CONTRASTIVE LABEL-EMBEDDING ALIGNMENT FOR TEXT CLASSIFICATION WITHOUT LABELED DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dual-encoder text embedding models provide efficient, reusable representations, but adapting them to a target classification task typically requires labeled documents to fit a classifier head. Alternatives that avoid labeled documents, such as NLI cross-encoders and prompted LLMs, process each document jointly with its candidate labels and therefore cannot cache document representations. We introduce *contrastive label-embedding alignment* (CLEA), which reaches a point on this supervision-cost / inference-cost frontier that, to our knowledge, no prior method occupies: no labeled documents, no task-specific head, and fully pre-encodable dual-encoder inference. CLEA lightly fine-tunes a shared encoder using only label verbalizers and a small set of natural-language descriptions per label, aligning each verbalizer with its description set through a symmetric multi-positive contrastive objective. Across four benchmarks spanning topic, sentiment, intent, and emotion classification and ten encoders from 22M to 600M parameters, CLEA improves macro-F_1 by **+0.09** on average over untuned zero-shot embeddings, and on every benchmark the best CLEA-adapted encoder matches or exceeds NLI cross-encoders, a reranker, and prompted LLMs while retaining pre-encodable inference. Five descriptions per label suffice, and LLM-generated descriptions recover most of the gains of hand-written ones. In budget-matched experiments, CLEA outperforms linear probes until several dozen labeled examples per class are available.

1 INTRODUCTION

Text embedding models have become a default tool for representing text, with strong off-the-shelf performance on retrieval, semantic similarity, and clustering (Reimers & Gurevych, 2019; Gao et al., 2021; Muennighoff et al., 2023). For classification, however, the same models are usually not used directly: zero-shot use of embeddings tends to underperform, and practitioners typically train a classifier head on top of frozen embeddings (Neelakantan et al., 2022; Enevoldsen et al., 2025). Doing so reintroduces a per-document annotation cost that the appeal of "off-the-shelf" embedding models was meant to avoid.

A range of alternatives sit between full supervised fine-tuning and pure zero-shot use. Linear probes on frozen embeddings are cheap to train but require enough labeled examples to fit a classifier head; few-shot methods like SetFit further fine-tune a contrastive embedding model on a handful of labeled examples per class (Tunstall et al., 2022); in-context learning with large language models avoids fine-tuning entirely but pays a high inference cost (Brown et al., 2020; Liu et al., 2023); and natural-language inference cross-encoders cast classification as entailment without any labeled documents but require a joint forward pass for each document-label pair, preventing document representations from being cached and reused (Yin et al., 2019). These approaches form a spectrum trading off two distinct costs: *supervision cost* (how many labeled documents are required) and *inference cost* (how expensive each prediction is). At the cheap end of both axes, dataless classification has long shown that label semantics alone, expressed via class names, glosses, or short descriptions, can be enough to compare documents against labels in a shared semantic space (Chang et al., 2008; Gao et al., 2023; Chai et al., 2020; Meng et al., 2020). Yet most realizations of this idea rely on cross-encoder

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

architectures, which break the dual-encoder inference profile that makes embedding models attractive in the first place.

We argue that there is a missing point on this spectrum: a method that uses no labeled documents and retains efficient pre-encodable dual-encoder inference, while reaching accuracy competitive with substantially more expensive baselines. We propose *contrastive label-embedding alignment* (CLEA), a description-based supervision method tailored to dual-encoder embedding models. CLEA uses a small set of natural-language descriptions per label (five suffice in our experiments, written manually or generated by an LLM) as the sole supervision signal. Because this supervision scales with the label set rather than with corpus size, CLEA requires only lightweight fine-tuning of a shared encoder to align label verbalizers and their descriptions in a common embedding space via a symmetric multi-positive contrastive objective.

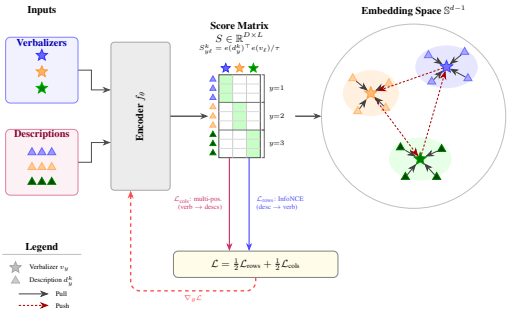


Figure 1: **Contrastive label-embedding alignment.** Descriptions and verbalizers are encoded by a shared encoder and optimized with rowwise and columnwise contrastive losses.

Across four classification benchmarks and ten encoders ranging from 22M to 600M parameters, CLEA improves macro-F_1 by an average of **+0.09** over zero-shot embedding use, and matches or exceeds NLI cross-encoders, prompted LLMs, and a few-shot SetFit baseline on every dataset we evaluate, while preserving the dual-encoder inference profile. CLEA also dominates few-shot linear probes up to non-trivial label budgets despite using no labeled documents. We further verify that the method scales to high-cardinality label spaces (the full 77-class Banking77 setting), is robust to noisy and imbalanced descriptions, and admits a fully label-free criterion for selecting the learning rate based on hyperspherical uniformity.

2 Related Work

Supervision forms for text classification. Methods for adapting embedding models to classification differ in what they require as supervision. *Linear probes* train a small classifier head on top of frozen embeddings: training is cheap but the head needs sufficient labeled documents to fit reliably (Neelakantan et al., 2022; Enevoldsen et al., 2025). *Few-shot methods* like SetFit further fine-tune the encoder contrastively on a small labeled set before training a classifier head, achieving strong results with modest compute (Tunstall et al., 2022); parameter-efficient variants reduce update cost but not annotation cost (Houlsby et al., 2019; Hu et al., 2022). *Dataless and zero-shot* methods replace labeled documents with semantic proxies: label names, seed words, knowledge bases (WordNet, Wikipedia), or short natural-language descriptions (Miller, 1995; Scott & Matwin, 1998; Gabrilovich & Markovitch, 2007; Chang et al., 2008; Wang et al., 2009; Yin et al., 2019; Chai et al., 2020; Meng et al., 2020; Gao et al., 2023). CLEA falls in this last category, using descriptions as supervision, but distinguishes itself in how that supervision is operationalized.

Inference profiles. Most entailment-based approaches to zero-shot classification use *cross-encoder* architectures, which jointly encode a document and a label hypothesis (Yin et al., 2019). This is accurate but expensive: inference cost grows linearly in the number of labels, and document representations cannot be cached. *Dual encoders* (Reimers & Gurevych, 2019; Gao et al., 2021; Wang et al., 2022; Li et al., 2023; Xiao et al., 2024; Choi et al., 2025; Zhang et al., 2025) encode documents and labels independently and reduce classification to similarity search, allowing both sides to be pre-computed and reused across deployments. The trade-off is well known: dual encoders are much cheaper at inference but typically weaker for fine-grained classification when used directly in zero-shot mode (Reimers & Gurevych, 2019; Fei et al., 2022). CLEA targets the dual-encoder regime specifically.

In-context learning with large models. Large language models can perform zero- or few-shot classification by listing label names and demonstrations directly in the prompt (Brown et al., 2020;

Dong et al., 2024; Luo et al., 2024). ICL avoids any task-specific training, but inference is expensive (the model processes the full prompt for every input), is sensitive to demonstration choice and ordering, and produces outputs that are non-trivial to constrain to a fixed label set.

Contrastive learning. InfoNCE-style contrastive objectives are widely used to align related views in a shared embedding space (van den Oord et al., 2018; Chen et al., 2020; Hadsell et al., 2006). In vision-language, CLIP applies this paradigm cross-modally to align images and captions (Radford et al., 2021). In text, contrastive objectives on augmented or paired sentences have produced strong sentence encoders (Reimers & Gurevych, 2019; Gao et al., 2021) and, scaled to large diverse corpora, the modern embedding models we evaluate in this paper (Wang et al., 2022; Li et al., 2023; Xiao et al., 2024). CLEA adopts this contrastive dual-encoder perspective but tailors the objective to a different kind of pair: label *verbalizers* aligned with sets of natural-language *descriptions* of the same label, with a symmetric multi-positive formulation that captures the many-to-one structure between a label and its descriptions.

3 Contrastive Label-Embedding Alignment

At a high level, our goal is to reshape an off-the-shelf dual-encoder text embedding model so that label verbalizers act as clean, well-positioned "representatives" of their labels in embedding space, purely based on natural-language descriptions of what those labels mean. Figure 1 provides a visual summary.

Setup and notation. Let $\mathcal{Y} = \{1, \dots, L\}$ denote the label set. For each $y \in \mathcal{Y}$ we assume two kinds of textual anchors:

- a short *verbalizer* v_y (e.g., for $y =$ "Sports" we might use $v_y =$ "This news snippet is about sports."), used at inference time to represent label y;[1]
- a small set of *label descriptions* $\mathcal{D}_y = \{d_y^k\}_{k=1}^{K_y}$, written as short paragraphs that clarify the kinds of documents y should cover.

We denote the union of all descriptions and its size by

$$\mathcal{D} = \bigcup_{y \in \mathcal{Y}} \mathcal{D}_y, \qquad D = \sum_{y \in \mathcal{Y}} K_y.$$

All training supervision comes from these verbalizers and descriptions, and no labeled documents are used.

Encoder and similarity scores. We use a single encoder f_θ with its native pooling map $\pi(\cdot)$[2]. Given a text t with T tokens, the encoder produces contextual token representations

$$f_\theta(t) \in \mathbb{R}^{T \times d}.$$

These are pooled and ℓ_2-normalized to obtain a sentence embedding

$$e(t) = \frac{\pi\big(f_\theta(t)\big)}{\big\|\pi(f_\theta(t))\big\|_2} \in \mathbb{R}^d,$$

so cosine similarity reduces to a dot product. With temperature $\tau > 0$, the similarity between a description d and a verbalizer v is

$$s(d, v) = \frac{e(d)^\top e(v)}{\tau}.$$

We reuse the base model's encoder architecture and pooling strategy, updating only θ and introducing no additional layers or task-specific heads.

[1] Appendix M studies variants that (i) omit the verbalizer and use the label text directly, or (ii) replace it with the mean embedding of the descriptions.

[2] We use the pooling native to the pretrained model, e.g., CLS-token, mean, or last-token pooling.

Batch structure and anchors. Each training batch considers the cross-product between all descriptions $\mathcal{D}$ and all verbalizers $\{v_1, \ldots, v_L\}$, forming the score matrix

$$S \in \mathbb{R}^{D \times L}, \qquad S^k_{y\ell} = s\big(d^k_y,\, v_\ell\big).$$

It is helpful to view this matrix from two complementary perspectives:

- *Row-anchors (descriptions).* Each row corresponds to a single description d^k_y and should assign high probability to its correct label y while treating other labels $\ell \neq y$ as negatives.
- *Column-anchors (verbalizers).* Each column corresponds to verbalizer v_ℓ and should collect probability mass from all of its positives $\{d^k_\ell\}_{k=1}^{K_\ell}$ while discounting descriptions of other labels.

3.1 Rowwise InfoNCE: Classifying descriptions into labels

From the rowwise viewpoint, each description d^k_y is a query that must identify its label y among all L options. The induced distribution over labels is

$$p(\ell \mid d^k_y) = \frac{\exp\{S^k_{y\ell}\}}{\sum_{j=1}^{L} \exp\{S^k_{yj}\}}.$$

The rowwise InfoNCE objective averages the cross-entropy against the correct label y:

$$\mathcal{L}_{\text{rows}} = \frac{1}{D} \sum_{y \in \mathcal{Y}} \sum_{k=1}^{K_y} \Big(\log \sum_{j=1}^{L} e^{S^k_{yj}} - S^k_{yy} \Big). \tag{1}$$

This is equivalent to a multiclass classifier over labels, where each description is a training example and the verbalizers serve as the representative embeddings for each class. Intuitively, equation 1 pulls each d^k_y toward its own verbalizer v_y while pushing it away from verbalizers $v_{\ell \neq y}$, tightening the alignment between descriptions and their labels.

3.2 Columnwise multi-positive InfoNCE: Aggregating over description sets

The rowwise objective treats each description independently. However, a label is not defined by a single canonical description, but by a *set* of complementary descriptions that cover different facets, edge cases, or typical failure modes. The columnwise objective explicitly models this many-to-one relation.

From the column perspective, each verbalizer v_ℓ has a set of positives $\mathcal{D}_\ell = \{d^k_\ell\}_{k=1}^{K_\ell}$, and all descriptions d^k_y with $y \neq \ell$ are negatives. We define the global and positive-partition normalizers

$$Z_\ell = \sum_{y \in \mathcal{Y}} \sum_{k=1}^{K_y} \exp\{S^k_{y\ell}\}, \qquad Z^+_\ell = \sum_{k=1}^{K_\ell} \exp\{S^k_{\ell\ell}\}.$$

The columnwise objective maximizes the aggregated positive mass relative to the global normalizer:

$$\mathcal{L}_{\text{cols}} = \frac{1}{L} \sum_{\ell=1}^{L} \big(\log Z_\ell - \log Z^+_\ell \big). \tag{2}$$

This is a *set-level* multi-positive term: it optimizes the combined probability of a label's positives rather than treating them as independent single-positive examples.

The log-sum-exp structure has two important consequences: the gradient

$$\frac{\partial \mathcal{L}_{\text{cols},\ell}}{\partial S^k_{\ell\ell}} = \frac{e^{S^k_{\ell\ell}}}{Z_\ell} - \frac{e^{S^k_{\ell\ell}}}{Z^+_\ell}$$

induces *adaptive within-positive weighting* proportional to $e^{S^k_{\ell\ell}}/Z^+_\ell$, emphasizing representative descriptions over noisy or idiosyncratic ones.[3] The ratio Z^+_ℓ/Z_ℓ is also normalized per label, keeping

[3] Appendix H provides an analysis of the impact of noisy descriptions.

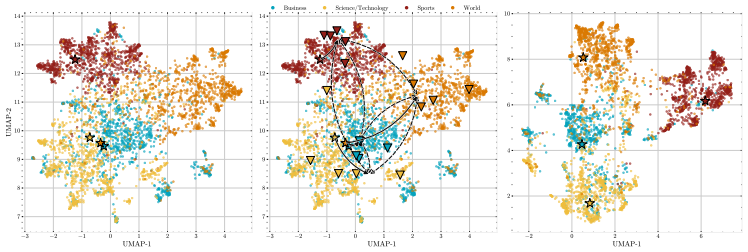


Figure 2: **AGNews (Zhang et al., 2015).** Left: embeddings before fine-tuning (stars denote label verbalizers). Middle: schematic of our training forces (triangles denote label descriptions). Right: embeddings after fine-tuning.

the objective stable when labels have different numbers of descriptions K_ℓ.[4] Geometrically, equation 2 pulls each v_ℓ toward the high-density region ("cloud") formed by its descriptions while repelling it from the description clouds of other labels.

3.3 Final objective and connection to standard InfoNCE

Our training loss is a simple symmetric combination of the rowwise and columnwise terms:

$$\mathcal{L} = \tfrac{1}{2}\,\mathcal{L}_{\text{rows}} + \tfrac{1}{2}\,\mathcal{L}_{\text{cols}}. \tag{3}$$

This symmetry encourages consistency in both directions: descriptions should clearly identify their label (rows), and each label should be well-explained by its description set (columns).[5]

In standard InfoNCE-style contrastive learning with paired batches, one forms a square similarity matrix whose diagonal entries are the unique positive pairs, and all off-diagonal entries act as in-batch negatives. In our setting, the score matrix $S \in \mathbb{R}^{D\times L}$ is rectangular, and positives are defined by label consistency rather than by the matrix diagonal: any cell (d_y^k, v_ℓ) with $y = \ell$ is a positive, while all cells with $y \neq \ell$ serve as in-batch negatives, yielding an $O(DL)$ softmax per batch. Unit-norm embeddings constrain optimization to the hypersphere, and the temperature τ controls the sharpness of the row- and column-softmax distributions. Following common practice in contrastive learning, we fix $\tau = 0.07$ (Gao et al., 2021; Chen et al., 2020; Radford et al., 2021).

3.4 Geometric intuition

Figure 2 illustrates the effect on AGNews with `all-MiniLM-L6-v2`. Before training (left), verbalizers (★) sit off-center relative to their document clouds and class regions overlap. The rowwise term pulls each description (▲) toward its verbalizer and away from the others, while the columnwise term moves each verbalizer toward the barycenter of its description cloud and away from other labels' descriptions (middle). Because the encoder is shared, the whole space is reshaped: after training (right), verbalizers sit in the densest part of their label's region and documents form tighter, better-separated clouds, although no document is used during training.

One way to read this is that dual encoders have a fixed representational budget (Weller et al., 2026), which off-the-shelf encoders allocate to open-ended retrieval and similarity rather than to any particular label partition. CLEA does not add capacity but rather reallocates existing capacity toward task-relevant directions, which helps explain why a small description-based signal can yield large classification gains.

3.5 Inference as dual-encoder classification

At test time, we use the encoder as a standard dual encoder for classification. Given a document x, we compute its embedding $e(x)$ and score labels by similarity to the verbalizers:

$$\text{score}(y \mid x) = e(x)^\top e(v_y), \qquad \hat{y} = \arg\max_{y\in\mathcal{Y}} \text{score}(y \mid x).$$

[4] Appendix I highlights this stability empirically.

[5] Appendix G provides an empirical analysis of the loss components.

Because both documents and verbalizers can be pre-encoded and stored, classification reduces to a nearest-neighbor search over label representations.

3.6 Label-free learning rate selection

With AdamW defaults and the batch fixed to one sweep over all description-verbalizer pairs, the learning rate (LR) is the main sensitive hyperparameter: overly large LRs can drive the embedding space toward representation collapse (Bardes et al., 2022), while overly small ones stall alignment. Following the view that contrastive learning balances alignment and uniformity on the hypersphere (Wang & Isola, 2020), and since alignment is directly enforced by the supervision signal, we select the LR that best preserves uniformity on an unlabeled pool $\mathcal{X}_u = \{x_i\}$ from the target domain.

Let $z_i = e(x_i)$ be ℓ_2-normalized embeddings and $t > 0$ a scale parameter. Define

$$\mathcal{L}_{\text{uni}}(t) = \log \mathbb{E}_{i \neq j}\left[e^{-t\|z_i - z_j\|_2^2}\right] \approx \log\left(M^{-1}\sum_{m=1}^{M} e^{-t\|z_{i_m} - z_{j_m}\|_2^2}\right), \tag{4}$$

where (i_m, j_m) are M random distinct pairs from $\mathcal{X}_u$. Lower values indicate a more uniform and less collapsed embedding space. We sweep candidate LRs and select the minimizer of $\mathcal{L}_{\text{uni}}$ with $t = 2$ (Wang & Isola, 2020). The criterion is label-free and computationally inexpensive, and lower values empirically correlate with stronger downstream performance across a range of models and datasets (Figure 11, Appendix O).[6]

Figure 3 illustrates AGNews document embeddings from *all-MiniLM-L6-v2* at the learning rate selected by this procedure. We reduce the embeddings to $\mathbb{R}^3$ with PCA and project them onto the unit sphere $\mathbb{S}^2$ using ℓ_2-normalization.[7]

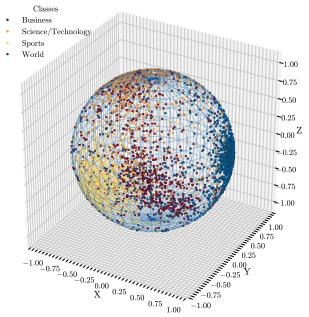


Figure 3: AGNews (Zhang et al., 2015) embeddings at the selected LR.

4 Experimental Setup

We evaluate on four text-classification benchmarks: topic (AGNews (Zhang et al., 2015)), emotion (EmotionDAIR (Saravia et al., 2018)), sentiment (RottenTomatoes (Pang & Lee, 2005)), and fine-grained intent (Banking77 (Casanueva et al., 2020)). For each dataset and class, we write exactly 5 short descriptions that characterize typical documents. Examples are provided in Appendix D, and the effect of the number of descriptions is studied in Appendix J. For *Banking77*, we report main results on a six-class subset of card-related intents to probe fine-grained distinctions within a narrow region of the label space.[8] High-cardinality behavior is studied separately on the full 77-class label space (Appendix N), where the dominant failure modes shift from boundary confusions between similar classes to coverage failures across many labels. We therefore focus on precision and recall alongside macro-F_1.

We test our method on ten pretrained dual-encoder text embedding models spanning a range of architectures and sizes, from roughly 22M to 600M parameters. Appendix E summarizes all models.

Training. Because description sets are small, we treat one sweep over all description–verbalizer pairs as an *epoch*. We use AdamW (Loshchilov & Hutter, 2019), training for at most 1000 iterations with early stopping on the training loss (patience $= 10$, tolerance $\Delta = 10^{-5}$, evaluated every 10 steps). Since the criterion is the training loss itself, no labeled validation data is required. Learning rates are swept over $\{1, 3, 5\} \times \{10^{-4}, 10^{-5}, 10^{-6}\}$ and selected using the uniformity criterion in equation 4, computed on 50,000 document pairs from an unlabeled held-out development split

[6] When an LR sweep is impractical, reusing the base model's pretraining learning rate provides a safe, non-optimized fallback. This heuristic performed reliably across many model–dataset combinations in our experiments.

[7] We avoid UMAP because its locality and crowding parameters can distort interpoint distances, making it less suitable for visualizing uniformity.

[8] We evaluate on: "expiration of a card", "arrival of a card", "card not working", "card payment fee", "payment the customer does not recognise", and "card swallowed by a machine".

Model	AGNews	Banking77	EmotionDAIR	RottenTomatoes	Avg
all-MiniLM-L6-v2	0.67	0.66	0.35	0.66	0.58
+ *CLEA*	0.79 (+0.12)	0.90 (+0.23)	0.43 (+0.09)	0.70 (+0.04)	0.70
e5-base-v2	0.75	0.79	0.44	0.83	0.70
+ *CLEA*	0.81 (+0.05)	0.96 (+0.17)	0.48 (+0.05)	0.82 (+0.00)	0.77
e5-large-v2	0.78	0.79	0.44	0.85	0.72
+ *CLEA*	0.82 (+0.03)	**0.96** (+0.17)	0.53 (+0.09)	0.86 (+0.00)	0.79
bge-base-en-v1.5	0.63	0.86	0.42	0.81	0.68
+ *CLEA*	0.82 (+0.19)	0.95 (+0.09)	0.47 (+0.06)	0.82 (+0.01)	0.77
bge-large-en-v1.5	0.75	0.84	0.44	0.82	0.71
+ *CLEA*	0.82 (+0.07)	0.95 (+0.10)	0.56 (+0.12)	0.85 (+0.03)	0.80
gte-base-en-v1.5	0.73	0.86	0.44	0.84	0.72
+ *CLEA*	0.83 (+0.09)	0.95 (+0.08)	0.49 (+0.06)	0.85 (+0.01)	0.78
gte-modernbert-base	0.75	0.88	0.45	0.82	0.73
+ *CLEA*	0.80 (+0.05)	0.94 (+0.06)	0.49 (+0.04)	0.84 (+0.03)	0.77
gte-large-en-v1.5	0.72	0.92	0.40	0.87	0.73
+ *CLEA*	0.83 (+0.11)	0.95 (+0.03)	0.50 (+0.10)	0.83 (-0.04)	0.78
Qwen3-Embedding-0.6B	0.63	0.88	0.48	0.76	0.69
+ *CLEA*	**0.85** (+0.22)	0.92 (+0.03)	0.57 (+0.08)	**0.88** (+0.12)	0.80
embeddinggemma-300m	0.53	0.80	0.49	0.64	0.61
+ *CLEA*	0.74 (+0.21)	0.94 (+0.14)	**0.57** (+0.09)	0.73 (+0.09)	0.74

Table 1: Main results by model family. Each model has a base zero-shot row and an + *CLEA* row; parentheses report absolute F_1 gains. Best + *CLEA* F_1 per dataset is **bold**, and each model's largest gain is underlined.

constructed from the training portion of each dataset.[9] To improve stability in this few-description supervision setting, we use linear warmup during the first 50% of training steps.

Evaluation. To ensure comparability across tasks with different label cardinalities and class balances, we use **macro F_1** as our primary metric, which gives equal weight to every class and is appropriate for both balanced and imbalanced multi-class settings (Sokolova & Lapalme, 2009). Reported macro-F_1 values are averaged over 3 random seeds.

5 Results and Analysis

We begin with the overall effect of label alignment across all models and datasets. Averaging over ten encoders and four benchmarks (Table 1), CLEA improves macro-F_1 from roughly 0.68 to 0.77, i.e. by about $+\mathbf{0.09}$ absolute. The gains are consistent: every (model, dataset) pair shows a non-negative change except for one slight regression (gte-large on RottenTomatoes, -0.04).

Where the gains are concentrated. Improvements vary across tasks in ways that align with label semantics. Topic and intent classification benefit the most, with mean gains of $+\mathbf{0.12}$ on **AGNews** and $+\mathbf{0.11}$ on **Banking77**. Several encoders gain more than 15–20 macro-F_1 points, suggesting that a few short descriptions can effectively characterize what makes a topic or intent distinct. **EmotionDAIR** (mean $+\mathbf{0.07}$) and **RottenTomatoes** (mean $+\mathbf{0.03}$) show smaller gains, in part because the top encoders already perform near their ceilings on RottenTomatoes (E5/GTE/BGE start at 0.81-0.87) and in part because emotion recognition involves subtle, context-dependent cues that may need richer or larger description sets, consistent with Appendix J where EmotionDAIR continues to benefit from increasing the description count well beyond $K = 5$. Across families, smaller models often realize the largest *relative* improvements (all-MiniLM-L6-v2 averages $+\mathbf{0.12}$, embeddinggemma-300m $+\mathbf{0.13}$), but strong encoders also benefit substantially: Qwen3-Embedding-0.6B averages $+\mathbf{0.12}$ and reaches the best overall macro-F_1 on AGNews (0.85) and RottenTomatoes (0.88). Averaging description embeddings into label prototypes without fine-tuning already helps on some tasks, but

[9] We compute $\mathcal{L}_{\text{uni}}$ from randomly sampled document pairs using only the raw inputs x_i (no labels are accessed or used at any point). For completeness, to mimic a deployment scenario in which no development set is available, we also repeat LR selection using only unlabeled texts from the target-domain test split as a proxy for the deployment distribution, and obtain the same qualitative conclusions and similar macro-F_1 scores.

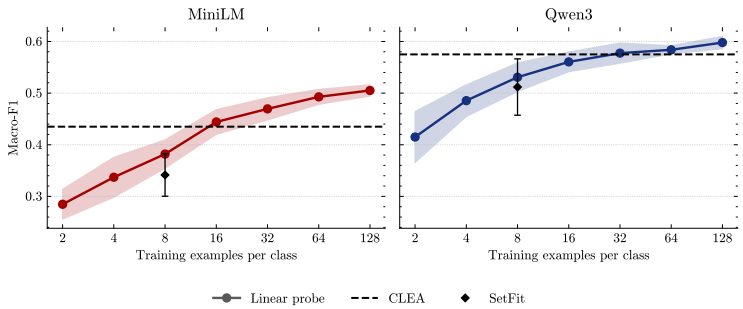


Figure 5: Supervision-cost frontier on EmotionDAIR. Linear probes on frozen embeddings (curves, ± 1 std. over 20 random label draws) at increasing labeled-example budgets, compared to CLEA (dashed) and SetFit-8 (diamond). CLEA uses no labeled documents.

inconsistently: CLEA exceeds the prototype in 18 of 20 (model, dataset) pairs, with the largest margin on EmotionDAIR, where averaged descriptions bring no improvement at all (Appendix K).

Comparison across inference regimes. We next compare CLEA against stronger zero-shot classification architectures: NLI cross-encoders, a reranker, and prompted LLMs (Appendix F, Table 5). Multiple CLEA-trained dual encoders match or exceed the best alternative regime on every dataset. On AGNews, Qwen3-Embedding + CLEA reaches 0.85, tying the best LLM and beating both NLI cross-encoders and the reranker, with several other CLEA encoders (e5-large, gte-base, gte-large) close behind and ahead of the strongest cross-encoder. On Banking77, e5-base + CLEA and e5-large + CLEA both reach 0.96, well above the best LLM (0.91), the best cross-encoder (0.86), and the reranker (0.87). On EmotionDAIR and RottenTomatoes the picture is similar, with CLEA-trained encoders leading by 5–13 macro-F_1 points. All while retaining the dual-encoder inference profile, where document and label representations are pre-encoded once and reused. Notably, the same Qwen3-Embedding-0.6B that underperforms its reranker counterpart Qwen3-Reranker-0.6B on all four datasets out of the box surpasses it on all four after CLEA.

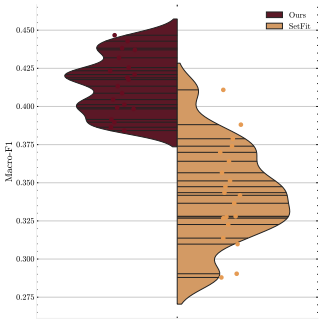


Figure 4: SetFit (Tunstall et al., 2022) comparison on EmotionDAIR.

Stability under data and description sampling. A second axis worth examining is run-to-run variability. We compare CLEA against SetFit (Tunstall et al., 2022), a widely used few-shot method that combines contrastive pretraining with a lightweight classifier head, on EmotionDAIR with 8 samples per class. Each method is run 20 times following its recommended configuration: SetFit varies the random draw of 8 labeled examples per class, and CLEA varies the random draw of 5 descriptions per label from a pool of 20 candidates. Figure 4 shows that CLEA achieves a higher mean macro-F_1 and substantially smaller variance: while SetFit can occasionally match it, it exhibits a long tail of poor outcomes that reflects sensitivity to the specific few-shot samples drawn.

Comparison across supervision budgets. To position CLEA on the supervision-cost spectrum more broadly, we add a linear probe baseline at increasing label budgets: 2, 4, 8, 16, 32, 64, and 128 examples per class, each run 20 times with different random label draws. Figure 5 reports the resulting curves on EmotionDAIR for two encoders at different scales (all-MiniLM-L6-v2, 22M; Qwen3-Embedding-0.6B, 600M). The picture is consistent across both: CLEA, using no labeled documents, outperforms linear probes up to roughly 16 examples per class on MiniLM and 32-64 examples per class on Qwen3, before linear probes eventually surpass it given enough labels. SetFit at 8 labeled examples sits below both CLEA and a linear probe at the same budget, suggesting that its contrastive fine-tuning stage does not provide a net benefit over a simple classifier head at this label budget. CLEA therefore occupies a useful point in the supervision-cost spectrum: the strongest method we evaluate when labeled documents are not available, and competitive with budget-matched supervised baselines until labels become abundant.

Robustness to description source. Manual description authoring is the main human effort required by CLEA. We test whether this cost is essential by replacing human-written descriptions with descriptions generated with GPT-5.1 (OpenAI, 2025) (Appendix L, Figure 9). Across two encoders and two datasets, LLM-generated descriptions recover most of the gains of human-written ones, with a small but consistent residual benefit from manual curation. The stability advantage over SetFit also persists: even with sampled descriptions, CLEA's macro-F_1 distributions are markedly tighter than those of SetFit at 8 labeled examples per class, indicating that the low-variance behavior of CLEA is not an artifact of using a single fixed human-authored description set.

6 Conclusion

We showed that the task-specific supervision needed to adapt a dual encoder to a classification task can be moved from documents to the label space. A handful of natural-language descriptions per label, aligned with the label verbalizers through a symmetric multi-positive contrastive objective, suffices to reshape the shared embedding space around the target labels. Although no document is used during training, documents inherit the new geometry, and the best adapted encoders match or exceed cross-encoders, rerankers, and prompted LLMs that incur substantially higher inference costs, while remaining competitive with supervised baselines until labeled documents number in the tens per class. Averaging the same descriptions into prototypes helps only inconsistently, indicating that the alignment step, rather than the descriptions alone, is what makes this signal reliable. Taken together, these results suggest that the weak direct classification performance of off-the-shelf embedding models reflects how their representational capacity is allocated rather than a lack of capacity, and that this allocation can be corrected cheaply without giving up pre-encodable inference.

Limitations and future work. Descriptions remain a form of supervision. They scale with the number of labels rather than with the corpus and can be generated by an LLM at a small cost in accuracy, but writing good ones still requires understanding the task. When labeled documents become abundant, linear probes and fully supervised fine-tuning eventually surpass CLEA; our method targets settings where labeled documents are scarce, costly, or unavailable. Our evaluation covers English benchmarks with short to medium-length texts, and extending it to multilingual encoders and longer documents is left to future work. The uniformity-based learning rate criterion is a practical label-free heuristic that correlates with downstream performance in most cases but is not a theoretically grounded selector, and a natural next step is to study its role more deeply, both empirically and as a possible component of the training objective. Broader impacts are discussed in Appendix A.

AI use statement

Generative AI models were used as objects of study to generate label descriptions in the experiments described in the "Robustness to description source" analysis and Appendix L. During manuscript preparation, generative AI tools were used to correct spelling and grammar and to rephrase sentences for clarity. The study design, analyses, interpretations, and conclusions are the authors' own. The authors take full responsibility for the final manuscript.

Ethics statement

The potential societal impacts of this work are discussed in Appendix A.

Reproducibility statement

We provide the full formulation of CLEA in Section 3, including the training objective and inference procedure, and describe the label-free learning-rate selection procedure in Section 3.6. Section 4 specifies the datasets, training procedure, hyperparameter search, early-stopping criterion, warmup schedule, and evaluation protocol. Further details on the datasets and pretrained encoders are provided in Appendices B and E, respectively, while the construction of the label verbalizers and descriptions, with examples for AGNews, is described in Appendix D. Compute resources are reported in Appendix C. All datasets and pretrained encoders used in the experiments are publicly available. We will release the code and experimental configurations with the camera-ready version.

REFERENCES

Ilias Aarab. BTZSC: A benchmark for zero-shot text classification across cross-encoders, embedding models, rerankers and LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IxMryAz2p3.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5/.

Duo Chai, Wei Wu, Qinghong Han, Fei Wu, and Jiwei Li. Description based text classification with reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1371–1382. PMLR, 2020. URL https://proceedings.mlr.press/v119/chai20a.html.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 830–835. AAAI Press, 2008.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Min Choi, Sahil Dua, and Alice Lisak. Introducing EmbeddingGemma: The best-in-class open model for on-device embeddings. Google Developers Blog, September 2025. URL https://developers.googleblog.com/en/introducing-embeddinggemma/. Accessed September 2025.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL https://aclanthology.org/2024.emnlp-main.64/.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, et al. MMTEB: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zl3pfz4VCV.

Yu Fei, Zhao Meng, Ping Nie, Roger Wattenhofer, and Mrinmaya Sachan. Beyond prompting: Making pre-trained language models better zero-shot learners by clustering representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8560–8579, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.587. URL https://aclanthology.org/2022.emnlp-main.587/.

Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1606–1611, 2007. URL https://www.ijcai.org/Proceedings/07/Papers/259.pdf.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. The benefits of label-description training for zero-shot text classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13823–13844, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.853. URL https://aclanthology.org/2023.emnlp-main.853/.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552/.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006. doi: 10.1109/CVPR.2006.100.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. Building efficient universal classifiers with natural language inference. arXiv preprint arXiv:2312.17543, 2023. URL https://arxiv.org/abs/2312.17543.

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

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):195:1–195:35, 2023. doi: 10.1145/3560815. URL https://doi.org/10.1145/3560815.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=NQPo8ZhQPa.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9006–9017, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.724. URL https://aclanthology.org/2020.emnlp-main.724/.

George A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. doi: 10.1145/219717.219748.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training. arXiv:2201.10005, 2022. URL https://arxiv.org/abs/2201.10005.

OpenAI. GPT-5.1. OpenAI API documentation, 2025. URL https://developers.openai.com/api/docs/models/gpt-5.1. Accessed: 2025-12-06.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 115–124, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219855. URL https://aclanthology.org/P05-1015/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3687–3697, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://aclanthology.org/D18-1404/.

Sam Scott and Stan Matwin. Text classification using WordNet hypernyms. In *Usage of WordNet in Natural Language Processing Systems*, 1998. URL https://aclanthology.org/W98-0706/.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts. arXiv preprint arXiv:2209.11055, 2022. URL https://arxiv.org/abs/2209.11055.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

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

Pu Wang, Jian Hu, Hua-Jun Zeng, and Zheng Chen. Using Wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, 19(3):265–281, 2009. doi: 10.1007/s10115-008-0152-4.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=k9CzIvzfaA.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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. 3914–3923, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1404. URL https://aclanthology.org/D19-1404.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176, 2025. URL https://arxiv.org/abs/2506.05176.

A Broader societal impacts

CLEA may make text classification more accessible by improving classification performance without requiring labeled documents while preserving efficient dual-encoder inference. This can lower the cost of prototyping and deploying classifiers in low-resource settings, specialized domains, and hardware-constrained environments. Its supervision signal also has a practical transparency advantage: rather than relying on large collections of labeled examples, CLEA uses a small set of natural-language descriptions per label, which practitioners can directly inspect, critique, and revise.

At the same time, the method does not remove the risks associated with automated classification. Incorrect or biased predictions can cause harm in consequential settings such as content moderation, hiring, or credit decisions. Moreover, label descriptions encode assumptions about the scope and meaning of each class, and model behavior will reflect these assumptions. Human-written descriptions may embed subjective or domain-specific biases, while LLM-generated descriptions may additionally inherit biases from the generating model and from the prompts used to produce them.

These risks can be partly mitigated because descriptions are short, human-readable artifacts. Before deployment, they should be reviewed by domain experts, audited for representational balance across relevant demographic and contextual dimensions, and updated as the deployment context changes. We recommend treating label descriptions as auditable model inputs rather than fixed ground truth, conducting robustness and bias checks across relevant subpopulations and distribution shifts, and limiting CLEA to decision-support roles in high-stakes settings rather than using it as the sole decision mechanism.

B Datasets Overview

Table 2 summarizes the datasets used in our experiments, including their domains, number of classes, sources, and license terms. All datasets are publicly available via Hugging Face Datasets[10]. For label verbalizers, we follow the setup of (Laurer et al., 2023).

Our selection deliberately covers four common zero-shot classification families: emotion recognition on social media (*EmotionDAIR*), intent detection in a narrow banking domain (*Banking77*), binary sentiment analysis for movie reviews (*RottenTomatoes*), and topic classification in news (*AGNews*). This yields a mix of short, informal texts (EmotionDAIR, Banking77), longer and more descriptive documents (AGNews), and highly domain-specific language (banking vs. movies vs. general news). For all datasets we use the official test splits provided by the original authors (as exposed through Hugging Face).

Task	Domain	Dataset	# Classes	Source	License
Emotion	Social media	EmotionDAIR	6	Saravia et al. (2018)	Research/education only
Intent	Banking	Banking77	77 (6 main[11])	Casanueva et al. (2020)	CC BY 4.0
Sentiment	Movies	RottenTomatoes	2	Pang & Lee (2005)	CC0 1.0
Topic	News	AGNews	4	Zhang et al. (2015)	Non-commercial

Table 2: Datasets used in the evaluation, covering emotion recognition, intent detection, sentiment analysis, and topic classification. For *Banking77* we report main results on a six-class card-related subset (Section 4); the full 77-class evaluation is in Appendix N.

C Compute resources

Experiments were run on an internal cluster of 12 nodes, each equipped with NVIDIA A100 80GB GPUs, with inference in `bfloat16`. The cluster was used to parallelize across the (model, dataset) grid; no individual experiment required more than a single A100. CLEA fine-tuning takes between one and a few minutes per (model, dataset) configuration. The most expensive baseline runs are the linear-probe budget sweeps in Figure 5 (seven budgets $\times$ 20 random label draws $\times$ 2 encoders). Prompted-LLM baselines (Qwen3-4B, Qwen3-8B) run on the same hardware as inference-only workloads and GPT-5.1 description generation was performed via the OpenAI API.

D Label verbalizers and descriptions

Our method assumes a short verbalizer and a small set of natural-language descriptions for each label. The verbalizer plays the role of a compact, sentence-level anchor (e.g., "This example news text is about world news"), while the descriptions provide richer, paragraph-level semantics that spell out the kinds of documents the label should cover. In practice, we construct these descriptions manually, starting from the label verbalizers of Laurer et al. (2023) and expanding each label into multiple complementary paraphrases. For AGNews, shown below, we write 5 descriptions per class that: (i) emphasize different aspects of the underlying topic (e.g., actors, events, context), (ii) avoid overlap with other labels (e.g., business vs. world news vs. sports), and (iii) remain generic enough to apply across articles and time periods.

[10] `https://huggingface.co/datasets`

[11] The six card-related intents are: "expiration of a card", "arrival of a card", "card not working", "card payment fee", "payment the customer does not recognise", and "card swallowed by a machine".

These descriptions are used only during the description-only training stage. No dataset examples or ground-truth labels are required to construct them. At inference time, the model receives only raw documents and the label representations (verbalizers, label names, or mean description embeddings; see Appendix M), with no labeled documents used at any stage. Table 3 illustrates the resulting description sets for AGNews. Analogous description collections are created for Banking77, EmotionDAIR, and RottenTomatoes.

Category (verbalizer)	Descriptions
World News *Verbalizer:* "This example news text is about world news."	• Coverage of international affairs and geopolitics: governments, elections, diplomacy, conflicts, treaties, and sanctions. Stories focus on cross-border events and their global implications rather than domestic business or sports outcomes. • News about countries interacting on the world stage: summits, UN resolutions, regional alliances, and humanitarian crises. Emphasis is on state actors, policy decisions, and shifts in international relations. • Reporting on wars, ceasefires, peace talks, and military deployments across regions. Articles highlight causes, stakeholders, civilian impact, and reactions from other nations or international bodies. • Global society and policy issues such as migration, human rights, climate diplomacy, and development aid. Pieces track how multiple countries respond and coordinate. • International incidents and disasters (natural or man-made) where response, accountability, and cross-national coordination are central. Focus remains on worldwide context rather than local business ramifications.
Sports *Verbalizer:* "This example news text is about sports."	• Results, previews, and analysis of professional or amateur competitions: matches, tournaments, standings, and championships. Content centers on performance, tactics, and outcomes on the field. • Athlete-focused updates including injuries, transfers, contracts, and retirements. Stories emphasize team dynamics and competitive impact. • Coverage of leagues and events: scheduling, rule changes, drafts, and officiating controversies. The angle is sporting governance and competitive fairness. • Game recaps and statistical breakdowns highlighting key plays, records, and milestones. The narrative ties individual performances to team results. • Profiles and human-interest features about coaches, players, and training methods. Emphasis is on preparation, strategy, and competitive psychology.
Business *Verbalizer:* "This example news text is about business news."	• Corporate news: earnings, revenue guidance, layoffs, executive changes, and strategic shifts. Articles assess company performance and shareholder impact. • Markets and finance coverage: stocks, bonds, commodities, currencies, and macro sentiment. Focus is on price moves, drivers, and investor reactions. • Mergers, acquisitions, IPOs, and venture funding. Pieces explain valuations, synergies, and regulatory hurdles. • Industry developments such as competition, supply chains, pricing, and business models across sectors. Reporting connects firm-level actions to market structure. • Policy and regulation affecting commerce: antitrust cases, trade policy, taxes, and compliance. The lens is how rules shape corporate behavior and profitability.
Science & Technology *Verbalizer:* "This example news text is about science and technology."	• Scientific research findings across fields like biology, physics, medicine, and climate science. Articles emphasize methods, evidence, and potential applications or limitations. • Technology product and platform news: hardware, software, mobile, cloud, and consumer gadgets. Coverage focuses on features, performance, and user impact. • AI, data science, and computing breakthroughs including models, chips, algorithms, and benchmarks. Stories discuss capabilities, risks, and real-world use cases. • Space and astronomy updates: launches, missions, telescopes, and planetary discoveries. Coverage highlights scientific goals and engineering challenges. • Cybersecurity and privacy incidents: vulnerabilities, breaches, hacks, and defenses. Reporting centers on technical cause, affected users, and mitigations.

Table 3: AGNews (Zhang et al., 2015) class verbalizer and class descriptions (5 per class).

E EMBEDDING MODELS

Our benchmark covers ten publicly available embedding models spanning different architectures, parameter scales, and training paradigms (Table 4). We include (i) smaller sentence-transformer style encoders such as `all-MiniLM-L6-v2`, (ii) recent E5 and BGE models trained on large synthetic contrastive corpora, (iii) the GTE family including a ModernBERT-based variant, and (iv) two more recent architectures targeting embedding quality, `embeddinggemma-300m` and `Qwen3-Embedding-0.6B`. This mix allows us to test CLEA on both lightweight models that are attractive in low-compute settings and larger, state-of-the-art encoders that already perform strongly in zero-shot retrieval and classification benchmarks.

For each model, Table 4 reports publication year, base architecture (encoder vs. decoder), backbone, main pre-training or fine-tuning corpus, parameter count, pooling strategy, and embedding dimensionality. We always use the pooling operation recommended by the model authors (e.g., mean pooling for E5, CLS pooling for GTE/BGE), and apply our description-only training on top of the publicly released checkpoints without any additional task-specific pretraining. The diversity of architectures and training recipes allows us to assess how robust our method is across model families, as presented in the main results (Table 1).

Model	Yr	Arch.	Backbone	FT / train data	# P	Pool	Dim
`all-MiniLM-L6-v2`	2021	enc.	MiniLM	1B paired sentences	22.7M	mean	384
`e5-base-v2`	2023	enc.	E5 (BERT)	270M synthetic contrastive	110M	mean	768
`e5-large-v2`	2023	enc.	E5 (BERT)	same as above	335M	mean	1024
`bge-base-en-v1.5`	2023	enc.	BGE (RoB.)	1.5B pair data, contrastive	109M	CLS	768
`bge-large-en-v1.5`	2023	enc.	BGE (RoB.)	same as above	434M	CLS	1024
`gte-base-en-v1.5`	2024	enc.+	GTE	MLM + contrastive pre-train	137M	CLS	768
`gte-large-en-v1.5`	2024	enc.+	GTE	same as above	434M	CLS	1024
`gte-modernbert-base`	2024	enc.	ModernBERT	same as above	149M	CLS	768
`embeddinggemma-300m`	2025	enc.	Gemma 3 (enc.)	Multiling. corpus (320B tok), contrastive	308M	mean	768*
`Qwen3-Embedding-0.6B`	2025	dec.	Qwen3	synthetic multiling. contrastive	0.6B	last	1024

Table 4: Architectural and training overview of the 10 embedding models used. Columns list publication year (Yr), encoder/decoder architecture (Arch.), backbone, principal fine-tuning (FT) or pre-training data, parameter count (#P), pooling strategy (Pool), and embedding dimensionality (Dim). *For `embeddinggemma-300m`, dimensionality 768 corresponds to Matryoshka Representation Learning (MRL) with nested sizes 512/256/128, a training scheme enabling shorter embeddings.

F COMPARISON ACROSS INFERENCE REGIMES

Table 5 compares CLEA against three alternative regimes that likewise require no labeled documents: NLI cross-encoders, rerankers, and prompted LLMs. We follow the experimental setup of (Aarab, 2026) for each regime, and summarize the relevant choices below for self-containment.

NLI cross-encoders. We evaluate `deberta-v3-large-nli` (Laurer et al., 2023) and `modernbert-large-nli`, both fine-tuned on natural-language inference and routinely used for zero-shot classification. For each (document, label) pair, the model scores whether the document entails the label hypothesis expressed as a verbalizer (e.g., "This news snippet is about sports."). The label with the highest entailment probability is selected as the prediction. Inference requires L joint forward passes per document, none of which can be cached across documents.

Reranker. We evaluate `Qwen3-Reranker-0.6B` (Zhang et al., 2025), the reranker counterpart of `Qwen3-Embedding-0.6B` from the same family. The text to be classified serves as the query, and the verbalized label descriptions are treated as the candidate "documents" to be reranked. For Qwen3-Reranker, which is generative, the relevance score for each (query, candidate label) pair is derived from the model's output token probabilities, and the predicted label is the one with the highest score. Including the dual-encoder embedding version of the same backbone in the dual-encoder block of Table 5 allows a controlled within-family comparison: the same backbone deployed as a (cached) embedding model versus deployed as a (joint-pass) reranker.

Model	Params	AGNews	Banking77	EmotionDAIR	RottenTom.
Dual encoders					
all-MiniLM-L6-v2	22M	0.67	0.66	0.35	0.66
+ *CLEA*		0.79	0.90	0.43	0.70
e5-base-v2	109M	0.75	0.79	0.44	0.83
+ *CLEA*		0.81	**0.96**	0.48	0.82
bge-base-en-v1.5	109M	0.63	0.86	0.42	0.81
+ *CLEA*		0.82	0.95	0.47	0.82
e5-large-v2	335M	0.78	0.79	0.44	0.85
+ *CLEA*		0.82	**0.96**	0.53	0.86
Qwen3-Embedding-0.6B	600M	0.63	0.88	0.48	0.76
+ *CLEA*		**0.85**	0.92	**0.57**	**0.88**
NLI cross-encoders					
deberta-v3-large-nli	304M	0.81	0.86	0.44	0.85
modernbert-large-nli	395M	0.76	0.83	0.30	0.82
Reranker					
Qwen3-Reranker-0.6B	600M	0.79	0.87	0.41	0.78
LLMs (multiple-choice prompting)					
Qwen3-4B	4B	0.82	0.91	0.43	0.87
Qwen3-8B	8B	**0.85**	0.91	0.40	0.86

Table 5: Comparison across inference regimes. Dual encoders allow label representations to be pre-encoded and reused, while cross-encoders and rerankers require a joint forward pass per document-label pair. LLMs process all labels in a prompt but use substantially larger models. Macro-F_1 is reported; best results per dataset are in **bold**, and the best dual-encoder results are underlined.

Prompted LLMs. We evaluate `Qwen3-4B` and `Qwen3-8B` (Yang et al., 2025) via multiple-choice prompting. For each document, we prompt the model with the input text followed by an enumerated list of labels expressed as their verbalizers, and assign the model's prediction to the option with the highest next-token probability among the option-letter tokens. This avoids free-form generation and post-hoc parsing, making the LLM's prediction a direct logit comparison over a closed candidate set. Each prediction processes the full prompt (input plus all L verbalized options) through a much larger model than the dual encoders.

Headline comparison. Multiple CLEA-trained dual encoders match or exceed the best alternative regime on every dataset (see Section 5 for the full discussion). The most direct same-backbone comparison is illustrative: the unmodified `Qwen3-Embedding-0.6B` dual encoder underperforms its reranker counterpart `Qwen3-Reranker-0.6B` on all four datasets, but after CLEA it surpasses the reranker on all four, while retaining the cached-inference advantage of dual-encoder deployment.

Qualitative LLM failure modes. Although the prompted LLMs are competitive overall, we observed a recurring qualitative failure mode in their predictions: a tendency to produce very plausible but incorrect classifications grounded in inferred external context. For instance, on AGNews the LLMs occasionally classify a clearly geopolitical World-news article as Business when it briefly mentions the financial implications of an event, despite the article's core topic being international affairs. On EmotionDAIR, the LLMs sometimes assign the emotion they would expect a person to feel given the situation described, rather than the emotion actually expressed in the text. These behaviors are consistent with the well-documented sensitivity of LLM zero-shot classification to label-set framing and demonstration choice (Liu et al., 2023).

G CONTRASTIVE LOSS VARIANTS

Effect of the loss components. Table 6 compares the three objectives introduced in Section 3: the rowwise InfoNCE loss $\mathcal{L}_{\text{rows}}$, the columnwise loss $\mathcal{L}_{\text{cols}}$, and their symmetric combination $\mathcal{L} = \frac{1}{2}\mathcal{L}_{\text{rows}} + \frac{1}{2}\mathcal{L}_{\text{cols}}$. Averaged over the four datasets, the symmetric loss is always at least as good as the best of the two single-sided losses and typically improves macro-F_1 by a small but consistent

margin. Across the ten backbones, $\mathcal{L}$ improves over $\mathcal{L}_{\text{rows}}$ by about $+0.01$ macro-F_1 on average and over $\mathcal{L}_{\text{cols}}$ by roughly $+0.03$ macro-F_1. The gains are particularly visible for the smaller or less specialized models (e.g., BGE-base, Gemma), where combining both directions yields up to $+0.04$ macro-F_1 compared to training only with $\mathcal{L}_{\text{cols}}$.

At the same time, the ablation confirms that $\mathcal{L}_{\text{rows}}$ is the stronger of the two components when used in isolation. For almost all backbones in Table 6, $\mathcal{L}_{\text{rows}}$ matches or slightly outperforms $\mathcal{L}_{\text{cols}}$ when trained alone, indicating that using description vectors as anchors (rows) already captures most of the benefit of CLEA. The columnwise objective by itself is thus not sufficient to reach the best performance, but it becomes beneficial once combined with the rowwise term.

Per-dataset behavior. The disaggregated results in Table 7 show that this pattern holds across tasks. Out of the 40 (model, dataset) combinations, the symmetric loss $\mathcal{L}$ achieves the best or tied-best macro-F_1 in 34 cases. In the remaining 6 cases, either $\mathcal{L}_{\text{rows}}$ or $\mathcal{L}_{\text{cols}}$ is better, but never by more than 0.03 absolute F_1. On coarse-grained topic classification (AGNews) and binary sentiment (RottenTomatoes), the three losses are often very close, with $\mathcal{L}$ typically providing a modest refinement on top of $\mathcal{L}_{\text{rows}}$ (e.g., e5-large, gte-base). The advantage of the symmetric objective becomes more pronounced on the fine-grained tasks Banking77 and EmotionDAIR. For instance, for BGE-large on EmotionDAIR, $\mathcal{L}$ reaches 0.56 macro-F_1, improving over both $\mathcal{L}_{\text{rows}}$ (0.52) and $\mathcal{L}_{\text{cols}}$ (0.53). For Gemma on EmotionDAIR, the symmetric loss lifts performance from 0.45 (rows-only) and 0.54 (cols-only) to 0.57. Similar trends hold for Qwen on EmotionDAIR and for several models on Banking77.

There are a few isolated cases where a single-sided loss slightly dominates the symmetric one (e.g., Qwen on AGNews, Gemma on AGNews and RottenTomatoes, GTE-large on EmotionDAIR and RottenTomatoes), but the margins are small (1-3 F_1 points) and not systematic across backbones or tasks. We therefore view these as noise-level fluctuations rather than evidence against the symmetric formulation.

Takeaways. Overall, the ablation supports our design choice. The rowwise term $\mathcal{L}_{\text{rows}}$ is the primary driver of performance: using description embeddings as anchors already yields strong classifiers and consistently outperforms the columnwise variant when used in isolation. The columnwise term $\mathcal{L}_{\text{cols}}$ plays a complementary role: it is weaker on its own but, when combined with $\mathcal{L}_{\text{rows}}$, acts as a regulariser that better structures the joint description-label space, leading to small but robust gains across most backbones and datasets. Consequently, we adopt the symmetric loss $\mathcal{L}$ as our default objective in all subsequent experiments.

Model	$\mathcal{L}_{\text{rows}}$	$\mathcal{L}_{\text{cols}}$	$\mathcal{L}$
MiniLM			
all-MiniLM-L6-v2	0.70 (0.20)	0.65 (0.18)	**0.70 (0.20)**
E5			
e5-base-v2	0.76 (0.20)	0.76 (0.21)	**0.77 (0.20)**
e5-large-v2	0.78 (0.19)	0.77 (0.19)	**0.79 (0.19)**
BGE			
bge-base-en-v1.5	0.76 (0.20)	0.74 (0.20)	**0.77 (0.21)**
bge-large-en-v1.5	0.78 (0.18)	0.78 (0.18)	**0.80 (0.17)**
GTE			
gte-base-en-v1.5	0.77 (0.19)	0.76 (0.19)	**0.78 (0.20)**
gte-modernbert-base	0.77 (0.19)	0.74 (0.19)	**0.77 (0.19)**
gte-large-en-v1.5	0.77 (0.19)	0.77 (0.19)	**0.78 (0.19)**
Qwen			
Qwen3-Embedding-0.6B	0.80 (0.16)	0.76 (0.18)	**0.80 (0.16)**
Gemma			
embeddinggemma-300m	0.71 (0.19)	0.72 (0.16)	**0.75 (0.15)**

Table 6: Mean macro-F1 (std. in parentheses) for each loss, averaged over the four datasets AGNews, Banking77, EmotionDAIR and RottenTomatoes. Best loss per model (row) is in bold.

Model	AGNews			Banking77			EmotionDAIR			RottenTomatoes		
	$\mathcal{L}_{\text{rows}}$	$\mathcal{L}_{\text{cols}}$	$\mathcal{L}$	$\mathcal{L}_{\text{rows}}$	$\mathcal{L}_{\text{cols}}$	$\mathcal{L}$	$\mathcal{L}_{\text{rows}}$	$\mathcal{L}_{\text{cols}}$	$\mathcal{L}$	$\mathcal{L}_{\text{rows}}$	$\mathcal{L}_{\text{cols}}$	$\mathcal{L}$
MiniLM												
all-MiniLM-L6-v2	0.78	0.71	**0.79**	0.89	0.80	**0.90**	0.43	0.38	**0.43**	0.69	0.70	**0.70**
E5												
e5-base-v2	0.78	0.80	**0.81**	0.95	0.95	**0.96**	0.48	0.45	**0.48**	0.82	**0.83**	0.82
e5-large-v2	0.81	0.80	**0.82**	0.95	0.95	**0.96**	0.52	0.51	**0.53**	0.85	0.84	**0.86**
BGE												
bge-base-en-v1.5	0.81	0.75	**0.82**	0.95	0.94	**0.95**	0.47	0.47	**0.47**	0.81	0.81	**0.82**
bge-large-en-v1.5	0.82	0.80	**0.82**	0.94	0.94	**0.95**	0.52	0.53	**0.56**	0.84	0.84	**0.85**
GTE												
gte-base-en-v1.5	0.82	0.80	**0.83**	0.94	0.92	**0.95**	0.49	0.48	**0.49**	0.83	0.84	**0.85**
gte-modernbert-base	**0.80**	0.71	0.80	0.94	0.92	**0.94**	0.49	0.48	**0.49**	0.83	0.84	**0.84**
gte-large-en-v1.5	0.82	0.79	**0.83**	0.93	0.94	**0.95**	0.50	**0.51**	0.50	0.84	**0.86**	0.83
Qwen												
Qwen3-Embedding-0.6B	**0.86**	0.78	0.85	0.90	0.90	**0.92**	0.56	0.50	**0.57**	0.87	0.87	**0.88**
Gemma												
embeddinggemma-300m	**0.75**	0.71	0.74	0.91	0.92	**0.94**	0.45	0.54	**0.57**	0.73	**0.74**	0.73

Table 7: Disaggregated macro-F_1 by dataset and loss. Entries are F_1 scores; for each model and dataset, the best loss is in bold.

H Robustness to noisy descriptions

To test how sensitive our method is to imperfect descriptions, we run a controlled corruption experiment on AGNews using `embeddinggemma-300m`. Starting from the original description set, we progressively replace a fraction of descriptions with overly vague, non-discriminative sentences (examples in Table 9). We vary the noise level in (0.0, 0.25, 0.5, 1.0), where (1.0) means that *all* descriptions are vague. For each noise level and each loss ($\mathcal{L}_{\text{rows}}$, $\mathcal{L}_{\text{cols}}$, and the symmetric $\mathcal{L}$), we train with the learning rate selected by the uniformity heuristic and track macro-F_1 over training (Figure 6). Table 8 reports the final F_1 scores (taking into account early stopping).

Noise level	$\mathcal{L}_{\text{rows}}$	$\mathcal{L}_{\text{cols}}$	$\mathcal{L}$
0.00	0.75	0.71	0.74
0.25	0.64	**0.71**	0.67
0.50	0.59	**0.64**	0.61
1.00	0.41	**0.45**	0.35

Table 8: Macro-F_1 on AGNews for `embeddinggemma-300m` under different description noise levels and losses.

With clean descriptions ((0.0) noise), the rowwise loss is strongest (0.75 vs. 0.71 for $\mathcal{L}_{\text{cols}}$), and the symmetric loss ($\mathcal{L}$) is only slightly behind at 0.74. Once we inject moderate noise (0.25 and 0.5), the behavior changes markedly. $\mathcal{L}_{\text{rows}}$ degrades sharply (down to 0.64 and 0.59, i.e. -0.11 and -0.16 absolute), whereas $\mathcal{L}_{\text{cols}}$ is much more stable: at 25% noise it remains essentially unchanged compared to the clean setting, and at 50% noise it still clearly dominates the other objectives (0.64 vs. 0.59 and 0.61). The symmetric objective $\mathcal{L}$ interpolates between these regimes: with clean descriptions it tracks $\mathcal{L}_{\text{rows}}$, but under moderate noise it moves closer to $\mathcal{L}_{\text{cols}}$, preserving much of the robustness provided by the column term. At the extreme noise level (1.0), when all descriptions are vague, all three objectives degrade substantially, as expected when the supervision signal is entirely uninformative.

These trends align with the analytical properties of $\mathcal{L}_{\text{cols}}$. By using a log-sum-exp over multiple positives per label, it (i) reweights descriptions so that strong, representative ones dominate the gradient while noisy/vague ones are down-weighted, and (ii) adds extra smoothing / regularization at

the label level. In the moderate-noise regime, arguably the most realistic setting, this makes $\mathcal{L}_{\text{cols}}$ significantly more robust than a pure rowwise InfoNCE. The symmetric loss $\mathcal{L}$ largely inherits this robustness while retaining strong performance in the clean setting, indicating that combining row- and columnwise objectives yields a good trade-off between peak accuracy and resilience to imperfect description sets.

#	Vague AGNews description
1	Coverage about the news, containing information about events and happenings in the world. The text discusses topics that may be of interest to readers who follow current events.
2	A piece of text about something that was reported by journalists or news organizations. It contains sentences and paragraphs describing various matters.
3	This is a news article that provides information to readers. The content covers subjects that are deemed newsworthy by editors and reporters.
4	Information presented in written form about things that have occurred or are occurring. The article is structured with a headline and body text.
5	Some content about a topic that readers might find relevant. The writing style follows journalistic conventions and presents facts or opinions.

Table 9: Examples of deliberately vague AGNews descriptions used to inject noise into the label description set.

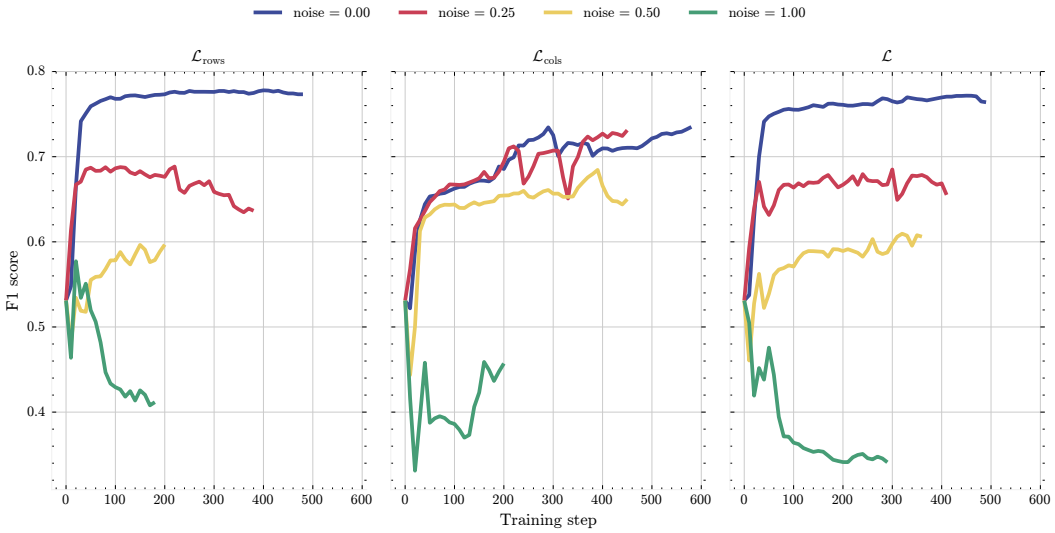


Figure 6: Macro-F_1 on AGNews for `embeddinggemma-300m` under increasing description noise. Each panel shows training curves for one loss ($\mathcal{L}_{\text{rows}}$, $\mathcal{L}_{\text{cols}}$, and the symmetric $\mathcal{L}$); colors denote noise levels (0.0, 0.25, 0.5, 1.0). The columnwise loss remains markedly more stable under moderate noise, and the symmetric loss largely inherits this robustness.

I Robustness to imbalanced description counts

Figure 7 studies whether the proposed method is sensitive to imbalances in the number of descriptions per label. We use `all-MiniLM-L6-v2` on AGNews and compare three setups: (i) the base model without description training, (ii) a balanced configuration with the same number of descriptions per label, and (iii) an imbalanced configuration where *Business* has 10 descriptions while the remaining labels have only 3. For the symmetric objective $\mathcal{L}$ (left panel), adding descriptions improves F_1 across all labels relative to the base model, and giving extra descriptions to *Business* yields a small additional gain for that label while *Science/Tech*, *Sports*, and *World* remain essentially unchanged between the balanced and imbalanced setups. In contrast, under the pure rowwise loss $\mathcal{L}_{\text{rows}}$ (right panel), *Business* again benefits from the extra descriptions, but the other three labels incur a small F_1 drop when moving from the balanced to the imbalanced configuration (on the order of 1-2 points),

although they still clearly outperform the base model. This pattern is consistent with the role of the columnwise term in $\mathcal{L}$. Because $\mathcal{L}_{\text{cols}}$ optimizes a per-label normalized quantity Z_ℓ^+/Z_ℓ, it is inherently less sensitive to how many descriptions a label has: each label's objective is scaled by its own normalizer rather than competing directly with other labels for probability mass. As a result, the symmetric loss $\mathcal{L}$, which includes $\mathcal{L}_{\text{cols}}$, can absorb an over-representation of *Business* without harming the other classes, whereas the purely rowwise objective shows mild cross-label interference when one label receives substantially more descriptions than the rest.

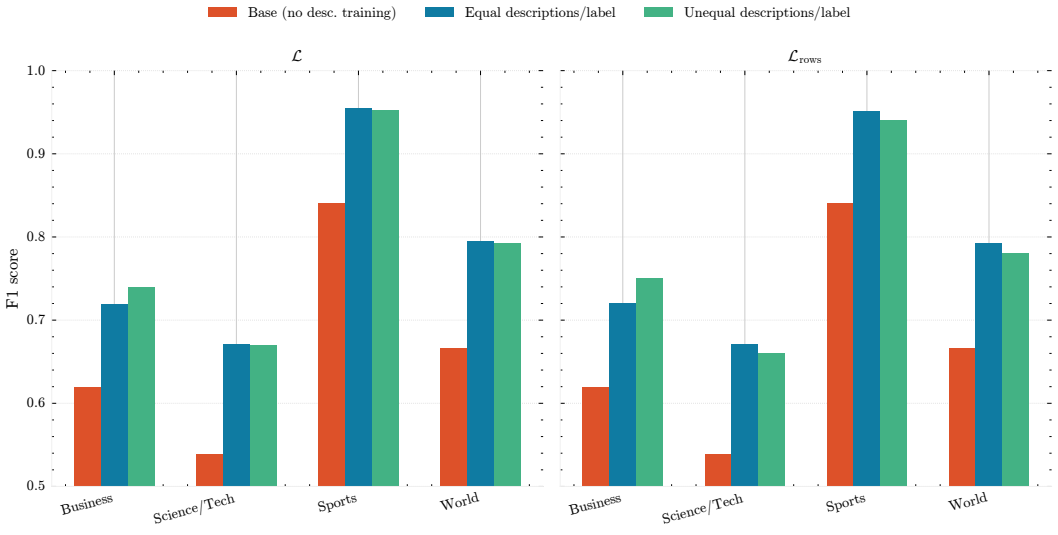


Figure 7: Per-label F_1 on AGNews for `all-MiniLM-L6-v2` under different description allocation schemes. We compare the base model (no description training), a balanced setup with an equal number of descriptions per label, and an imbalanced setup where *Business* has 10 descriptions while the other labels have only 3. Left: symmetric loss $\mathcal{L}$, where extra descriptions for *Business* slightly improve that label without affecting the others. Right: rowwise loss $\mathcal{L}_{\text{rows}}$, where *Business* gains while the remaining labels incur a small F_1 drop but still outperform the base model.

J Effect of the number of descriptions per label

Figure 8 investigates how performance varies with the number of descriptions per label when using the default symmetric loss $\mathcal{L}$. We vary $K \in \{1, 3, 5, 10, 20\}$ while keeping verbalizers fixed; rows correspond to `e5-base-v2` and `e5-large-v2`, columns to the four datasets. To avoid confounding description count with curation order, we sample five random permutations of the description pool for each K, train on the first K descriptions from each permutation, and report the mean and standard deviation of macro-F_1. Across model-dataset pairs, the largest gains usually occur when moving from one description to 3-5 descriptions per label, especially on Banking77 and AGNews. Beyond that point, the returns are dataset-dependent: Banking77 largely saturates after the first few descriptions, RottenTomatoes changes only marginally across the range, while EmotionDAIR and e5-large-v2 on AGNews continue to benefit from larger description sets. The variance bands show that description ordering matters most at small K, but the overall trend is stable under random permutations: a small set of 3-5 informative descriptions captures much of the gain, with additional descriptions useful when the task requires finer semantic coverage.

K Mean-description prototype baseline

Table 10 separates the effect of richer label descriptions from the effect of contrastive alignment. The mean-description baseline embeds each label's descriptions, averages them into a prototype, and classifies without any fine-tuning. CLEA exceeds this prototype in 18 of the 20 (model, dataset) pairs, ties in one (e5-base-v2 on RottenTomatoes), and trails in one (Qwen3-Embedding-0.6B on Banking77, 0.92 vs. 0.93). The gap between the two varies by task. It is smallest on Banking77, where

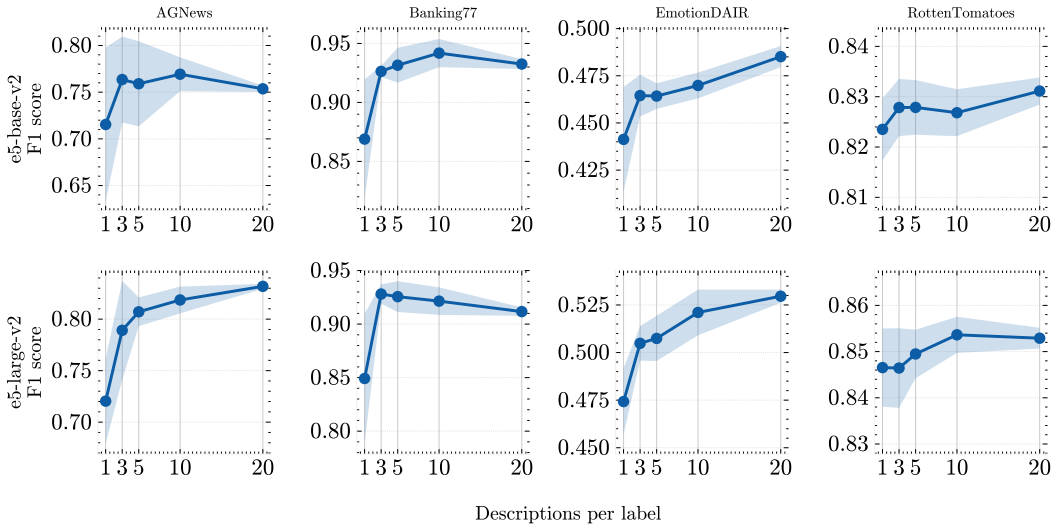


Figure 8: Effect of the number of descriptions per label on macro-F_1 under randomized description ordering. Rows correspond to `e5-base-v2` (top) and `e5-large-v2` (bottom); columns correspond to AGNews, Banking77, EmotionDAIR, and RottenTomatoes. For each K, five random permutations of the description pool are evaluated, selecting the first K descriptions from each permutation. Lines show means across permutations; shaded regions indicate ± 1 standard deviation. Gains are strongest from $K = 1$ to $K = 3$–5, with later improvements depending on the dataset.

Model	Method	AGNews	Banking77	EmotionDAIR	RottenTom.
all-MiniLM-L6-v2	base	0.67	0.66	0.35	0.66
	mean-desc	0.78	0.87	0.39	0.69
	+ *CLEA*	**0.79**	**0.90**	**0.43**	**0.70**
e5-base-v2	base	0.75	0.79	0.44	0.83
	mean-desc	0.78	0.93	0.40	0.82
	+ *CLEA*	**0.81**	**0.96**	**0.48**	0.82
bge-base-en-v1.5	base	0.63	0.86	0.42	0.81
	mean-desc	0.81	0.94	0.39	0.81
	+ *CLEA*	**0.82**	**0.95**	**0.47**	**0.82**
e5-large-v2	base	0.78	0.79	0.44	0.85
	mean-desc	0.69	0.94	0.42	0.84
	+ *CLEA*	**0.82**	**0.96**	**0.53**	**0.86**
Qwen3-Emb.-0.6B	base	0.63	0.88	0.48	0.76
	mean-desc	0.81	**0.93**	0.52	0.82
	+ *CLEA*	**0.85**	0.92	**0.57**	**0.88**

Table 10: Mean-description representations vs. contrastive alignment. *Base*: zero-shot classification using the label verbalizer as the sole label representation. *Mean-desc*: mean-pooled description embeddings as label representations, without fine-tuning. + *CLEA*: contrastive label-embedding alignment. Macro-F_1 is reported; the best result per model-dataset pair is in **bold**.

descriptions of narrowly defined intents already yield good prototypes, and largest on EmotionDAIR, where averaged descriptions do not improve over the verbalizer at all (0.42 vs. 0.43 on average across encoders), while CLEA gains +0.07. The prototype is also less reliable: it falls below the base encoder in 6 of 20 pairs (e.g., e5-large-v2 on AGNews, 0.78 $\rightarrow$ 0.69), whereas CLEA is never more than 0.01 below it. Richer label anchors therefore help, but inconsistently, while contrastive alignment turns the description signal into reliable gains across encoders and tasks, including those where averaging descriptions fails.

L LLM-GENERATED DESCRIPTIONS AND SETFIT COMPARISON

We next evaluate whether CLEA depends on manually written descriptions, and compare this description-only setup against a few-shot SetFit baseline (Tunstall et al., 2022). SetFit combines contrastive fine-tuning with a lightweight classifier head and therefore requires labeled documents. We run SetFit 20 times with different random draws of eight labeled examples per class, and compare it with 20 CLEA runs that use GPT-5.1 (OpenAI, 2025) generated descriptions: for each label, each CLEA run samples five descriptions without replacement from a pool of 20 descriptions and uses no labeled documents.

Figure 9 reports this standalone comparison on AGNews and EmotionDAIR for `all-MiniLM-L6-v2` and `Qwen3-Embedding-0.6B`. LLM-generated descriptions are competitive with SetFit overall and show substantially lower variance, although SetFit has a higher mean in the MiniLM/AGNews setting. Human-written descriptions provide a small but consistent additional gain, visible as the dashed CLEA reference lines. This suggests that manual curation is beneficial but not essential: LLM-generated descriptions can substantially reduce the authoring burden while preserving the label-description supervision signal. In our experience, writing five human descriptions per label took roughly 45–90 minutes, conditional on prior familiarity with the task.

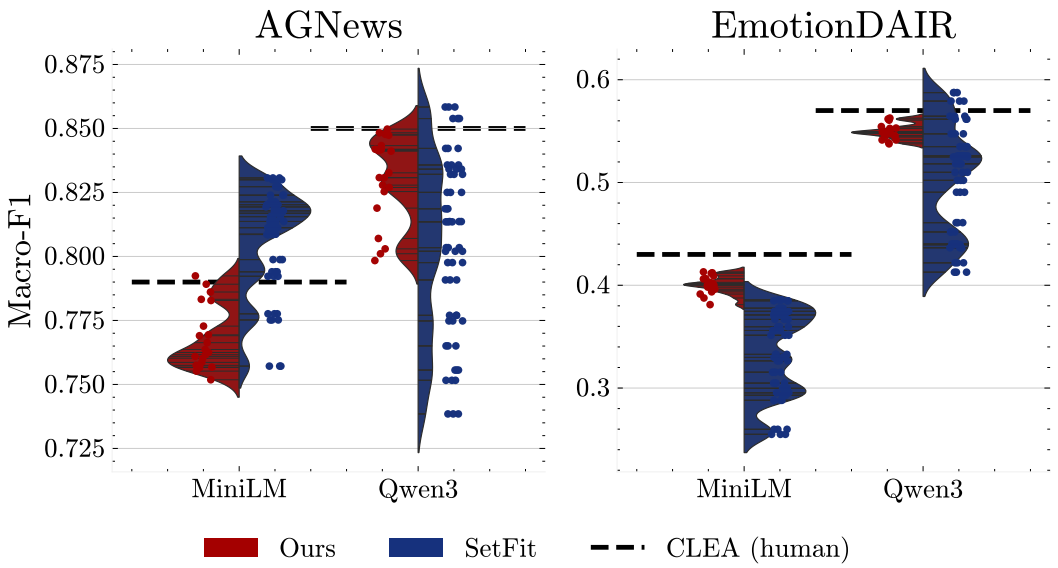


Figure 9: Comparison between CLEA with GPT-5.1-generated descriptions and SetFit (Tunstall et al., 2022). CLEA uses no labeled documents; each run samples five descriptions per label from a pool of 20 generated descriptions. SetFit uses eight labeled examples per class. Dashed lines show CLEA with human-written descriptions.

M CHOICE OF INFERENCE ANCHOR

We also ablate the choice of label representation ("anchor") used during training and inference. We compare three modes:

1. **Label.** We replace the verbalizer sentence by the raw label string (e.g., "Sports") and use this as the anchor both during training and at inference. 2. **Verbalizer.** Our default setting: we train and evaluate with the full verbalizer sentence (e.g., "This news snippet is about sports."). 3. **Mean.** We train as in the default setting (with verbalizers), but at inference we instead score documents against the mean embedding of each label's descriptions, $\frac{1}{K_y}\sum_k e(d_y^k)$.

Figure 10 reports macro-F_1 for `e5-base-v2` and `e5-large-v2` across the four datasets. No single mode dominates everywhere. For AGNews, the verbalizer performs best. For Banking77 and

EmotionDAIR, the raw label string tends to outperform the verbalizer. For RottenTomatoes, the mean of descriptions is strongest. Across all model-dataset pairs, however, the differences are small (typically within 1-2 F_1 points), and we do not observe statistically significant differences between the three modes. This indicates that the method is not fragile to the precise choice of anchor. Taken together, these results suggest that (i) training directly on label names is a viable alternative to full verbalizers, and (ii) even when training with verbalizers, practitioners can safely switch to mean description embeddings at inference time, which typically match the performance of the default.

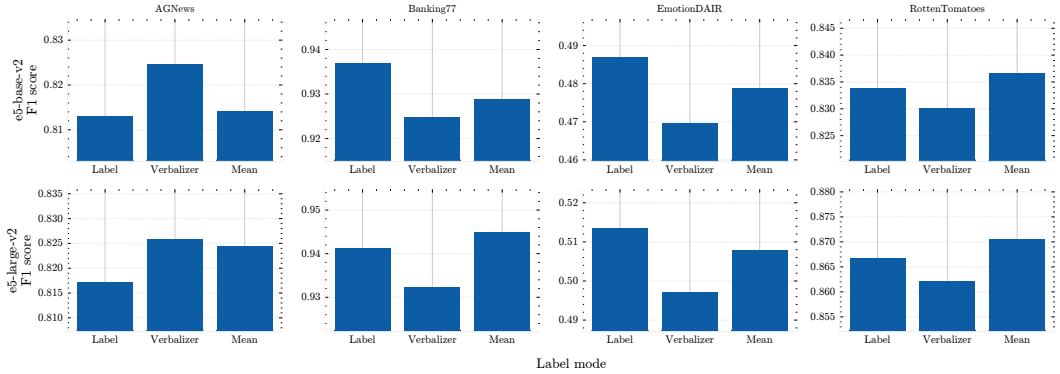


Figure 10: Ablation over label representation. Bars show macro-F_1 when training and evaluating on the raw label string (*Label*), training and evaluating on the full verbalizer sentence (*Verbalizer*, our default), or training on verbalizers but using the mean description embedding at inference (*Mean*). Rows correspond to `e5-base-v2` (top) and `e5-large-v2` (bottom); columns correspond to AGNews, Banking77, EmotionDAIR, and RottenTomatoes. The best choice varies mildly by dataset, but the small differences are not statistically significant, indicating robustness to the particular inference anchor.

N Performance on large label spaces

We additionally test whether CLEA remains effective when the label space is large. We use the full 77-way *Banking77* label set, writing only three descriptions per label, and evaluate on the official test split. As in the main experiments, learning rates are selected with the label-free uniformity criterion from Section 3.

Table 11 shows that CLEA yields consistent gains across four encoders under this harder setting. For `all-MiniLM-L6-v2`, macro-F_1 and accuracy both increase by $+0.07$ (0.59→0.66), while precision improves from 0.65 to 0.69 and recall from 0.62 to 0.69. The effect is strongest for `embeddinggemma-300m`, where macro-F_1 rises by $+0.09$ (0.52→0.61), accuracy by $+0.08$, precision by $+0.03$, and recall by $+0.10$. The remaining encoders show the same pattern: `bge-base-en-v1.5` gains $+0.03$ macro-F_1 and $+0.03$ recall, while `e5-large-v2` gains $+0.07$ macro-F_1 and $+0.10$ recall with unchanged precision. Thus, the gains are driven primarily by better recall, indicating that CLEA reduces false negatives in the larger label space without simply over-predicting more labels.

Overall, these results indicate that CLEA scales to high-cardinality label spaces even with a very small description budget. The gains in macro-F_1 and recall suggest that aligning labels to compact descriptions helps the model cover many fine-grained intents rather than merely sharpening predictions for the most frequent ones.

O Learning-rate selection via uniformity

Uniformity-based learning rate selection. As discussed in Section 3, we treat the learning rate (LR) as the main sensitive hyperparameter and select it using a *label-free* criterion based on the uniformity loss $\mathcal{L}_{\text{uni}}$ in equation 4. For each model-dataset pair, we sweep over a small grid of candidate LRs during training and evaluate $\mathcal{L}_{\text{uni}}$ on an unlabeled pool $\mathcal{X}_u$ from the target domain. We

Model	F_1	Acc.	Prec.	Rec.
all-MiniLM-L6-v2	0.59	0.59	0.65	0.62
+ *CLEA*	0.66 (+0.07)	0.66 (+0.07)	0.69 (+0.04)	0.69 (+0.07)
embeddinggemma-300m	0.52	0.52	0.60	0.55
+ *CLEA*	0.61 (+0.09)	0.60 (+0.08)	0.64 (+0.03)	0.64 (+0.10)
bge-base-en-v1.5	0.64	0.65	0.69	0.68
+ *CLEA*	0.67 (+0.03)	0.67 (+0.02)	0.70 (+0.01)	0.71 (+0.03)
e5-large-v2	0.57	0.55	0.69	0.59
+ *CLEA*	0.65 (+0.07)	0.65 (+0.10)	0.69 (+0.00)	0.68 (+0.10)

Table 11: Performance on the full 77-class Banking77 label space with three descriptions per label. + *CLEA* denotes contrastive label-embedding alignment. Absolute improvements over the base model are shown in parentheses.

then choose the LR that minimizes $\mathcal{L}_{\text{uni}}$, i.e., that yields the most uniform (least collapsed) document embeddings on $\mathcal{X}_u$. Table 12 lists the selected learning rates. These are the hyperparameters used for all results reported in the main table (Table 1).

Model	AGNews	Banking77	EmotionDAIR	RottenTomatoes
all-MiniLM-L6-v2	$1 \cdot 10^{-5}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$3 \cdot 10^{-5}$
e5-base-v2	$1 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$1 \cdot 10^{-4}$
e5-large-v2	$1 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$5 \cdot 10^{-5}$
bge-base-en-v1.5	$5 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$3 \cdot 10^{-5}$
bge-large-en-v1.5	$1 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$5 \cdot 10^{-6}$
gte-base-en-v1.5	$1 \cdot 10^{-4}$	$1 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$1 \cdot 10^{-5}$
gte-modernbert-base	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$
gte-large-en-v1.5	$3 \cdot 10^{-4}$	$5 \cdot 10^{-5}$	$3 \cdot 10^{-4}$	$1 \cdot 10^{-4}$
Qwen3-Embedding-0.6B	$3 \cdot 10^{-5}$	$1 \cdot 10^{-5}$	$3 \cdot 10^{-5}$	$1 \cdot 10^{-5}$
embeddinggemma-300m	$1 \cdot 10^{-4}$	$3 \cdot 10^{-5}$	$5 \cdot 10^{-5}$	$5 \cdot 10^{-5}$

Table 12: Learning rates selected by minimizing the uniformity loss ($\mathcal{L}_{\text{uni}}$) on an unlabeled pool ($\mathcal{X}_u$) for each model–dataset pair. The LRs are used in the main results (Table 1).

Relationship between uniformity and downstream performance. Appendix Figure 11 examines how $\mathcal{L}_{\text{uni}}$, measured on $\mathcal{X}_u$, relates to downstream macro-F_1 after description-only training. Our goal is not to treat $\mathcal{L}_{\text{uni}}$ as a perfect surrogate for macro-F_1, but as a practical label-free *heuristic* for LR selection. Across the 40 model-dataset pairs in Figure 11, **29/40** exhibit a statistically significant *negative* Pearson correlation between $\mathcal{L}_{\text{uni}}$ and macro-F_1 ($p \leq 0.10$): runs that yield more uniform embeddings on $\mathcal{X}_u$ (lower $\mathcal{L}_{\text{uni}}$) tend to achieve higher macro-F_1. The remaining **11/40** pairs show weak, non-significant correlations, which are typically slightly negative or close to zero but never strongly positive. In other words, lower uniformity does not guarantee higher macro-F_1, yet it is empirically helpful in the majority of cases and, critically, does not appear to systematically harm performance in the remainder. This supports using $\mathcal{L}_{\text{uni}}$ as a robust, label-free signal for picking a reasonable learning rate, while acknowledging that it is empirically useful but not universally predictive.

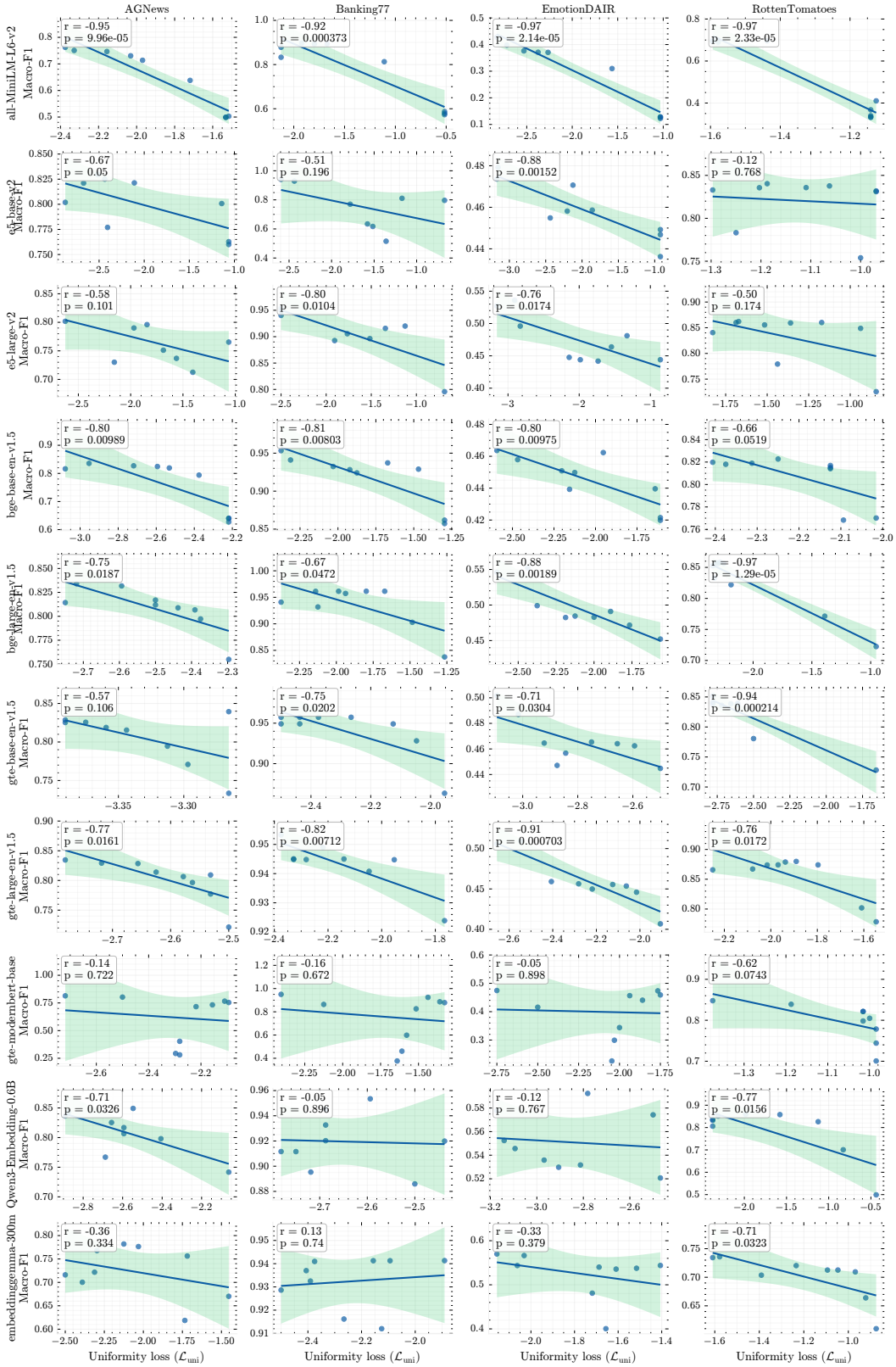


Figure 11: Scatter plots of uniformity loss $\mathcal{L}_{\text{uni}}$ against Macro-F1 performance across datasets. Rows correspond to embedding models, while columns correspond to datasets. Each subplot shows individual runs with a different learning rate (dots), an ordinary least squares regression line with 95% confidence interval (shaded), and the Pearson correlation coefficient between $\mathcal{L}_{\text{uni}}$ and Macro-F1.