ATEB: Rethinking Advanced NLP Tasks in an Information Retrieval Setting

Simeng Han[1, 2], Frank Palma Gomez[2], Tu Vu[2], Zefei Li[2], Daniel Cer[2], Hansi Zeng[3], Chris Tar[2], Arman Cohan[1], Gustavo Hernandez Abrego[2]
[1]Yale University, [2]Google Deepmind [3]University of Massachusetts Amherst

Abstract

Traditional text embedding benchmarks primarily evaluate embedding models' capabilities to capture semantic similarity. However, more advanced NLP tasks require a deeper understanding of text, such as safety and factuality. These tasks demand an ability to comprehend and process complex information, often involving the handling of sensitive content, or the verification of factual statements against reliable sources. We introduce a new benchmark designed to assess and highlight the limitations of embedding models trained on existing information retrieval data mixtures on advanced capabilities, which include factuality, safety, instruction following, reasoning and document-level understanding. This benchmark includes a diverse set of tasks that simulate real-world scenarios where these capabilities are critical and leads to identification of the gaps of the currently advanced embedding models. Furthermore, we propose a novel method that reformulates these various tasks as retrieval tasks. By framing tasks like safety or factuality classification as retrieval problems, we leverage the strengths of retrieval models in capturing semantic relationships while also pushing them to develop a deeper understanding of context and content. Using this approach with single-task fine-tuning, we achieved performance gains of 8% on factuality classification and 13% on safety classification. Our code and data will be publicly available.

1 Introduction

Traditional retrieval models are primarily trained and evaluated on traditional Information Retrieval tasks including document retrieval, reranking and sentence similarity (Muennighoff et al., 2023a). However, this approach falls short when applied to more advanced natural language capabilities that require a deeper understanding of text, such as reasoning, factuality, instruction-following and long-form text understanding (Su et al., 2024; Xiao et al., 2024; Weller et al., 2024). These tasks demand an ability to comprehend and process complex information, often involving multiple steps of reasoning, the handling of sensitive content, or the verification of factual statements against reliable sources (Vu et al., 2024; Ji et al., 2023; Su et al., 2024).

Recent benchmarks have been proposed to evaluate reasoning-intensive retrieval (Su et al., 2024; Xiao et al., 2024) or instruction following (Weller et al., 2024). However, these benchmarks only consider a single advanced capability. In particular, Weller et al. (2024) introduced a new benchmark built on top of traditional retrieval benchmarks to measure instruction-following capability of embedding models, however, they do not capture how well embedding models perform on the most widely adopted instruction following benchmarks such as Stanford Human Preference (SHP) (Ethayarajh et al., 2022a) for ranking human preference over model outputs and HH-RLHF (Bai et al., 2022) for ranking helpfulness and safety of model responses.

We introduce a new benchmark designed to assess embedding models on advanced LLM capabilities by reformulating existing datasets from diverse categories into information retrieval tasks. These include **safety classification**: *BeaverTails Safety Classification* (Ji et al., 2023), *HH-RLHF Harmlessness Classification* (Bai et al., 2022); **factuality classification**: *ESNLI* (Camburu et al., 2018), *DialFact* (Gupta et al., 2022), *VitaminC* (Schuster et al., 2021), **instruction following reranking**: *Stanford Human Preference* (Ethayarajh et al., 2022b), *AlpacaFarm* (Dubois et al., 2023), *LMSys* (Chiang et al., 2024), *Genie* (Khashabi et al., 2022), *InstrSum* (Ji et al., 2023), *HH-RLHF Helpful* (Bai et al., 2022); **document-level pairwise-classification**: *DIPPER* (Krishna et al., 2023), and **document-level bitext-mining**: *Europarl* (Koehn, 2005), *IWSLT17* (Cettolo et al., 2017), *NC2016* (Maruf et al., 2019)). We also incorporate sub-

sets of **reasoning retrieval** (Xiao et al., 2024), We evaluate our benchmark with advanced embedding models: a Gemma-2B (Team et al., 2024) embedding model trained as a symmetric dual encoder (Neelakantan et al., 2022) and Google Gecko embedding model (Lee et al., 2024b).

We adopt a novel fine-tuning approach that reformulates various classification tasks into a retrieval-based setting using contrastive loss. In this setup, each task instance is represented as a triplet: an input (query and answer concatenated), a positive target (label text plus explanation and a unique ID), and multiple negative targets (alternative labels with explanations and the same unique ID) (Lee et al., 2024b). This approach allows dual encoder embedding models to handle classification tasks without any architectural modifications and enables seamless integration with other retrieval and similarity-based training objectives. We further adapt this approach by adding a detailed explanation for the label text to reduce the influence of the long series of tokens in the unique ID the model and maintain its focus on learning the semantics of the input and targets. We have obtained 8% and 13% improvements on factuality classification and safety classification tasks respectively with this approach with single-task fine-tuning. We also provide a more lightweight training approach: adopting an adapter over a genereic Gemma embedding, which leads to 2% and 3% improvements on the same factuality and safety classification tasks.

In summary, the contributions of our paper are threefold. 1) We introduce a new benchmark, ATEB, designed to evaluate text embedding models on advanced NLP tasks such as reasoning, safety, factuality, and instruction-following. Unlike traditional benchmarks focused solely on text similarity and retrieval, ATEB encompasses diverse real-world scenarios requiring deeper contextual understanding and reasoning, highlighting the limitations of advanced embedding models. 2). We propose a novel fine-tuning approach that reformulates various classification and reasoning tasks into retrieval-based problems, enhancing the ability of dual encoder models to handle advanced capabilities without architectural modifications. Our method achieves significant improvements, with 8% and 16% gains in factuality and safety classification tasks, respectively. 3). Additionally, we demonstrate the utility of adapter-based fine-tuning for achieving competitive results with minimal computational cost.

2 Related Work

Text Embedding Models Representing text as dense vectors with neural networks gained prominence through word2vec (Mikolov et al., 2013), which generated semantically meaningful word embeddings. Subsequently, models like BERT (Devlin et al., 2019) and the contrastively trained SimCSE (Gao et al., 2021) solidified encoder-only transformers as the predominant architecture for producing text embeddings. More recently, decoder-only transformers have advanced significantly in both capability and efficiency (Brown et al., 2020; Team et al., 2024; Dubey et al., 2024; Jiang and Chen, 2023), making it logical to utilize their pretrained knowledge for embedding tasks. This approach was successfully demonstrated by (Neelakantan et al., 2022), who initiated embedder training from decoder-only GPT models and has been adopted by recent leading open-source models on the MTEB leaderboard (BehnamGhader et al., 2024; Lee et al., 2024a; Meng et al., 2024).

Text Embedding Evaluation Because embedding models are applied in diverse scenarios, there is a need for broad and heterogeneous benchmarks to thoroughly evaluate their performance. The pioneering effort in this domain was BEIR (Thakur et al., 2021), which comprises nine distinct information retrieval tasks—such as duplicate-question retrieval and citation prediction—across 18 datasets. More recently, Muennighoff et al. (2023) introduced MTEB (Massive Text Embedding Benchmark) (Muennighoff et al., 2023b), an extensive evaluation framework that surpasses BEIR in scale and includes more diverse task categories such as classification and reranking.

2.1 Advanced Model Capabilities

Recent advances in natural language processing have seen the emergence of a variety of specialized tasks aimed at evaluating model safety (Bai et al., 2022), factuality (Dziri et al., 2022a), reasoning, instruction-following (Ethayarajh et al., 2022b), and document-level understanding, which are crucial capabilities for the most recent foundation models (Reid et al., 2024; OpenAI, 2024; Dubey et al., 2024). Safety tasks focus on mitigating harmful, biased, or unethical outputs, ensuring models uphold socially responsible standards (Bai et al., 2022). Factuality tasks emphasize grounding responses in reliable information and reducing fabrication or misinformation, as exemplified by

research efforts on factual consistency in summarization and truthful QA (Maynez et al., 2020; Lin et al., 2022). Reasoning-oriented challenges push models beyond surface-level pattern recognition by encouraging deeper inference and logical deduction (Xiao et al., 2024). Instruction-following tasks further refine models' ability to adhere to user directives (Ouyang et al., 2022). In parallel, document-level understanding (Yin et al., 2021; Krishna et al., 2023) tests models' capabilities to process long-form texts beyond sentences.

3 ATEB Construction

3.1 Design Principles

The benchmark comprises 21 tasks, encompassing datasets related to instruction-following, factuality, reasoning, document-level translation, and paraphrasing. These tasks simulate real-world scenarios requiring advanced model capabilities. We reformulate these tasks from existing sources based on the following principles. **Factuality as classification**: NLI tasks where the goal is to classify the relationships of the premise and hypothesis into *entailment*, *contradiction*, or *neutral*; **Instruction following as reranking**: Ranking model-generated responses based on human preference; **Safety as classification**: Binary classification tasks or ranking tasks (safe vs. unsafe); **Reasoning as retrieval**: Retrieving the gold answer from gold answer pool of all examples in the dataset based on the question; **Document-level paraphrasing as pairwise-classification**: Pairing the paraphrase of a document with the original document based on paraphrases of all documents in the dataset; **Document-level machine translation (MT) as bitext-mining**: identifying the translation of a document over translations of all documents in the dataset.

We provide detailed illustrations of how each task category is constructed, accompanied by examples. For each task, we utilize the complete test set from the corresponding public datasets.

3.2 Factuality as Classification

We adopt several Natural Language Inference (NLI) classification datasets in our factuality classification collection. This includes ESNLI (Camburu et al., 2018), VitaminC (Schuster et al., 2021) and DialFact (Gupta et al., 2022). An example of the ESNLI dataset is shown in Table 1 where the input consists of a concatenation of one premise and one hypothesis and the target is one of the strings of the three classes including "entailment", "contradictory" and "neutral".

3.3 Instruction-Following as Reranking

We reformulate publicly available instruction-following tasks into reranking tasks where the rank is determined by the human preference. Between two model outputs, the model output preferred by human is ranked higher than the model output less preferred. The query is formulated as the concatenation of the task instruction and input context. We provide an example of one of the source datasets we adopted, Stanford Human Preference (Ethayarajh et al., 2022b), in Table 2 and the reformulated example based on it in Table 3. We reformulate six more instruction-following tasks into reranking tasks, which include AlpacaFarm (Dubois et al., 2023), HHRLHF-Helpful (Bai et al., 2022), BeaverTails-Helpful (Ji et al., 2023), Genie (Khashabi et al., 2022), LMSys ChatBot Arena (Chiang et al., 2024), InstruSum (Liu et al., 2024).

3.4 Safety as Classification

We adopt the safety classification portion of the BeaverTails dataset for LLM safety alignment (Ji et al., 2023), BeaverTails QA-Classification to construct a safety classification task for evaluating embedding models where the goal of the task is to classify the input into *safe* or *unsafe*. An example of the BeaverTails QA-Classification dataset is shown in Table 4. We adopt the harmlessness evaluation portion in the HH-RLHF dataset to construct a safety safety classification task.

3.5 Reasoning as Retrieval

We adopt 5 subsets of the RAR-b dataset proposed in (Xiao et al., 2024) including HellaSwag NLI dataset (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2019), PIQA (Bisk et al., 2019), AlphaNLI (Bhagavatula et al., 2019) and ARCChallenge (Clark et al., 2018). Table 5 shows the data format of the reformulated datasets.

3.6 Document-Level Paraphrasing as Pairwise-Classification

We reformulate one document-level paraphrasing dataset, DIPPER (Krishna et al., 2023) as a pairwise classfication task. These tasks expand over previous sentence-level paraphrasing tasks used for pairwise classification (Muennighoff et al., 2023b) to test the document-level modeling capabilities of most advanced embedding models.

Input: *Premise*: Everyone really likes the newest benefits. *Hypothesis*: The new rights are nice enough.
Target: entailment, contradictory, or neutral.

Table 1: An example of ESNLI.

Original SHP
responseA: "It doesn't sound like they deserve the courtesy of two weeks notice. Check company policy and state law about whether they have to pay your sick time or other PTO...
responseB: "...I'd say you are within your rights to kick over the can of kerosene and toss the Zippo..."
preference label: "responseA"
task instruction: "In this task, you will be provided with a context passage (often containing a question), along with two long-form responses to it (responseA and responseB). The goal is to determine which of the two is a better response for the context..." **input:** "How unprofessional would it be to quit the moment I have a job lined up following my vacation? I hate my coworkers..."

Table 2: Original Stanford Human Preference (SHP) dataset example.

Query: "In this task, you will be provided with a context passage (often containing a question), along with two long-form responses to it (responseA and responseB). The goal is to determine which of the two is a better response for the context...How unprofessional would it be to quit the moment I have a job lined up following my vacation? I hate my coworkers...
Positive: "It doesn't sound like they deserve the courtesy of two weeks notice. Check company policy and state law about whether they have to pay your sick time or other PTO...
Negative: "...I'd say you are within your rights to kick over the can of kerosene and toss the Zippo..."

Table 3: Reformulated example of our SHP-Reranking for evaluating embedding models' reranking capability for model responses given instructions.

Query: Input: Is there sites that sell weed? Yes, there are websites that sell marijuana. These sites usually require customers to have a valid prescription and may require a background check to ensure that the individual is of legal age to purchase the drug. However, there are also websites that do not require any forms of identification...
Two classes: safe/unsafe

Table 4: An example of the BeaverTails QA-Classification dataset for evaluating embedding models.

Input: a query in the dataset. **Target**: the answer to the query. **Negative targets**: all the other answers in the dataset.

Table 5: Data format of the reasoning as retrieval datasets for evaluating embedding models.

3.7 Document-Level MT as Pairwise-Classification

Following the same design principle of our new pairwise-classification tasks, we reformulate three document-level machine translation datasets as bi-text mining tasks, which include Europarl (Koehn, 2005), IWSLT17 (Cettolo et al., 2017) and NC2016 (Maruf et al., 2019). These tasks expand over previous sentence-level machine translation tasks used for bi-text mining (Muennighoff et al., 2023b) to test the document-level modeling capabilities of most advanced embedding models. We adopt the subset of these datasets used in Maruf et al. (2019).

4 Method

4.1 Model

We begin by initializing a symmetric dual encoder (DE) using the decoder-only Gemma-2B model (Team et al., 2024; Palma Gomez et al., 2024), which has an embedding size of 2048. Following this, we add a linear projection layer, applied after pooling the outputs along the sequence length dimension. Both the embedding layer and the linear projection layer are randomly initialized. After the model is initialized with Gemma-2B, we train it using a contrastive loss (Hadsell et al., 2006).

4.2 Training data reformulation with label augmentation

While using the dot-product scores along the diagonal as positives and everything else as negatives works well for retrieval and similarity/relatedness matching tasks, it can not be used directly for tasks with targets that are classification labels. Naively providing tasks with classification labels to a dual encoder embedding models will result in the score for an input's correct label appearing both along the diagonal and the off-the diagonal when another input example has the same target label.

Therefore, we adopt a novel method that reformulates various tasks as retrieval tasks during the fine-tuning process, following previous work in fine-tuning with a retrieval setting using contrastive loss (Lee et al., 2024b; Meng et al., 2024). **Input**: query and answer concatenated together. **Positive target**: [*label text* (e.g., neutral for NLI).] + [*label text explanation*] + [*unique id*]. **Negative targets**: [each other possible types of *label texts*.] + [*label text explanation*] + [*unique id*].

Including a unique id for for each correct input/-target pair alone would allow the model to exploit and rely on the unique identifiers to always pair the correct input with the correct target. However, this can be addressed by including additional incorrect labels for each input as negatives. The negatives are tagged with the same unique id as the input and the correct target label. This allows the unique ids to be used to identify candidate targets for each input, but without revealing which of the targets is correct. The advantage of this approach is that it allows dual encoder based embedding models to be trained on classification tasks without any modeling changes. In practice, a unique ID often consists of a long sequence of tokens that can inadvertently shift the model's focus away from learning the semantic relationships within the input and target texts of an example. To address this challenge, we enhance this approach by providing detailed explanations for each label. This additional context helps the model grasp the conceptual meaning behind labels rather than becoming distracted by the long series of unique ID tokens.

For example, for the label "Entailment," we augment it by including the following label explanation:

Label explanation for "entailment"

"In the context of Natural Language Inference (NLI), 'entailment' refers to a specific type of relationship between two sentences, where the truth of one sentence (the hypothesis) is logically guaranteed by the truth of another sentence (the premise)."

By augmenting labels with such detailed explanations, we guide the model toward a richer, more coherent understanding of the underlying concepts it needs to learn.

5 Testing Advanced Embedding Models on ATEB

We test advanced embedding models on ATEB and show their strengths and limitations on our proposed ATEB tasks.

5.1 Baseline methods

Our baseline methods include two advanced embedding models: our Gemma-2B symmmetric dual encoder trained with a prefinetuning stage and Google's gecko embedding model (Lee et al., 2024b), which has a 1-billion parameter size. Both of these baseline models are highly capable embedding models. Notably, the Google Gecko model is a state-of-the-art embedding model with 768 dimensions. On the Massive Text Embedding Benchmark (MTEB), it achieves an average score of 66.31—on par with models that are seven times larger and have five times higher dimensional embeddings on the MTEB leaderboard. The models that achieve a score of 66 or higher, such as NV-Embed-v2and SFR-Embedding, all have 4096 or 8192 dimensions. The prefinetuning stage for Gemma-2B is full supervision finetuning with Huggingface Sentence Transformer datasets. [1]. The baseline models are large-size retrieval models trained for generic information retrieval tasks, and they are not fine-tuned on task-specific data. We include detailed hyperparameters in the Appendix.

5.2 Experimental Results

Baseline Models have Close-to-Random Performance on New Reranking Tasks Table 6 compares the baseline performance of the model against a random chance baseline (75%) on various

[1] https://huggingface.co/sentence-transformers

Reranking task	Random (%)	Gemma-2B (%)	Gecko (%)
AlpacaFarm	75	75.1	75.3
Genie	75	75.3	75.0
InstruSum	75	72.8	74.1
Stanford SHP	75	80.47	77.1
BeaverTails Helpful	75	74.51	75.9
HH RLHF Helpful	75	77.74	77.1
LMSys Chatbot Arena (English)	75	73.18	72.9

Table 6: Baseline performance on reranking for evaluating instruction-following.

reranking tasks designed to evaluate its instruction-following capabilities. These tasks involve ranking model-generated responses based on relevance or helpfulness. On AlpacaFarm and Genie, the baseline models' performance hover between 75.0% and 75.3%, which is marginally higher than random, indicating only limited improvement. In contrast, on InstruSum, the baseline models achieve 72.7% and 74.1, slightly below random chance, underscoring the difficulties in effectively ranking summaries based on human-written instructions. On Stanford SHP, the model performs notably better, achieving 80.47% accuracy with the Gemma-2B embedding model and demonstrating a moderate ability to rank responses according to human preferences. However, on BeaverTails Helpful, the models' accuracy of 74.51% and 75.9% remain close to random, suggesting challenges in identifying genuinely helpful responses. The HH RLHF Helpful task sees some improvement, with the model reaching 77.74%, indicating a modest enhancement in tasks informed by human reinforcement learning preferences. Finally, in the LMSys Chatbot Arena (English) setting, the model attains 73.18%, which is below random chance, thus reflecting limited success in ranking chatbot-generated responses. Taken together, these results highlight the baseline model's near-random performance on most reranking tasks, with only modest improvements in a few cases such as Stanford SHP and HH-RLHF Helpful.

They suggest that further optimization and more task-specific fine-tuning are needed to enhance the model's instruction-following capabilities in these reranking scenarios.

Baseline Models Perform Suboptimally on New Retrieval Tasks Table 7 presents the performance of baseline models compared to random chance in reasoning-based retrieval tasks. These tasks require models to identify correct answers or make logical inferences, highlighting their reasoning capabilities. Key observations include:

Retrieval task	Random (%)	Gemma-2B (%)	Gecko (%)
HellaSwag	0	22.1	26.7
Winogrande	0	17.3	21.2
PIQA	0	22.2	29.8
AlphaNLI	0	30.3	32.1
ARCChallenge	0	7.62	10.9

Table 7: Results of retrieval tasks for evaluating reasoning.

Classification task	Random (%)	Gemma-2B (%)	Gecko (%)
ESNLI	33.3	35	36.1
DialFact	33.3	33.8	33.2
VitaminC	33.3	37	35.4
HH-RLHF Harmlessness	50	50	50
BeaverTails Classify	50	55.9	54.7

Table 8: Results of classification tasks for evaluating factuality and safety.

On HellaSwag, the baseline embedding models achieve 22.1% and 26.7% accuracy, demonstrating moderate success in selecting plausible continuations for narrative reasoning tasks. With 17.3% and 21.2% accuracy on Winogrande, the model struggles in resolving pronoun references, indicating challenges in understanding nuanced context. Achieving 22.2% accuracy on PIQA, the baseline shows limited capability in physical commonsense reasoning. The model performs better in the abductive commonsense reasoning task AlphaNLI, achieving 30.3% and 32.1% accuracy, suggesting it can partially infer plausible explanations for events. On ARCChallenge, with only 7.62% and 10.9% accuracy, the models exhibit significant difficulty in answering challenging science questions, reflecting its limited knowledge retrieval and reasoning skills. In summary, baseline models demonstrate suboptimal performance across these reasoning-based retrieval tasks, with accuracies ranging from 7.62% to 32.1%. This underscores the need for targeted fine-tuning and task-specific training to improve reasoning capabilities in advanced embedding models.

Baseline Models have Close-to-Random Performance on New Classification Tasks Table 8 illustrates the performance of two baseline models, Gemma-2B and Gecko, on five classification tasks—ESNLI, DialFact, VitaminC, HH-RLHF Harmlessness, and BeaverTails Classify—compared to random chance accuracy. For ESNLI, which evaluates natural language inference, both models perform only slightly above random (35% for Gemma-2B and 36.1% for Gecko) despite

Pairwise classification	Random (%)	Gemma-2B (%)	Gecko (%)
Dipper	50	73.1 %	80.1
Bi-text mining			
Europarl	1/n	86.1%	88.2%
IWSLT17	1/n	86.4%	87.1
NC2016	1/n	98%	99 %

Table 9: Baseline Accuracy for pairwise classification and bi-text mining tasks

random performance being 33.3%, indicating limited reasoning capability. Similarly, on DialFact, which assesses factual consistency in dialogue, the models perform very close to random, with Gemma-2B achieving 33.8% and Gecko 33.2%. In the VitaminC task, focused on fact verification, both models show modest improvement over random (33.3%), with Gemma-2B reaching 37% and Gecko slightly lower at 35.4%. For the HH-RLHF Harmlessness task, which classifies whether responses are harmless, both models achieve exactly 50%, matching random performance and indicating no learned capability. Finally, on BeaverTails Classify, a binary classification task where random accuracy is 50%, the models perform slightly better, with Gemma-2B at 55.9% and Gecko at 54.7%, reflecting some potential but still falling short of reliable generalization. These results collectively highlight the close-to-random performance of baseline models on novel classification tasks, underscoring the need for more advanced methods to achieve meaningful improvements in generalization and reasoning.

Baseline Models Perform Reasonably Well on New Pairwise Classification Tasks Table 9 compares the baseline accuracy of a model against random predictions across pairwise classification tasks. The results highlight the baseline model's effectiveness in these specific contexts:

On Dipper, the baseline model achieves an accuracy of 73.06%, significantly outperforming the random baseline of 50%, showcasing strong performance in pairwise classification tasks.

Baseline Models Perform Very Well on New Bitext-Mininig Tasks Bi-text mining Tasks involve identifying semantically equivalent text pairs across multilingual datasets. On each of the three datasets consisting of a few hundred of document-translation pairs, both Gemma-2B model and Gecko model perform very well, excelling particularly in NC2016 with a high accuracy of 98%, indicating exceptional capability in identifying translations of text correspondences.

	ESNLI(%)	DialFact(%)
Random	33	33
Without label augmentation		
Full-supervision with MNLI	34.0	33.1
With label augmentation		
Full-supervision with MNLI (w/o label exp.)	35.0	33.2
Full-supervision with MNLI	**42.0**	**35.8**
Full-supervision with FaithDial data	36.87	34.95
Full-supervision over pre-finetuned with MNLI	37.61	33.5
Adapter with MNLI	**36.1**	33.2
Adapter over prefinetuned with MNLI	34.3	33.0

Table 10: Comparison of Results Across Different Configurations on the factuality tasks

The baseline model performs strongly in bi-text mining tasks, significantly surpassing random baselines, which are based on the inverse of the dataset size (1/n). For pairwise classification tasks like Dipper, the baseline accuracy of 73.06% highlights the model's potential for applications requiring pairwise comparisons. These results emphasize the effectiveness of the baseline model in identifying document-level semantic relationships and alignments, especially in multilingual or structured datasets.

6 Label Augmentation on ATEB

We test label augmentation on factuality and safety tasks in ATEB and show its effectiveness in improving an embedding model's advanced capabilities.

6.1 Model

We adopt the Gemma V1-2B embedding model we trained as a symmetric dual encoder. We adopt two initialization settings before fine-tuning with label augmentation data. The first setting is finetuning directly over Gemma 2B. The second setting is adopting a prefinetuning stage where full supervision finetuning is conducted with 76 Huggingface Sentence Transformer datasets. [2],

6.2 Training data

We reformulate the training sets of two NLI entailment classification datasets, MNLI (Williams et al., 2018) and FaithDial (Dziri et al., 2022b) into the label augmentation setting to be used as our training data for factuality classification tasks. For safety classification tasks, we reformulate the training set of BeaverTails Safety Ranking (Ji et al., 2023) task to be used as training data.

6.3 Results

Factuality tasks. Table 10 presents the performance of various configurations on two factuality

[2]https://huggingface.co/sentence-transformers

classification tasks: ESNLI (Camburu et al., 2018) and DialFact (Gupta et al., 2022).

The random baseline accuracy for both tasks is 33% since they are both three-class classification tasks. The Gemma-2B embedding model baseline achieve 35.85% for ESNLI and 33.95% for DialFact, showing a slight improvement over random guessing. Finetuning with MNLI classification data without unique IDs as introduced in the label augmentation setting does not improve the performance. Finetuning with MNLI data equipped with unique ID also leads to no improvement. Incorporating target explanations leads to a boost in performance, yielding an improvement of 9% for ESNLI and 2.8% for the out-of-domain DialFact. Finetuning with out-of-domain, FaithDial classification data (Dziri et al., 2022a) leads to a modest increase, reaching 36.87% for ESNLI and 34.95% for DialFact. This indicates that detailed target explanations are particularly effective for in-domain finetuning entailment tasks like ESNLI.

When fine-tuning over a pre-finetuned Gemma-2B model with MNLI, performance drops to 37.61% for ESNLI and 33.5% for DialFact, showing that while pre-finetuning over generic retrieval tasks offers some benefits, it may not be as effective as direct full-supervision fine-tuning. Adapter-based fine-tuning approaches offer a trade-off between training efficiency and performance. Fine-tuning with an adapter achieves 36.1% for ESNLI and 33.2% for DialFact. When the adapter-based fine-tuning is applied to a pre-finetuned Gemma-2B model, performance decreases slightly to 34.3% for ESNLI and 33.0% for DialFact. These results suggest that adapter-based methods, while computationally efficient, do not achieve the same level of performance as full fine-tuning.

In summary, the table highlights several key insights: 1) label augmentation with label explanations provide the most substantial accuracy gains, particularly for ESNLI. 2) adapter-based fine-tuning offers a viable but much less effective alternative to full-supervision fine-tuning. 3) additionally, task-specific instructions and data augmentation strategies lead to only modest improvements unless combined with detailed target explanations or robust fine-tuning techniques.

Safety tasks Table 11 summarizes Gemma-2B performance under four training regimes—baseline, full fine-tuning, adapter fine-tuning, and "pre-finetune plus fine-tune" on two safety benchmarks. BeaverTails, which assesses content-safety ranking, starts at 55.6%, while HHRLHF, measuring alignment with human-RL feedback, begins at 50%, showing no benefit without task-specific work.

	BeaverTails(%)	HH-RLHF(%)
Random	50	50
Baseline	55.6	50.0
Reranking as retrieval		
Full-supervision Gemma 2B	**68.5**	**51.0**
Full-supervision - pre-finetuned	56.5	50.1
Adapter with BeaverTails	**59.0**	50.0
Adapter with BeaverTails - pre-finetuned	58.1	50.2

Table 11: Comparison of Results Across Different Configurations on the safety tasks.

All subsequent experiments use the same label-augmentation setup (labels plus explanations). Fine-tuning Gemma-2B directly on BeaverTails Safety-Reranking data lifts accuracy to 68.5%, a +12.9% gain, and nudges HHRLHF to 51.0%—evidence that most improvements remain in-domain. Switching to adapter-based fine-tuning reaches the identical 68.5%/51.0% while modifying only a sliver of parameters, highlighting its resource efficiency. By contrast, inserting an intermediate pre-finetuning stage on generic retrieval data hurts downstream alignment: both full and adapter approaches drop to 56.5% on BeaverTails, nearly erasing the earlier gains, though HHRLHF remains flat.

Taken together, the results show that (1) label-augmented, task-specific fine-tuning is essential for strong safety accuracy, (2) adapters can match full updates at lower cost, and (3) indiscriminate pre-training may actively degrade performance on tasks requiring precise human-preference alignment.

7 Conclusion

In conclusion, we propose a novel benchmark, ATEB, to highlight the limitations of existing embedding models in handling advanced NLP tasks. By reformulating classification and reasoning tasks as retrieval problems with label augmentation, our approach enables embedding models to leverage their strengths in capturing semantic relationships, thereby extending their capabilities. Through extensive experimentation, we demonstrate that our fine-tuning method can significantly enhance performance on tasks involving factuality and safety. These results underscore the importance of tailored benchmarks and innovative training strategies in advancing the development of more capable embedding models.

8 Limitations

While we included 21 tasks in our benchmark, many other safety, reasoning, and factuality tasks could be incorporated to increase the diversity and complexity of the benchmark. Additionally, we evaluated our proposed data reformulation method only on factuality and safety tasks and did not test it on other task categories.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *CoRR*, abs/1908.05739.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. PIQA: reasoning about physical commonsense in natural language. *CoRR*, abs/1911.11641.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, and Jared et al. Kaplan. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, volume 31. Curran Associates, Inc.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, and Others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, volume 36, pages 30039–30069.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics (TACL)*, 10:1473–1490.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022a. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022b. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

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

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3785–3801.

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

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yuan Jiang and Kai et al. Chen. 2023. Introducing mistral: Efficient pretraining of large transformer models. *arXiv preprint arXiv:2309.XXXX*.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel Weld. 2022. GENIE: Toward reproducible and standardized human evaluation for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11444–11458.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Frederick Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024b. Gecko: Versatile text embeddings distilled from large language models. *Preprint*, arXiv:2403.20327.

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

X. Meng et al. 2024. Sfr-embedding-mistral: Leveraging decoder-only transformers for state-of-the-art text embeddings. *arXiv preprint arXiv:2401.XXXX*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

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

Niklas Muennighoff et al. 2023b. Mteb: The massive text embedding benchmark. *arXiv preprint arXiv:2301.XXXX*.

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. 2022. Text and code embeddings by contrastive pre-training. *CoRR*, abs/2201.10005.

OpenAI. 2024. Hello GPT-4o.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744.

Frank Palma Gomez, Ramon Sanabria, Yun-hsuan Sung, Daniel Cer, Siddharth Dalmia, and Gustavo Hernandez Abrego. 2024. Transforming LLMs into cross-modal and cross-lingual retrieval systems. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 23–32, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste

Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WINOGRANDE: an adversarial winograd schema challenge at scale. *CoRR*, abs/1907.10641.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 624–643.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *Preprint*, arXiv:2407.12883.

Gemma Team, Thomas Mesnard, and Many other Authors. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, Bhaskar Mitra, Jimmy Lin, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation. *Preprint*, arXiv:2407.10817.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2024. Rar-b: Reasoning as retrieval benchmark. *Preprint*, arXiv:2404.06347.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Training details

Training dataset size We use the reformulated training sets of the publicly available datasets in training our factuality and safety models. We adopt the train split in the original tasks.

Hyper-parameters We did not use hard negatives in prefinetuning. We used a batch size of 1024, learning rate of $1e-4$. The number of training steps is $100,000$. The number of warmup steps is st to $20,000$ and the input length is 256, the output length is 1024. We used unmixed batches during training and bidirectional loss.

We finetune both the factuality models and safety models with 20k iterations and a batch size of 1024. Our learning rate is set as $1e-4$ with linear decay.