, C. Cse-sfp: Enabling unsupervised sentence representation learning via a single forward pass. *arXiv preprint arXiv:2505.00389*, 2025a.

Zhang, C., Zhang, Q., Li, K., Nuthalapati, S. V., Zhang, B., Liu, J., Li, S., Zhang, L., and Fan, X. Gem: Empowering llm for both embedding generation and language understanding. *arXiv preprint arXiv:2506.04344*, 2025b.

Zhang, D., Li, J., Zeng, Z., and Wang, F. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*, 2024a.

Zhang, J., Lan, Z., and He, J. Contrastive learning of sentence embeddings from scratch. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3916–3932, 2023a.

Zhang, M., Zhang, X., Zhao, X., Huang, S., Hu, B., and Zhang, M. On the role of pretrained language models in general-purpose text embeddings: A survey. *arXiv preprint arXiv:2507.20783*, 2025c.

Zhang, X., Ma, X., Shi, P., and Lin, J. Mr. TyDi: A multilingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 127–137, 2021. doi: 10.18653/v1/2021.mrl-1.12.

Zhang, X., Li, Z., Zhang, Y., Long, D., Xie, P., Zhang, M., and Zhang, M. Language models are universal embedders. *arXiv preprint arXiv:2310.08232*, 2023b.

Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., and Lin, J. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1114–1131, 2023c.

Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*, 2024b.

Zhang, Y., Bai, J., Cai, Z., Qin, S., Chen, Z., Guan, J., and Rong, W. Your dense retriever is secretly an expeditious reasoner. *arXiv preprint arXiv:2510.21727*, 2025d.

Zhao, D., Huang, Q., Yan, D., Sun, Y., and Yu, J. Partially shared concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13117–13125, 2026.

Zhao, K., Wu, Q., Miao, Z., and Tsuruoka, Y. Prompt tuning can simply adapt large language models to text encoders. In *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, pp. 38–50, 2025.

Zheng, Z., Qiu, S., Fan, L., Zhu, Y., and Zhu, S.-C. Grice: A grammar-based dataset for recovering implicature and conversational reasoning. In *Findings of the Association for Computational Linguistics: ACL 2021*, pp. 2074–2085, 2021.

Zhuang, H., Emma Zhang, W., Yang, J., Chen, W., and Sheng, Q. Z. 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)*, pp. 3591–3601, 2024a.

Zhuang, S., Wang, S., Koopman, B., and Zuccon, G. Starbucks: Improved training for 2d matryoshka embeddings. *arXiv preprint arXiv:2410.13230*, 2024b.

Zhuang, Y., Trinh, A., Qiang, R., Sun, H., Zhang, C., Dai, H., and Dai, B. Towards better instruction following retrieval models. *arXiv preprint arXiv:2505.21439*, 2025.

Zhuo, W., Sun, Y., Wang, X., Zhu, L., and Yang, Y. WhitenedCSE: Whitening-based contrastive learning of sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12135–12148, 2023.

Zinvandi, E., Alikhani, M., Sarmadi, M., Pourbahman, Z., Arvin, S., Kazemi, R., and Amini, A. Famteb: Massive text embedding benchmark in persian language. *arXiv preprint arXiv:2502.11571*, 2025.

A. Implementation Details

A.1. Model Checkpoints

We use the model checkpoints listed in Table 2 for all experiments reported in Section 6. To ensure a fair and reproducible comparison, we adopt the **official checkpoints** used by the MTEB benchmark and evaluate every open-source model under the **default configurations** provided by the Sentence Transformers library (Reimers & Gurevych, 2019). Importantly, we do not apply any additional parameter tuning, task-specific calibration, or prompt engineering; this design choice isolates the intrinsic capability of each embedding model under standard deployment settings.

Proprietary Embeddings For OpenAI's proprietary embedding models, we obtain representations using OpenAI's official client library[3] and follow the recommended embedding API usage. This ensures that the results reflect the intended inference behavior of these models rather than custom wrappers or ad hoc post-processing.

Complementary Baselines To contextualize performance, we include two complementary baselines:

- **Random:** For each dataset, we implement a random predictor by sampling labels according to the empirical label distribution. This provides a hardness reference that accounts for class imbalance and prevents overstating performance on skewed datasets.
- **Bag-of-Tokens (Harris, 1954; Sun et al., 2025b):** As a lexical-feature baseline, we include a Bag-of-Tokens representation using the `google-bert/bert-base-uncased` tokenizer. This baseline serves as a strong surface-matching reference point, helping quantify how much modern embeddings improve beyond lexical overlap.

Together, these implementation choices are designed to make comparisons consistent across model families, while keeping the evaluation protocol aligned with widely used community standards.

Table 2. **List of evaluated embedding models and the corresponding checkpoints used in our experiments.**

Model	Model Size	Checkpoint
S-BERT (Reimers & Gurevych, 2019)	22.7M	`sentence-transformers\all-MiniLM-L6-v2`
GIST-Small (Solatorio, 2024)	33.4M	`avsolatorio\GIST-small-Embedding-v0`
BGE-Base (Xiao et al., 2024b)	109M	`BAAI\bge-base-en-v1.5`
Angle (Li & Li, 2024a)	335M	`WhereIsAI\UAE-Large-V1`
BGE-Large (Xiao et al., 2024b)	335M	`BAAI\bge-large-en-v1.5`
MXBAI-Large (Lee et al., 2024c)	335M	`mixedbread-ai\mxbai-embed-large-v1`
GIST-Large (Solatorio, 2024)	335M	`avsolatorio\GIST-large-Embedding-v0`
Stella (Zhang et al., 2024a)	435M	`NovaSearch\stella_en_400M_v5`
Linq-Mistral (Kim et al., 2024)	7.11B	`Linq-AI-Research\Linq-Embed-Mistral`
E5-Mistral (Wang et al., 2023; 2022)	7.11B	`intfloat\e5-mistral-7b-instruct`
GTE-Qwen (Li et al., 2023b)	7.61B	`Alibaba-NLP\gte-Qwen2-7B-instruct`
Jasper (Zhang et al., 2024a)	1.99B	`NovaSearch\jasper_en_vision_language_v1`
OpenAI-Small	N.A.	`text-embedding-3-small`
OpenAI-Large	N.A.	`text-embedding-3-large`

A.2. Implicit Semantics Tasks

Our evaluation targets **implicit semantics** across diverse task formulations and data sources. Because the underlying datasets differ in structure and annotation format, we organize the evaluation into three 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.

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

Table 3. **Accuracy (%) of embedding models on the Pragmatics Understanding Benchmark (PUB) tasks.** Each column corresponds to one of the 14 PUB tasks (T1–T14), covering diverse pragmatic phenomena.

Model	Pragmatics Understanding Benchmark (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 (Sun et al., 2025b)	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 (Reimers & Gurevych, 2019)	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 (Solatorio, 2024)	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 (Xiao et al., 2024b)	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 (Li & Li, 2024a)	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 (Xiao et al., 2024b)	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 (Lee et al., 2024c)	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 (Solatorio, 2024)	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 (Zhang et al., 2024a)	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
LLM-based Embeddings														
Linq-Mistral (Kim et al., 2024)	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 (Wang et al., 2023; 2022)	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 (Li et al., 2023b)	**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 Encoders														
Jasper (Zhang et al., 2024a)	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 Embeddings														
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

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**.

A.3. Evaluation Protocols

For **classification** and **pair classification** tasks, we follow the standard evaluation protocol used in MTEB (Muennighoff et al., 2023). In this setting, embeddings are computed for inputs (and, when applicable, for candidate labels or verbalized label descriptions), and predictions are made using the corresponding embedding-based classification procedure consistent with MTEB's implementation.

For **zero-shot classification**, we follow the embedding-based approach[4] described in OpenAI's documented use case: the input query (or question) and its associated text are embedded jointly, while each candidate answer option is embedded separately. The model selects the option with the highest similarity to the input embedding. This protocol provides a standardized way to test "label selection by semantic similarity," allowing comparison across both open-source and proprietary embeddings.

B. 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.

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

Table 4. **Accuracy (%) of embedding models on additional implicit meaning benchmarks. P-Stance** includes stance detection 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)** consists of target identification (Tar.), intent (Int.), sexism (Sex.), offensiveness (Off.), and bias detection (Bias) tasks. **Political Bias (Pol. Bias)** denotes political ideology classification.

Model	P-Stance			IHS			SBIC					Pol. Bias
	Trump	Biden	Bernie	Det.	Cat.	Tar.	Tar.	Int.	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 (Sun et al., 2025b)	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 (Reimers & Gurevych, 2019)	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 (Solatorio, 2024)	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 (Xiao et al., 2024b)	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 (Li & Li, 2024a)	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 (Xiao et al., 2024b)	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 (Lee et al., 2024c)	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 (Solatorio, 2024)	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 (Zhang et al., 2024a)	79.2	80.0	70.2	76.9	56.7	47.6	81.2	80.9	93.4	83.2	81.5	54.4
LLM-based Embeddings												
Linq-Mistral (Kim et al., 2024)	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 (Wang et al., 2023; 2022)	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 (Li et al., 2023b)	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 Encoders												
Jasper (Zhang et al., 2024a)	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 Embeddings												
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

Widespread Variance Across Models A central observation is that performance varies substantially across both models and tasks, often in ways that are not predicted by surface-level benchmark rankings. For example, many models achieve near-perfect accuracy on *Task 1 (Direct/Indirect Classification)* from PUB, suggesting that some pragmatic cues can be captured by surface correlates. However, several widely used embedding models, such as **GIST-Small**, **S-BERT**, and **BGE-Base**, perform only marginally better than **Bag-of-Tokens** on more challenging tasks, and in some cases fall below it.

This pattern becomes even more pronounced on certain implicature-oriented tasks. On *Task 10 (Implicature NLI)*, multiple models, including OpenAI's proprietary models and the LLM-based models like **E5-Mistral** and **Jasper**, underperform the **Bag-of-Tokens** baseline. This indicates that higher capacity or modern training recipes do not automatically translate into reliable pragmatic competence, and that lexical heuristics can sometimes outperform dense representations when the latter fail to encode the relevant inferential structure. Overall, these results reinforce a key theme of the paper: **Success on conventional surface-level benchmarks does not reliably transfer to tasks requiring deeper interpretive understanding.**

Strengths of Large and Multimodal Models While variance is widespread, the results also reveal a consistent advantage for large-scale and multimodal embedding models. **Jasper**, for example, ranks among the top across a wide range of tasks, with particularly strong performance on socially grounded datasets such as IHS and SBIC. Similarly, large-scale models like **E5-Mistral** and **OpenAI-Large** perform strongly across multiple domains, including social bias classification and pragmatic reasoning tasks.

These outcomes suggest that scale and broader pretraining signals can improve robustness for complex semantic phenomena. However, the advantage is not uniform: even high-performing models exhibit sharp drops on specific pragmatic tasks, indicating that capacity helps but does not fully resolve the underlying challenge.

Persistent Challenges in Implicature and Reference Tasks Despite improvements from scale and multimodality, certain pragmatic phenomena remain consistently difficult across all evaluated models. In particular, *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**.

This trend points to a deeper limitation of current training pipelines. Many dominant training signals–whether self-supervised, NLI-based, or retrieval-based–do not systematically require recovering unstated intent or bridging pragmatic gaps. As a result, even powerful models may rely on shallow correlates rather than encoding the inferential structure needed for implicature and reference resolution.

Implications for Benchmark and Model Design Taken together, the appendix results expose persistent blind spots in current embedding models, especially for tasks involving implicature, figurative language, presupposition, reference, and socially grounded inference. More importantly, they clarify that these weaknesses are not isolated edge cases: they reflect a systematic mismatch between (i) what embeddings are trained to optimize and (ii) what is required to represent implicit meaning.

Addressing these gaps will likely require both linguistically grounded supervision that targets implicit semantics and benchmarks that directly measure interpretive competence beyond surface similarity. In this sense, the appendix results provide a detailed empirical foundation for the agenda proposed in the paper.