OwlBERT: Domain-Adapted Encoders for Product Understanding and Relevance Ranking in E-commerce

Anonymous ACL submission

Abstract

E-commerce retrieval and recommendation require encoders that preserve product-specific distinctions, including substitutes versus complements, attribute equivalence, brand/category constraints, and query–product intent. General purpose encoders trained on broad web text often underrepresent these fine grained compatibility and specification signals. We introduce OwlBERT, a family of encoder models with 17M–400M parameters for e-commerce product understanding. Starting from the public Ettin long context encoder, OwlBERT is domain annealed on EcomWeb, a 350B-token e-commerce domain language corpus, while retaining an 8,192-token context window. We further introduce Guided MLM, a masking strategy that preferentially selects salient product spans without changing the masked language modeling objective. We evaluate OwlBERT on e-commerce domain language modeling, ESCI graded query–product similarity, WANDS relevance ranking, and GLUE retention checks. Domain annealing substantially improves held out e-commerce modeling: at Base scale, perplexity drops from 6.04 to 3.99 after Phase 3 training with Guided MLM. On Amazon ESCI, OwlBERT-base already matches E5-large with fewer than half the parameters, while OwlBERT-large reaches 0.457 Spearman correlation, outperforming E5-large and BGE-large under the same graded similarity protocol. On WANDS, OwlBERT-large achieves 0.924 NDCG@5, 0.933 NDCG@10, and 0.977 MAP, while maintaining competitive GLUE retention. These results show that targeted domain annealing and span-guided masking can deliver stronger e-commerce relevance than general purpose encoders at compact encoder scales.

1 Introduction

Encoder-only language models remain central to production retrieval and reranking because they offer a favorable trade off between quality, latency and cost. BERT (Devlin et al., 2019) and its successors support high throughput matching systems, while recent work such as ModernBERT (Warner et al., 2025) shows that encoders can still improve through better recipes, longer contexts and more efficient attention. The open question we study is whether such encoders can be adapted to the product centric semantic relations that determine e-commerce relevance.

E-commerce semantics differ from generic sentence similarity in systematic ways. Queries are often short and underspecified, product text is semi-structured, and small attributes such as brand, size, color, material, quantity or compatibility can determine whether two items are interchangeable. For example, "iPhone 14 case" and "iPhone 14 phone" share many tokens but refer to complementary rather than substitutable products, while "women's waterproof trail running shoes size 8" depends on gender, use case, material and size constraints. Generic encoders trained on broad web corpora often blur these distinctions, especially for long tail entities and attribute value constructions.

Our approach combines EcomWeb, a domain-annealing strategy, and Guided MLM (§4.1). Given a strong long context checkpoint, is a focused annealing phase enough to recover product specific semantics? We release OwlBERT as pre-trained encoders and evaluate them on downstream benchmarks rather than pretraining loss alone.

Contributions Our work makes the following contributions:

1. We introduce **EcomWeb**, a 350B-token product language corpus curated from FineFineWeb (M-A-P et al., 2024) through domain selection, LLM validation and classifier-based filtering; and **OwlBERT**, a family of compact e-commerce encoders domain-adapted on this corpus, and release anonymized checkpoints and corpus data at `https:`

//huggingface.co/buckets/doubleblind000/owlbert.

2. We study Phase 3 domain annealing from a strong long context checkpoint as a practical e-commerce adaptation strategy (§4).

3. We propose **Guided MLM** and validate it with controlled perplexity ablations (§6.4).

4. We evaluate OwlBERT checkpoints against general-domain and task-matched baselines on product-language, relevance, and general NLU benchmarks (§6).

2 Related Work

Modern encoder pretraining. BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) established encoder-only pretraining as a strong foundation for classification and retrieval. More recent work revisits the encoder recipe itself: DeBERTa (He et al., 2020) improves attention parameterization, MosaicBERT (Portes et al., 2023) and Cramming (Geiping and Goldstein, 2023) study training efficiency, and ModernBERT (Warner et al., 2025) combines RoPE, GeGLU, alternating local/global attention, unpadding and flash attention for long-context encoders. OwlBERT adopts this modern encoder direction but asks whether the training signal can be specialized for product-centric semantic relations.

Domain-adaptive encoders. Domain-specific pretraining has improved encoders in biomedical, clinical, scientific, financial and legal NLP, including BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), BioClinical ModernBERT (Sounack et al., 2025), SciBERT (Beltagy et al., 2019), FinBERT (Yang et al., 2020) and LEGAL-BERT (Chalkidis et al., 2020). These studies show that targeted corpora can improve downstream behavior, but evaluation typically targets domain-specific classification or tagging rather than product retrieval. E-commerce search additionally requires graded query-product relevance and attribute-sensitive matching.

E-commerce and product representation. E-BERT augments BERT with phrase-level and product-level pretraining tasks for e-commerce (Zhang et al., 2020), while CatBERT trains incrementally on catalog text and argues that product data differs sharply from sentence-like corpora (Mallavarapu et al., 2022). OwlBERT follows this domain-specialization motivation but emphasizes open data, compact encoder scales, and Guided MLM as an attribute aware masking mechanism.

Retrieval, reranking and product benchmarks. Query-product matching is fundamentally a retrieval and ranking problem rather than generic sentence-pair classification. Because of this framing, standard evaluations compare dense embedding models such as E5 (Wang et al., 2022), BGE (Xiao et al., 2024), and GTE (Li et al., 2023), cross-encoder rerankers (Nogueira and Cho, 2019), and general encoders including ModernBERT (Warner et al., 2025). For e-commerce, Amazon ESCI (Reddy et al., 2022) provides graded query–product similarity labels and WANDS (Chen et al., 2022) evaluates query-level ranking over human relevance judgments.

3 Data and Corpus Curation

Fine-grained domain representation is critical for accurate e-commerce encoding. We therefore curate **EcomWeb**, a large corpus of commerce-related text. Our source is *FineFineWeb* (M-A-P et al., 2024), a 4.4 trillion token CommonCrawl-derived corpus organised into fifty categories. Each entry consists of a text snippet and domain label. To isolate retail content we manually inspect domains and select those with substantial overlap with commerce. We group them into sources requiring domain-specific filtering and sources with near-complete commerce overlap (Appendix A). Notably, the Hobby domain contributes 21.4%, News contribute 11.7%, and Fashion and Beauty combined contribute 15.3%. Figure 1 visualises the resulting distribution.

After selecting domains, we apply language detection to retain English text and remove boilerplate, advertisements and low-quality pages. We deduplicate near-duplicate documents using MinHash and filter profanity and adult content. The resulting corpus comprises more than 350 billion tokens of e-commerce text, providing a large-scale domain-specific resource for pretraining.

Source corpus and cleaning pipeline. After selecting commerce overlapping FineFineWeb domains (Appendix A), we sample examples from each for LLM-assisted binary relevance labeling. A stronger auditor model reviews a subset of labels to identify prompt and guideline failures. The re-

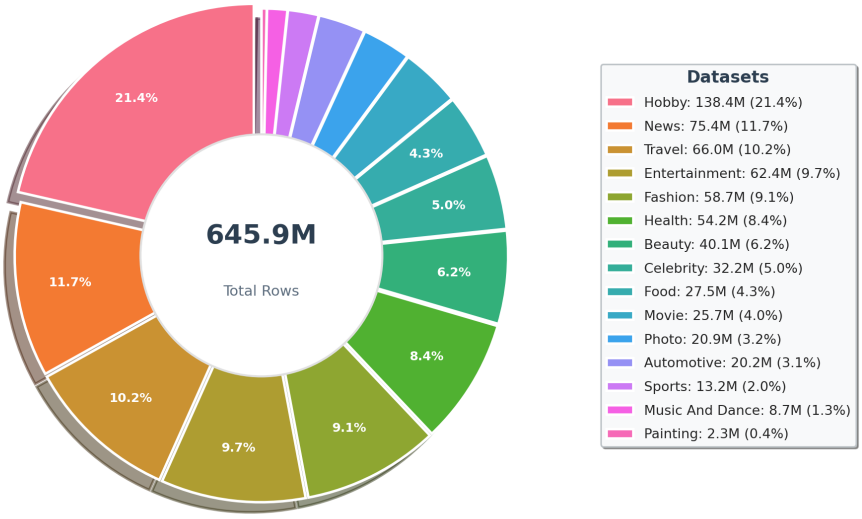

Figure 1: Domain distribution of EcomWeb. Sizes represent the amount of filtered data contributed by each FineFineWeb domain and additional corpora. Hobby and News supply the largest portions.

sulting labels are distilled into per-domain fastText classifiers, which score the full domain partitions at scale; domain-specific thresholds determine the retained pretraining subset. Before Phase 3 training, we decontaminate the retained pretraining corpus against the public evaluation benchmarks used in this paper, including Amazon ESCI and WANDS.

4 Pretraining

OwlBERT continues from the public Ettin (Weller et al., 2025) encoder checkpoint after its long context phase. Phases 1–2 are inherited initialization (Table 1); this paper's contribution is Phase 3 domain specialization on EcomWeb with annealed sampling and Guided MLM. Our central question is how much e-commerce behavior Phase 3 alone can induce without full domain pretraining from scratch. Let $X = (x_1, \ldots, x_T)$ denote a token sequence and θ the model parameters. The masked language modeling (MLM) objective is

$$\mathcal{L}_{\text{MLM}}(\theta) = -\,\mathbb{E}_{X\sim\mathcal{D},\, m\sim\mathcal{M}}\big[\log p_\theta(X_m \mid X_{\setminus m})\big]. \tag{1}$$

where m indexes masked positions drawn from a masking distribution $\mathcal{M}$ and $X_{\setminus m}$ denotes the unmasked context. Masking is performed dynamically with span-aware sampling to better capture multi-token lexical units and to reduce information leakage from partially observed subword fragments.

4.1 Guided MLM

Guided MLM is a targeted masking variant that prioritizes information-rich product spans. Guidance comes from lightweight named-entity recognizers: for each FineFineWeb source domain we curate training labels with GLiNER (Zaratiana et al., 2024) and fine-tune a BERT-mini tagger on domain-specific attribute and product-entity annotations. At Phase 3 training time, sequences are tagged with the NER model matched to their source domain; recognized spans highlight characteristic attributes and product features that carry retrieval-critical signal. Let $p(s)$ denote the base span sampling probability for candidate span s. For tagged spans we set $\tilde{p}(s) = \alpha\, p(s)$ and leave $\tilde{p}(s) = p(s)$ otherwise, then renormalize within each sequence to obtain the guided masking distribution. The MLM objective is unchanged. During training, 5% of sequences per batch are masked under this guided distribution and the remainder use standard random span-masking. This preserves the generalization benefits of unbiased MLM while allocating additional learning signal to rare, high-value tokens and structured product descriptors.

Tokenizer and optimizer. OwlBERT preserves the tokenizer and long context architecture of the Ettin Phase 2 checkpoint. The tokenizer is a modern BPE vocabulary of 50,368 tokens, improving token efficiency over the original BERT WordPiece; the inherited 8,192-token context window

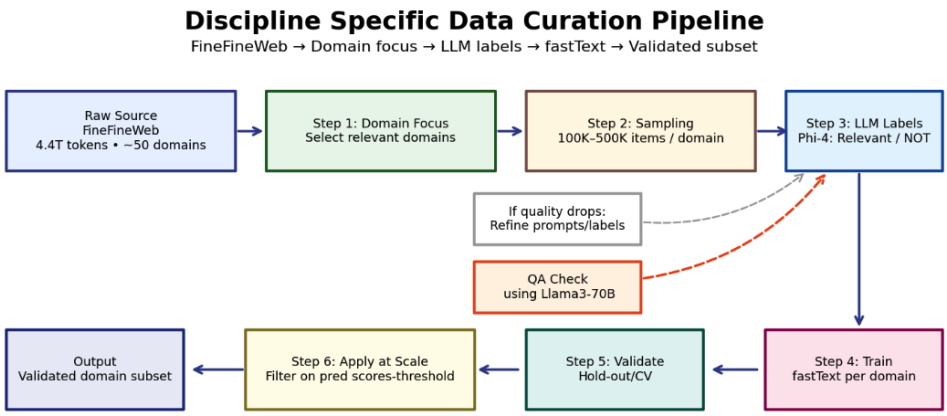

Figure 2: EcomWeb curation pipeline: domain selection, sampling, LLM labeling, QA auditing, fastText distillation, and thresholded filtering at scale.

Table 1: Training provenance and curriculum. Phases 1–2 are inherited from the public Ettin long context encoder checkpoint; this paper performs the Phase 3 e-commerce domain specialization training.

Phase	Starting checkpoint	Data	Tokens	Seq. length	Trained here?
1: General pretraining	Ettin recipe/checkpoint lineage	Mixed open data	1.7T	512	No
2: Context extension	Ettin Phase 1 checkpoint	Mixed long context data	250B	8,192	No
3: Domain annealing	Ettin Phase 2 8,192-token checkpoint	EcomWeb	350B	8,192	Yes

Table 2: Examples of Guided MLM span selection.

Text	Guided token spans
Removable, washable cover Easy-carry handle Soft, warm and luxurious! Filled with reclaimed poly fiber to keep waste out of landfills. Denim adds a touch of softness to the rugged look of this comfortable pet bed. #39103 blue denim wash	Removable; washable; Easy-carry; reclaimed poly fiber; Denim; blue;

supports long product text. During our Phase 3 domain annealing, optimization uses StableAdamW with Adafactor-style update clipping and a trapezoidal Warmup–Stable–Decay schedule with a final inverse-square-root ($1/\sqrt{t}$) decay; hardware, batch size, steps and seeds are omitted from this submission for anonymity.

Phases 1–2 inherit Ettin's 1.7T-token general pretraining and 250B-token context extension to 8,192 tokens; OwlBERT starts from the released Phase 2 checkpoint.

E-commerce annealing. We specialize the Phase 2 checkpoint on EcomWeb for approximately 350 billion tokens. We reduce the masking ratio to 10%–15%, anneal sampling weights toward e-commerce data, and interleave Guided MLM with standard span masking. The learning rate follows the same Warmup–Stable–Decay schedule with a final $1/\sqrt{t}$ decay while preserving the inherited 8,192-token context (Figure 3).

5 Model Architecture

OwlBERT retains the BERT-style encoder architecture and modern long context enhancements inherited from the Ettin Phase 2 checkpoint. Key differences from the original BERT architecture include:

- **Biasless layers and pre-normalisation.** Following ModernBERT, we remove bias terms from linear layers and layer norms and use pre-layer normalisation to improve training stability.
- **Rotary positional embeddings.** We replace absolute positional embeddings with RoPE, enabling extrapolation to long contexts and efficient implementation.
- **GeGLU activations.** GeGLU activations provide better optimisation compared to GELU.
- **Alternating global/local attention.** Attention layers alternate between full attention and lo-

Figure 3: Training curriculum for OwlBERT. Phases 1–2 are inherited from the public Ettin long context encoder checkpoint; this work performs the Phase 3 annealing on EcomWeb for 350 billion tokens.

cal sliding windows, reducing quadratic complexity while preserving global context.

- **Unpadding and flash attention.** We adopt unpadding to remove padding tokens and use Flash Attention for both global and local attention, improving throughput.
- **Hardware-aware depth.** The Micro, Mini, Base and Large variants follow a deep-and-narrow design to maximise GPU utilisation within parameter budgets.

Table 9 in Appendix B summarises the configuration of OwlBERT models.

6 Experiments

We compare OwlBERT checkpoints against general-domain and task matched baselines on downstream tasks (§6.2–§6.4), testing whether domain annealing transfers beyond pretraining perplexity. We provide anonymized OwlBERT checkpoints and EcomWeb artifacts for reproducibility during review at `https://huggingface.co/buckets/doubleblind000/owlbert`.

Datasets. We evaluate on Amazon ESCI (Reddy et al., 2022) (English `small` split: 419,653 train / 181,701 test), WANDS (Chen et al., 2022) (140,068 / 46,690 / 46,690 train/dev/test), and GLUE (Wang et al., 2018) as a general NLU retention check.

Evaluation protocols. All models start from publicly available or OwlBERT pre-trained checkpoints. We evaluate masked product token recovery as an intrinsic probe, fine-tune on ESCI and WANDS for graded similarity and query level ranking, and fine-tune on GLUE for retention. Task-specific metrics are given in the subsections below; fine-tuning settings are summarized next and detailed in Appendix C.

Fine-tuning. Dense retrievers are fine-tuned on ESCI and cross-encoder rerankers on WANDS using the dataset-provided splits without resampling, with HuggingFace `Trainer` (seed 42) and all reported scores on the held out test split. ESCI uses a Sentence-Transformers bi-encoder with CoSENT-Loss (Huang et al., 2024) on graded query-product pairs; WANDS uses a CrossEncoder trained with BCE on concatenated query-product inputs. For each model we run a lightweight hyperparameter sweep over learning rate and epoch count and report the best saved checkpoint (Appendix C).

6.1 Masked Product Token Recovery

We follow the masked-token recovery protocol of ModernBERT and FineFineWeb (M-A-P et al., 2024). Product titles and descriptions are truncated to 128, 256 or 512 tokens; a span-aware mask covering 15% of tokens is applied; and the model predicts the original tokens at top-k with $k \in \{1, 3, 5\}$. Appendix D reports the full title and description matrices. OwlBERT-base consistently outperforms ModernBERT-base across all block sizes and top-k settings.

6.2 Graded Query-Product Similarity

We evaluate whether domain adapted MLM encoders produce better query-product embeddings than general purpose retrieval models when relevance grade is used as the similarity target. Starting from each OwlBERT checkpoint, we train an embedding model on the English *train* split of Amazon ESCI. Each example is a (query, product title) pair, truncated when needed to fit the encoder context window. We use the ESCI relevance grade as

Table 3: ESCI graded similarity after embedding fine-tuning on the train split. Scores are Spearman correlations between predicted cosine similarity and grade-derived targets on held-out query-product pairs.

Family	Model	Params	Spearman
OwlBERT	Micro	17M	0.2991
OwlBERT	Mini	68M	0.3955
OwlBERT	Base	150M	0.4208
OwlBERT	Large	400M	0.4571
Ettin	encoder-17m	17M	0.253
Ettin	encoder-68m	68M	0.302
ModernBERT	base	150M	0.319
ModernBERT	large	400M	0.369
EmbeddingGemma	300m	300M	0.411
E5	small	33M	0.3857
E5	base	110M	0.4076
E5	large	335M	0.4186
BGE	small	33M	0.3915
BGE	base	110M	0.4056
BGE	large	335M	0.4316

the supervision signal for similarity by mapping Exact, Substitute, Complement and Irrelevant to 1.0, 0.66, 0.33 and 0.0, respectively, and optimize with CoSENT loss (Huang et al., 2024) so that cosine similarity in the embedding space preserves these graded relations. Evaluation reports Spearman correlation between predicted cosine similarities and the grade-derived target scores on held-out query-product pairs.

We compare OwlBERT-initialized embedding models against Microsoft E5 and BAAI BGE encoders trained under the same ESCI protocol and metric. E5 and BGE are established retrieval embedding families that already undergo multiple stages of contrastive pretraining before task-specific fine-tuning; despite this, OwlBERT models still outperform them at comparable scale. As shown in Figure 4 and Table 3, performance improves from Micro through Large, with OwlBERT-large reaching 0.457 Spearman, suggesting that domain MLM pretraining transfers to stronger product-relevance embeddings when grade-based similarity is the evaluation objective.

6.3 Product Search Relevance on WANDS

WANDS (Chen et al., 2022) provides 480 queries, 42,994 products and 233,448 human relevance judgments labeled *Exact*, *Partial* or *Irrelevant*. Unlike the ESCI graded-similarity task in §6.2, evaluation is query-level ranking over each query's labeled candidate products rather than pair-level correlation. We fine-tune OwlBERT, ModernBERT-base/large and BERT-mini/small on the WANDS train split (dev used for checkpoint selection; Appendix C), representing each product by its name, description and structured catalog features.

We also adapt Alibaba-NLP/gte-reranker-modernbert-base and BAAI/bge-reranker-base as cross-encoder baselines; both are already strong rerankers trained on large-scale relevance supervision before WANDS fine-tuning. At test time, models score every labeled query-product pair, candidates are ranked within each query, and we report NDCG@5, NDCG@10 and MAP using *Exact* > *Partial* > *Irrelevant* as graded gain.

Table 4 compares the OwlBERT family against ModernBERT, BERT and the GTE/BGE reranker baselines. OwlBERT models outperform the general encoder and pre-trained reranker baselines across NDCG@5, NDCG@10 and MAP, with performance generally improving from Micro through Large.

6.4 Guided MLM Ablation

To isolate Guided MLM, we run a same-checkpoint ablation that varies only the masking strategy while holding the Phase 2 initialization, data subset and token budget fixed. We evaluate the Base and Large ablation runs after 20B training tokens using held-out e-commerce perplexity. Table 5 shows consistent PPL reductions when Guided MLM is enabled. Unlike the full EcomWeb Phase 3 comparison in Table 7, this ablation is restricted to 20B tokens so the effect of guided masking can be isolated.

Table 5: Guided MLM perplexity ablation for Base and Large variants after 20B training tokens, not the full EcomWeb run. Lower PPL is better.

Variant	Guided MLM	PPL	Δ PPL
Base	no	5.67	–
Base	yes	5.14	0.53
Large	no	5.32	–
Large	yes	4.68	0.64

6.5 General NLU Retention

We use GLUE as a retention check to verify that e-commerce specialization does not catastrophically degrade general language understanding. Table 6 presents results across four model size categories. These scores are not the main claim of the paper; they show that the specialization gains on product tasks do not come at the cost of broad collapse on standard NLU probes.

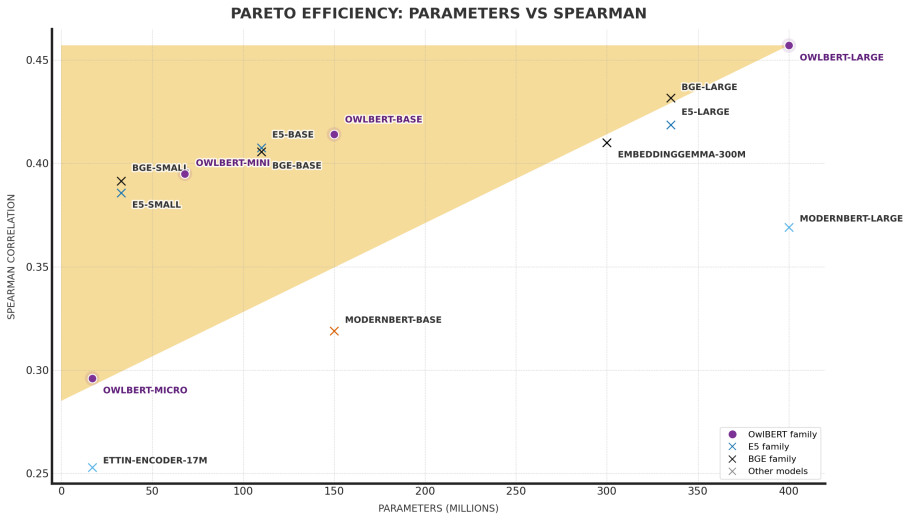

Figure 4: Pareto efficiency of ESCI graded similarity: Spearman correlation versus model size after embedding fine-tuning on the train split. OwlBERT models (Micro to Large) lie on the efficiency frontier and outperform Microsoft E5, BAAI BGE, ModernBERT, Ettin and EmbeddingGemma at comparable parameter scales.

Table 4: WANDS product search relevance after fine-tuning on the train query split and evaluation on held-out test queries. Bold marks the best score in each column.

Model	Params	NDCG@5	NDCG@10	MAP
OwlBERT				
Micro	17M	0.9122	0.9207	0.9722
Mini	68M	0.9157	0.9271	0.9739
Base	150M	0.9187	0.9314	0.9770
Large	400M	0.9239	0.9331	0.9774
General encoder baselines				
ModernBERT-base	150M	0.9113	0.9243	0.9729
ModernBERT-large	400M	0.9198	0.9310	0.9749
BERT-mini	11M	0.8771	0.8962	0.9609
BERT-small	29M	0.8940	0.9108	0.9688
Pre-trained reranker baselines				
GTE reranker–ModernBERT-base	150M	0.9137	0.9271	0.9740
BGE reranker-base	278M	0.9127	0.9227	0.9725

Table 6: GLUE retention results comparing OwlBERT models to contemporary baselines across five tasks. Bold indicates best performance within each model size category.

	Large		*Base*		*Mini*		*Micro*	
Task	OwlBERT	ModernBERT	OwlBERT	ModernBERT	OwlBERT	DistilBERT	OwlBERT	ettin-17m
SST-2	95.76	**96.67**	94.27	**94.72**	**92.78**	89.23	**90.37**	88.42
QNLI	94.00	**94.82**	**92.75**	90.88	**91.51**	87.99	**85.59**	85.30
QQP	**89.84**	89.73	**89.28**	89.26	**88.52**	87.42	**85.90**	85.54
MNLI	**92.06**	90.14	87.23	**88.71**	**85.23**	80.75	**78.26**	77.66
STS-B	**91.80**	91.33	**89.55**	85.92	**88.53**	84.84	**83.00**	82.62

7 Discussion

The results in §6 show that when evaluation is strongly e-commerce–shaped, *data and curriculum* dominate raw scaling: OwlBERT improves over general-purpose encoders at comparable or larger parameter counts, suggesting that representation quality in this domain is bottlenecked less by model capacity and more by exposure to the right long-tail entities, attribute-value constructions, and training distribution.

Why does specialization help? E-commerce language differs from generic web text in systematic ways: it is entity-dense, highly compositional, and often expressed in semi-structured fragments. The

Table 7: Held-out e-commerce perplexity comparison at the Base scale after full EcomWeb Phase 3 training. Lower is better.

Checkpoint	Guided MLM	PPL
Ettin Phase 2	no	6.04
Domain MLM, no Guided MLM	no	4.93
OwlBERT Phase 3	yes	3.99

EcomWeb corpus is explicitly designed to cover this space, and the gains we observe are most pronounced in settings that stress these properties. This supports the hypothesis that the encoder benefits from learning (i) robust lexical representations for domain terms and structure, and (ii) contextual rules that bind attributes to the correct head entities.

Is annealing enough for adaptation? Because OwlBERT starts from the Ettin Phase 2 checkpoint, its improvements are attributable to the domain-annealing stage rather than to full e-commerce pretraining from scratch. This is a practical advantage: many domains may not have the compute budget to repeat general pretraining, but can afford a focused distribution shift over high quality in-domain data. Downstream transfer (§6) and held out e-commerce perplexity at the Base scale provide the key evidence: comparing the un-annealed Ettin Phase 2 checkpoint with continued domain MLM and the final Phase 3 OwlBERT checkpoint before any task-specific fine-tuning, perplexity drops monotonically from 6.04 to 4.93 to 3.99, showing that domain annealing shifts the inherited encoder toward product-language distributions.

Guided MLM as a domain signal amplifier. Guided MLM concentrates MLM signal on rare but retrieval-critical spans such as product names, materials, sizes and colours. The ablations in §6.4 and Table 7 show that this bias improves product-language fit without replacing standard random masking entirely.

Practical deployment implications. The observed gains translate directly to several production facing use cases. First, stronger token recovery on titles/descriptions suggests improved representations for attribute completion and normalization, which can support catalog quality pipelines. Second, higher Spearman correlation on graded relevance implies more faithful embedding geometry for retrieval and candidate generation, especially in distinguishing *substitute* vs. *complement* relations that generic models often conflate.

8 Conclusion

We presented **OwlBERT**, domain-adapted encoders for e-commerce product understanding initialized from Ettin and specialized on **EcomWeb** via Phase 3 annealing and **Guided MLM**. Downstream evaluations show consistent gains on product-language and relevance benchmarks over general encoders and task-matched baselines, with competitive GLUE retention. These results support domain annealing as a practical recipe: inherit a strong long context encoder, shift its training distribution toward curated in-domain data, and target salient product spans without repeating general pretraining from scratch. Future work includes broader multilingual coverage and tighter contamination reporting.

Limitations

OwlBERT is evaluated primarily on English e-commerce data, so the results do not establish comparable gains for other languages, marketplaces, regions or time periods. Product relevance can vary across sellers and geographies, and category imbalance may cause the model to represent popular verticals better than sparse ones. The corpus curation pipeline depends on domain labels, LLM-assisted relevance judgments and classifier thresholds, each of which can introduce coverage gaps or systematic filtering bias. For reproducibility during review, we release anonymized OwlBERT checkpoints, EcomWeb artifacts, and Phase-3 pretraining code.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3615–3620.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law

school. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 2898–2904.

Yan Chen, Shujian Liu, Zheng Liu, Weiyi Sun, Linas Baltrunas, and Benjamin Schroeder. 2022. Wands: Dataset for product search relevance assessment. In *European Conference on Information Retrieval*, pages 128–141. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, volume 1 (long and short papers)*, pages 4171–4186.

Jonas Geiping and Tom Goldstein. 2023. Cramming: Training a language model on a single gpu in one day. In *International Conference on Machine Learning*, pages 11117–11143. PMLR.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Xiang Huang, Hao Peng, Dongcheng Zou, Zhiwei Liu, Jianxin Li, Kay Liu, Jia Wu, Jianlin Su, and Philip S Yu. 2024. Cosent: Consistent sentence embedding via similarity ranking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2800–2813.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

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

M-A-P, Ge Zhang, Xinrun Du, Zhimiao Yu, Zili Wang, Zekun Wang, Shuyue Guo, Tianyu Zheng, Kang Zhu, Jerry Liu, Shawn Yue, Binbin Liu, Zhongyuan Peng, Yifan Yao, Jack Yang, Ziming Li, Bingni Zhang, Minghao Liu, Tianyu Liu, and 6 others. 2024. Fine-FineWeb: A comprehensive study on fine-grained domain web corpus. Hugging Face. Version v0.1.0.

Tejaswini Mallavarapu, Ying Xie, and Simon Hughes. 2022. Catbert: An incrementally trained language representation model for e-commerce applications. In *Proceedings of the Interactive and Scalable Information Retrieval Methods for E-commerce Workshop*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Jacob Portes, Alexander Trott, Sam Havens, Daniel King, Abhinav Venigalla, Moin Nadeem, Nikhil Sardana, Daya Khudia, and Jonathan Frankle. 2023. Mosaicbert: A bidirectional encoder optimized for fast pretraining. *Advances in Neural Information Processing Systems*, 36:3106–3130.

Chandan K Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping queries dataset: A large-scale esci benchmark for improving product search. *arXiv preprint arXiv:2206.06588*.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J Pollard, Eric Lehman, Alistair EW Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

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

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders. *arXiv preprint arXiv:2507.11412*.

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

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

Denghui Zhang, Zixuan Yuan, Yanchi Liu, Fuzhen Zhuang, Haifeng Chen, and Hui Xiong. 2020. E-bert: A phrase and product knowledge enhanced language model for e-commerce. *arXiv preprint arXiv:2009.02835*.

A Data and Corpus Details

This appendix supplements §3. Table 8 lists FineFineWeb (M-A-P et al., 2024) domains kept directly when commerce-aligned, or filtered with LLM relevance labels distilled into per-domain fastText classifiers. The 350B-token mixture is deduplicated, English-filtered, and decontaminated against Amazon ESCI and WANDS before Phase 3 training.

Table 8: Selected domains from FineFineWeb. The left block lists domains requiring filtering, while the right block lists domains with almost complete commerce overlap.

Filter required	**Directly selected**
Hobby	Fashion
News	Beauty
Health	Celebrity
Entertainment	Movie
Travel	Photo
Food	Painting
Automotive	
Sports	
Music & Dance	

B Model Configurations

Table 9 lists the OwlBERT model configurations referenced in §5.

Table 9: OwlBERT model configurations.

Param.	Micro	Mini	Base	Large
Parameters	17M	68M	150M	400M
Layers	7	19	22	28
Hidden	256	512	768	1,024
FFN	384	768	1,152	2,624
Heads	4	8	12	16
Context	8,192	8,192	8,192	8,192
LR	$3e{-}3$	$3e{-}3$	$8e{-}4$	$5e{-}4$
WD	$3e{-}4$	$3e{-}4$	$1e{-}5$	$1e{-}5$

C Fine-tuning Details

This appendix supplements §6 with downstream fine-tuning protocols for Amazon ESCI and WANDS.

Data splits. ESCI train and test pairs are taken from the public Hugging Face release `https://huggingface.co/datasets/tasksource/esci`, retaining English (`product_locale`) examples with `small_version=1`: 419,653 train and 181,701 test query–product pairs (≈70/30; no dev split is provided). WANDS train, dev, and test splits are used directly from the public Hugging Face release `https://huggingface.co/datasets/napsternxg/wands` (60/20/20 by query: 140,068 train, 46,690 dev, and 46,690 test judgments; 233,448 labeled pairs in total). We do not resample or relabel either benchmark.

Amazon ESCI (bi-encoder). We train a Sentence-Transformers bi-encoder with CoSENTLoss on (query, `product_title`) pairs. ESCI labels {E, S, C, I} are mapped to {1.0, 0.66, 0.33, 0.0}. Optimization uses AdamW, a `NO_DUPLICATES` batch sampler, warmup ratio 0.1, bf16, and `load_best_model_at_end=True` with `save_strategy="epoch"`. Evaluation on the test split reports Pearson and Spearman correlation between predicted cosine similarity and the gold score.

WANDS (cross-encoder). We train a CrossEncoder with `num_labels=1` and BinaryCrossEntropyLoss on concatenated (query, `product_name`) pairs. Relevance labels {0, 1, 2} map to {0.0, 0.5, 1.0}. Training uses `max_seq_length=256`, warmup ratio 0.1, bf16, and `flash_attention_2` for ModernBERT-family checkpoints. At test time we score all labeled pairs, rerank within each query, and report NDCG@10, NDCG@5, MAP at relevance ≥ 1, and Spearman on queries with ≥ 2 distinct labels (n=421).

Hyperparameter search. Both setups use HuggingFace `Trainer` with `seed=42`. We sweep learning rate over {1e-5, 2e-5, 3e-5, 5e-5} (plus 5e-6 for the ESCI large model only) and epochs over {2, 3, 4, 5, 8, 10}. Per-GPU batch size is set by model capacity: 512/256/128 for bi-encoders and 128/64/32/16 for cross-encoders. We report the best peak/saved checkpoint per model.

D Full Token Recovery Results

This appendix reports the complete masked product token recovery results referenced in §6. Table 10 gives top-k accuracy on product titles and Table 11 on product descriptions at context lengths of 128, 256 and 512 tokens.

Table 10: Masked product token recovery accuracy on product titles.

Block Size	128			256			512		
Top-k token Accuracy	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
Large Models									
OwlBERT-large	**72.0**	**82.3**	**85.1**	**74.4**	**84.3**	**87.0**	**76.1**	**85.6**	**88.2**
ModernBERT-large	65.0	75.0	78.2	68.4	78.1	81.1	69.8	79.4	82.3
Base Models									
OwlBERT-base	**69.2**	**79.7**	**82.6**	**71.4**	**81.6**	**84.5**	**72.6**	**82.7**	**85.5**
ModernBERT-base	60.5	70.6	74.1	64.2	74.1	77.4	65.5	75.4	78.7
Mini Models (∼70M Parameters)									
OwlBERT-mini	**65.58**	**76.4**	**79.67**	**67.15**	**77.85**	**81**	**66.75**	**77.62**	**80.81**
DistilBERT	34.88	43.99	48.53	33.48	42.57	42.27	29.41	37.78	42.49
Micro Models (10–17M Parameters)									
OwlBERT-micro	**55.73**	**66.75**	**70.56**	**55.49**	**66.57**	**70.44**	**51.14**	**62.6**	**66.78**
BERT-mini	39.94	47.94	51.47	43.85	51.94	55.38	46.09	54.18	57.57

Table 11: Masked product token recovery accuracy on product descriptions.

Block Size	128			256			512		
Top-k token Accuracy	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
Large Models									
OwlBERT-large	**75.7**	**86.8**	**89.8**	**77.9**	**88.4**	**91.1**	**79.7**	**89.6**	**92.1**
ModernBERT-large	71.9	82.6	85.7	75.0	85.3	88.2	77.4	87.2	89.8
Base Models									
OwlBERT-base	**73.1**	**84.6**	**87.8**	**75.5**	**86.3**	**89.3**	**77.4**	**87.7**	**90.4**
ModernBERT-base	67.8	78.8	82.2	71.2	81.8	85.0	74.0	84.1	87.0
Mini Models (∼70M Parameters)									
OwlBERT-mini	**68.26**	**80.39**	**83.93**	**71.99**	**83.49**	**86.73**	**74.28**	**85.21**	**88.22**
DistilBERT	48.19	59.93	64.58	49.4	61.39	66.08	49.1	61.32	66.16
Micro Models (10–17M Parameters)									
OwlBERT-micro	**57.18**	**70.28**	**74.7**	**61.63**	**74.16**	**78.24**	**64.01**	**76.08**	**79.94**
BERT-mini	49.6	60.11	64.23	52.37	62.67	66.62	54.74	64.86	68.76