FORTIFY: Generative Model Fine-tuning with ORPO for ReTrieval Expansion of InFormal NoisY Text

Dan DeGenaro[1] **Eugene Yang**[2,3] **David Etter**[3] **Cameron Carpenter**[2]
Kate Sanders[2] **Alexander Martin**[2] **Kenton Murray**[2,3] **Reno Kriz**[2,3]
[1]Georgetown University; [2]Johns Hopkins University;
[3]Human Language Technology Center of Excellence
Correspondence: drd92@georgetown.edu

Abstract

Despite recent advancements in neural retrieval, representing text fragments or phrases with proper contextualized embeddings is still challenging. Particularly in video retrieval, where documents are text extracted through OCR from the frames or ASR from audio tracks, the textual content is rarely complete sentences but only a bag of phrases. In this work, we propose FORTIFY, a generative model fine-tuning approach for noisy document rewriting and summarization, to improve the downstream retrieval effectiveness. By experimenting on MultiVENT 2.0, an informational video retrieval benchmark, we show Llama fine-tuned with FORTIFY provides an effective document expansion, leading to a 30% improvement over prompting an out-of-box Llama model on nDCG@10. Zero-shot transferring the model tailored for MultiVENT 2.0 to two out-of-distribution datasets still demonstrates competitive retrieval effectiveness to other document preprocessing alternatives. Our training script and generated preference training data are publicly available at `https://available.after.acceptance/`.

1 Introduction

In typical ad hoc retrieval, documents are usually assumed to be well-formed and informative, such as news articles, blog posts, or social media threads (Craswell et al., 2020; Lawrie et al., 2023a, 2024; Thakur et al., 2021). While some may be more structured and readable than others, they generally convey information in a way that is easily understandable to human readers. Since neural retrieval models, such as Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) and ColBERT (Khattab and Zaharia, 2020), leverage pretrained language models (Devlin et al., 2019; Zhuang et al., 2021) trained on natural language to encode documents, they typically achieve strong performance on such tasks.

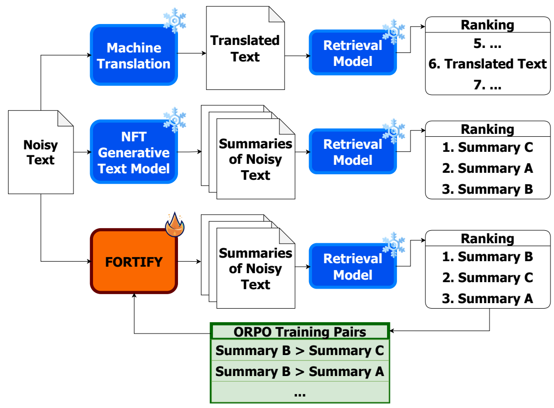

Figure 1: Overview of our document expansion approaches. Machine translation serves as a baseline. In the NFT (no fine-tuning) approach, we use a generative text model to generate fluent, keyword-dense summaries of noisy, multilingual text. In FORTIFY, we further rank the generated summaries using a retrieval model to create training pairs for preference optimization and fine-tune with Odds Ratio Preference Optimization (ORPO).

However, in many real-world settings, documents contain noisy or fragmented text, which does not resemble typical human communications. While this is relatively rare in traditional ad hoc retrieval, it is much more common when text is extracted from other modalities, such as automated speech recognition (ASR) from audio, or optical character recognition (OCR) from images or videos. Because this textual content is automatically generated, it may contain recognition errors, misidentifications, and incorrect reading order (de Oliveira et al., 2023), often resulting in disjointed sentence fragments or even incomplete words. As a result, neural retrieval models struggle to represent these texts effectively, leading to weaker retrieval performance.

To address this challenge, we propose a document expansion and rewriting approach using a generative model to transform fragmented text into coherent passages. We first explore a zero-shot

Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025), pages 100–115
August 1, 2025

prompting approach and demonstrate the innate ability of generative models like Llama3 (Dubey et al., 2024) to reconstruct text. While this method is promising, generating meaningful summaries from unordered, disjointed tokens remains a significant challenge. To further instill retrieval-driven preferences into the generative model, we fine-tune it using Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024), a technique that does not require an explicit reference model or reward function. We name this method FORTIFY, or **F**ine-tuning with **O**RPO for **R**e**T**rieval expansion of **I**n**F**ormal nois**Y** text.

We evaluate our approach on multiple video and cross-language retrieval benchmarks, and demonstrate that expanding raw documents with generated summaries leads to significant and robust performance improvements. Additionally, we find that FORTIFied summaries further boost retrieval effectiveness. To our knowledge, this is the first work to apply preference optimization to document expansion for retrieval.

Our contributions are threefold:

1. We introduce a novel document expansion approach which leverages a generative model to reconstruct fragmented text into coherent passages.

2. We propose FORTIFY, a fine-tuning mechanism using ORPO to encourage a language model to learn retrieval-driven preferences.

3. We conduct extensive experiments across multiple retrieval modalities and settings, demonstrating the effectiveness and robustness of our methods.

2 Related Work

Text Retrieval Recently developed neural retrieval models leverage pretrained language models to encode documents into one (Karpukhin et al., 2020; Formal et al., 2021; Nguyen et al., 2023) or multiple (Khattab and Zaharia, 2020; Li et al., 2023) contextualized embeddings to achieve better (Thakur et al., 2021) and more robust retrieval effectiveness, even in multilingual retrieval (Lawrie et al., 2023a, 2024). However, because of their pretraining data (Chari et al., 2023), they are not well-tuned for retrieving informal or even fragmented text (DeLucia et al., 2022; Lawrie et al., 2023b; Thakur et al., 2021). While recent work, such as RAPTOR (Sarthi et al., 2024), tries to preprocess text through layers of summarization, these models still anticipate well-formed text as the input. Particularly in video retrieval, text is extracted from different modalities and thus may be ill-formed. Neural text retrieval models suffer when dealing with this kind of text.

Video Retrieval Traditional benchmarks for video retrieval (Chen and Dolan, 2011; Krishna et al., 2017; Xu et al., 2016) generally involve generic web images or three to five-second video clips paired with web-scraped or automatically generated captions. Methods typically compute visual features from these images or from sampled video frames that can be mapped to these natural language captions (Cao et al., 2024; Luo et al., 2022; Reddy et al., 2025; Wang et al., 2024). However, there has been a shift away from these tasks to harder tasks requiring multimodal understanding, like audio and overlaid text, and longer videos (Kriz et al., 2024; Wang et al., 2019). This has lead to a rise in multimodal models that jointly incorporate modalities (Chen et al., 2023; Liu et al., 2025; Wu et al., 2025). However, these approaches are not robust to these challenging benchmarks, with one significant factor being the fusion of noisy outputs from OCR and ASR compounding errors and decreasing performance.

Multimodal Text Extraction Alongside visual captioning, optical character recognition (OCR) and automatic speech recognition (ASR) are two of the primary approaches to map multimodal data to natural language descriptions.

Recently, vision-language foundation models, such as PaliGemma (Beyer et al., 2024), InternVL (Chen et al., 2024), Idefics2 (Laurençon et al., 2024), and LLaVa (Liu et al., 2023), have been explored for modeling OCR content implicitly and effectively, rendering standard OCR approaches unnecessary, e.g., MMOCR (Kuang et al., 2021), and TrOCR (Li et al., 2022). Recent work has also explored using document screenshots for retrieval (Ma et al., 2024), an approach that relies heavily on the quality and the format of the screenshots. Retrieving documents with noisy OCR content (or otherwise working with such content) remains challenging.

Recent advances in ASR have achieved impressively low word error rates (Kheddar et al., 2024). However, speech involving code-switching (Yan et al., 2023), multiple speakers (Watanabe et al., 2020), or noisy environments (Dua et al., 2023; Li et al., 2014) all still present significant challenges to producing clean transcripts. Such transcripts are frequently incoherent despite low word error rates, motivating works involving post-hoc correction to the ASR output (Ma et al., 2023).

Preference Optimization Preference optimization (Rafailov et al., 2024; Shao et al., 2024; Xu et al., 2024; Meng et al., 2024; Hong et al., 2024) has arisen as a common alternative to reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020) to alleviate the multi-stage procedure requiring a reward model (Casper et al., 2023). Many recent works have built on DPO: replacing pair-wise preference data (Cai et al., 2024; Ethayarajh et al., 2024), with sets of reference responses in a log-likelihood loss (Xu et al., 2024; Park et al., 2024). In this work, we adopt Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024), which incorporates an odds ratio-based loss for differentiating the generation styles between preferred and non-preferred responses. Compared to ordinary DPO, ORPO aligns better with the goal of producing fluent, coherent generations for downstream retrieval due to its inclusion of an additional language modeling loss term, along with the odds ratio term.

3 Methods

In this section, we describe our initial document expansion approach without fine-tuning (No-fine-tune – NFT), along with FORTIFY, a novel method for optimizing machine-generated document expansion for information retrieval.

Given a noisy document d, NFT involves zero-shot prompting a generative model for one or more summaries $\hat{d}_{1,\ldots,N}$ from d, focusing on maximizing the inclusion of synonyms and keywords to enhance retrieval performance. These summaries are then used to augment the original document, producing an expanded version in the form $d + \hat{d}_1 + \ldots + \hat{d}_N$, where $+$ denotes concatenation.

FORTIFY further refines this expansion by optimizing machine-generated summaries based on their relevance to corresponding queries. Given a retrieval method, NFT summaries are scored against the corresponding queries, and training pairs are constructed by pairing the highest-scoring summary with several lower-scoring alternatives. This enables a retrieval-driven preference optimization.

3.1 Challenges in Noisy Text Retrieval

With frequency-based approaches such as BM25 (Robertson et al., 1995, 2009), retrieval performance degrades significantly in the presence of typographical errors, text recognition errors (e.g., substitution of visually similar characters), speech transcription errors (e.g., substitution of phonetically similar letters), and other character-level inaccuracies (de Oliveira et al., 2023). For example, if we attempt to retrieve a noisy document containing song lyrics that were recognized via OCR from a music video using the name of the musical artist as a query, we are unlikely to succeed, as the artist's name may not appear in the video. However, by leveraging a generative model to produce a summary, we not only correct character-level errors but also elaborate on the content and introduce useful keywords and phrases. An example is shown in Appendix C, Figure 5.

While neural retrieval models are more robust to character-level errors, they still struggle with higher-level structural issues, particularly ill-formed sentences and unrelated, adjacent phrases. This is because such noisy documents are rarely seen in the training data used for modern neural retrieval models (Nguyen et al., 2016). Consider a single video frame containing multiple distinct spans of text, such as two lines on a blackboard, each containing a chemical equation. To retrieve this video from the extracted text, we must flatten or concatenate all text spans to apply standard text retrievers. This process often produces incoherent outputs. Such text is likely to suffer not only from recognition errors, but also a lack of coherence, sentence structure, or recognizable words. By applying a generative model, we can reconstruct meaning from the fragmented text prior to indexing. A strong generative model can correctly identify the text as chemical equations and even suggest relevant elements and compounds. Notably, it can also extract and contextualize useful keywords such as *chemical*, *reactions*, and *compounds*, further improving retrievability. See Appendix C, Figure 6 for an example of this.

3.2 Zero-Shot Expansion of Noisy Text

We propose expanding noisy documents with such machine-generated summaries by leveraging modern generative models' abilities to produce clean, coherent, and keyword-dense text. As an initial setting, we adopt a zero-shot approach, where we provide the noisy text and prompt a generative model to produce a keyword-dense summary. The generated summaries can then either be indexed directly or concatenated with the original text; in later sections, we utilize the concatenation approach.

This method provides several advantages. Since modern generative models are highly multilingual, noisy documents can be expanded into any language, potentially improving the alignment between documents and expected queries for both term frequency and neural retrieval models. For instance, in cross language retrieval, where queries are primarily in English, we can prompt the model to produce English summaries of multilingual documents, effectively translating key phrases while preserving retrieval relevance. Additionally, by explicitly prompting the model to focus on synonyms, keywords, and retrieval relevance, summary-based document expansion introduces semantically related terms, improving retrieval effectiveness when queries lack important keywords.

Beyond improving term matching, generative document expansion also addresses structural issues in noisy documents. By generating coherent, well-formed summaries, the model compensates for disjointed or ill-structured inputs, producing text that is more suitable for retrieval. While generative model inference is computationally expensive, document expansion occurs at indexing time rather than search time, minimizing computational overhead during retrieval.

3.3 FORTIFY Preference Optimization

Zero-shot inference on generative models is heavily dependent on the prompt, which leads to instability in the generation (Jiang et al., 2020; Gao et al., 2021; Errica et al., 2024; Chakraborty et al., 2023). To improve the robustness of the generation process, we further fine-tune the model with preference examples based on the downstream retrieval task. Typically, fine-tuning the generative model for document expansion through reinforcement learning requires an explicit reward function on the final retrieval effectiveness and a preference model on the retrieval system. However, defining the reward is challenging as the query distribution is often unknown at training and indexing time, leaving great uncertainty in the direction of optimization. Therefore, we use Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024), a variant of Direct Preference Optimization (Rafailov et al., 2024) without defining a reference model, to provide preference signals during fine-tuning.

Specifically, let $\hat{d}_x$ and $\hat{d}_y$ be two generated summaries of a raw document d. For a pointwise retrieval model $f(q, d)$ and a query q that document d is relevant to, we define the preference of the retrieval model $f(q, d)$ as

$$\hat{d}_x \succ \hat{d}_y \quad \text{if and only if} \quad f(q, \hat{d}_x) > f(q, \hat{d}_y) \tag{1}$$

where $\succ$ indicates the left operand is more preferable than the right operand.

Following Hong et al. (2024), the odds ratio loss of the preference $\hat{d}_x \succ \hat{d}_y$ can be written as

$$\mathcal{L}_{OR} = -\log \sigma \left(\log \frac{\mathbf{odds}_\theta(\hat{d}_x|d)}{\mathbf{odds}_\theta(\hat{d}_y|d)} \right) \tag{2}$$

where the function $\mathbf{odds}_\theta$ indicates the odds of generating such a sequence of text based on the parameter θ. Such odds ratio losses promote the generative model to generate $\hat{d}_x$ over $\hat{d}_y$ when given the document d based on the preference of the retrieval model f and the query q. Intuitively, the distribution of the training query q and pre-defined retrieval model f are critical to this process since the model would be biased toward the two after fine-tuning. In our experiments, we provide empirical evidence that the resulting generative model is actually robust to the downstream retrieval models.

4 Experiments

4.1 Data

We evaluate FORTIFY on two video retrieval datasets as well as a cross-language text retrieval dataset as an out-of-domain evaluation. The statistics are summarized in Table 3 in the Appendix.

- MultiVENT2.0 (Kriz et al., 2024) consists of 218K YouTube videos, with text and speech content primarily in Arabic, Chinese, English, Korean, Russian, and Spanish. The videos vary heavily in terms of production quality, from unprocessed recordings taken on mobile phones to professionally edited news broadcasts. Queries are designed to approximate what a user might search for in order to find a video about a specific event. We evaluate on the test split (2,546 queries over 109K videos) and report nDCG@10 and R@1000 following Kriz et al. (2024).

- TextVR (Wu et al., 2025) consists of 42.2K queries over 10.5K videos from across eight domains: Street View (indoor), Street View (outdoor), Game, Sports, Driving, Activity, TV Show, and Cooking. We evaluate on the test split, containing 2.7K videos, with one query each, and report R@1 and R@10 to align with the online shared task associated with TextVR.

- NeuCLIR Chinese Technical CLIR Collection (Lawrie et al., 2024) contains about 396K journal abstracts from 1,980 Chinese academic journals spanning 67 disciplines. The NeuCLIR Technical document collection has two corresponding sets of topics from the 2023 and 2024 TREC NeuCLIR tracks, respectively. To ensure the summarization process is not trivially easy, we use only the abstract without the title as the raw document. We report the official evaluation metrics of the NeuCLIR track, which are nDCG@20 and R@1000.

4.2 Text extraction from video

In order to create textual indices for retrieval, we extract text from the videos using two main approaches: Automatic Speech Recognition (ASR) and Optical Character Recognition (OCR). Except where explicitly indicated, we do not perform machine translation on either the ASR or OCR text.

ASR Videos frequently contain audio, and for our ASR system, we rely on a powerful multilingual model, Whisper Large v2 (Radford et al., 2023) without speech translation (that is, audio detected by Whisper as language x is transcribed in language x, not in English). As Whisper Large v2 is among the top-performing open-source ASR models (even outperforming proprietary models as shown in the authors' appendix), and as it is highly multilingual and trained on diverse sources of data, its outputs are fairly accurate across domains and more commonly used languages. If the speech extracted from a video is indeed useful for retrieval, Whisper is likely to give the strongest baseline for retrieval using ASR.

OCR We further extract text OCR using the hybrid model described in Etter et al. (2023). This is a state-of-the-art multilingual model which was found to significantly outperform many popular open-source OCR models and toolkits on the test split of the highly multilingual CAMIO OCR dataset (Arrigo et al., 2022), including Tesseract (Smith et al., 2009), EasyOCR, TrOCR (Li et al., 2022), and MMOCR (Kuang et al., 2021) across a variety of different scripts.

4.3 Baseline Document Expansion

As a baseline, ASR and OCR texts are summarized by prompting Llama-3-8B-Instruct (Dubey et al., 2024; AI@Meta, 2024) without additional fine-tuning (*No-fine-tune (NFT) summaries*). For each video, the ASR content is placed into a prompt template that explicitly directs Llama to produce a keyword-dense summary useful for information retrieval. This prompt is shown in Appendix B, Figure 3.

Summaries are generated by passing the ASR or OCR text to the Llama-3-8B-Instruct model with a generation limit of 512 tokens, no repeated trigrams, and using top-p sampling with $p = 0.9$ and a temperature of 0.6. The raw ASR or OCR (or the concatenation of both) text is expanded with the summaries by concatenation. Processing MultiVENT 2.0's test split (109K videos), assuming the text is already extracted, took approximately 36 hours on eight 40GB A100 GPUs.

Alternatively, we expand the raw documents with their machine translation since the extracted ASR or OCR text is not necessarily English, which is the query language of the three evaluation collections. For MultiVENT 2.0, since the collection is large, we use NLLB (Costa-jussà et al., 2022), an open-source machine translation model that covers more than 200 languages, to translate the extracted ASR and OCR text. For TextVR, we use Google Translate to obtain the translation through their Web APIs. Finally, for NeuCLIR Technical Documents, we use the official translation provided by the NeuCLIR track, which is also produced by Google Translate.

4.4 FORTIFY Fine-tuning Setup

We fine-tune Llama-3-8B-Instruct to produce more useful summaries using an original dataset of preferred and dispreferred summaries (contrastive training pairs, as required to proceed with ORPO). The summaries included in this dataset were produced using the subset of the training split of MultiVENT 2.0, totaling 2,000 videos, for which training queries were written. For each of the unique query-video pairs having OCR content, we prompt Llama to produce a keyword-dense summary suited to information retrieval, given the OCR content.

To ensure high quality summaries in the training set, we use a one-shot prompt template, shown in Appendix B, Figure 4, containing the extracted OCR text from a manually selected video in MultiVENT's training set, along with a manually written summary to produce more accurate summaries for training.

We sample from Llama-3-8B-Instruct five times to produce five distinct summaries of the OCR content with the same generation setting. We then score each of the generated summaries against their relevant queries using the PLAID-X implementation of ColBERT (Khattab and Zaharia, 2020; Santhanam et al., 2022; Yang et al., 2024b) (details are discussed below). Finally, we construct training summary pairs by pairing the highest-scoring summary for a particular video's OCR with each of the lower-scoring summaries. We repeat a nearly identical process to produce summaries of the ASR content but with a different prompt template containing the extracted ASR text from a particular video along with a manually written summary. This dataset is split into 80-20 train-dev splits for FORTIFY fine-tuning.

We perform a LoRA (Hu et al., 2021) fine-tuning process on Llama-3-8B-Instruct with ORPO using the implementation provided by Huggingface[1], with LoRA matrices of rank 16, $\alpha = 32$, and dropout probability 0.05. We target the up, down, Q, K, V, and O projection layers during fine-tuning. We train for three epochs over 12K training pairs, sampling randomly from the training pairs. We employ a paged AdamW 8-bit optimizer with a learning rate of $8 \cdot 10^{-6}$, $\beta = 0.1$ (called λ in the ORPO paper), and 10 linear warmup steps. We accumulate gradients over 4 batches of size 2.[2]

4.5 Retrieval Models and Pipeline

We test FORTIFY on three retrieval models, BM25 (Robertson et al., 1995, 2009), DPR (Karpukhin et al., 2020), and ColBERT (Khattab and Zaharia, 2020), while only fine-tuning Llama with FORTIFY on ColBERT. For BM25, we use the implementation provided by PyTerrier (Macdonald et al., 2021) with $k_1 = 1.2$, $k_3 = 8$, and $b = 0.75$. For DPR, we use Tevatron (Gao et al., 2022) with a multilingual DPR model based on DistilBERT (Sanh, 2019) provided by `sentence-transformers` (Reimers and Gurevych, 2019) that is fine-tuned on the Quora dataset.[3] Documents are encoded and indexed with FAISS (Douze et al., 2024) without approximation. Finally, we use the PLAID-X (Yang et al., 2024c) implementation for ColBERT with 1-bit residual compression. Documents are encoded with a Multilingual ColBERT-X (Nair et al., 2022; Lawrie et al., 2023c) model trained with Multilingual Translate Distill (Yang et al., 2024a) from the Mono-mT5-XXL cross-encoder (Jeronymo et al., 2023). [4] Additionally, we report results using an English-to-Chinese cross-language ColBERT-X model [5] on the NeuCLIR Technical Document task for comparison. Results can be seen in the Appendix, Table 4.

5 Results and Analysis

For MultiVENT 2.0 (the dataset on which FORTIFY is trained), presented at the left part of Table 1, expanding the original OCR, ASR, or both (OCR+ASR) with summaries generated by FORTIFY provides a significant improvement over no expansion or expansion with their machine translation. When using ColBERT on the FORTIFY-expanded OCR and ASR documents, it provides a 76% improvement in nDCG@10 (0.324 to 0.569) over LanguageBind (Zhu et al., 2023), a state-of-the-art video encoding language model reported in the MultiVENT 2.0 dataset paper (Kriz et al., 2024), and 30% over no expansion (0.437 to 0.569).

Regardless of the source of text (OCR or ASR), expanding with generated summaries is more effective than using machine translation, which is an alternative document processing method (with similar hardware requirements) since the extracted text is not necessarily in the query language. Such improvements are consistent across multiple settings, indicating that the summaries are useful for a wide range of retrieval models, including statistical models like BM25.

However, since FORTIFY is trained to tailor the expansion for retrieval using ColBERT, documents expanded with FORTIFY summaries are more advantageous for ColBERT, resulting in improvement in both nDCG@10 and R@1000 over zero-shot prompting, though nDCG@10 is not statistically significant. However, the differences in

[1] https://huggingface.co/docs/trl/main/en/orpo_trainer

[2] Hyperparameter choices largely retained from this tutorial: https://huggingface.co/blog/mlabonne/orpo-llama-3

[3] https://huggingface.co/sentence-transformers/quora-distilbert-multilingual

[4] https://huggingface.co/hltcoe/plaidx-large-eng-tdist-mt5xxl-engeng

[5] https://huggingface.co/hltcoe/plaidx-large-zho-tdist-mt5xxl-engeng

Table 1: Retrieval effectiveness with different document expansion approaches. nDCG in the table uses a rank cutoff at 10. Superscript of w, x, y and z indicates the metric value using the corresponding expansion approach is statistically significantly better than the **same retrieval model** using *No Expansion* (w), *Machine Translation* (x), *No-Fine-tuned (NFT) Summary* (y), and FORTIF*ied Summary* (z), respectively (also indicated in the first column) with 95% confidence. The statistical test uses a paired t-test with multiple testing corrections over datasets and retrieval models. Rows in light gray indicate retrieval methods relying on features other than text, which is unfair to compare methods only using the extracted text but are included for border comparisons.

		MultiVENT 2.0						TextVR (*Zero-shot Transferred*)					
Expansion Approach	Retrieval Model	OCR		ASR		OCR+ASR		OCR		ASR		OCR+ASR	
		nDCG	R@1K	nDCG	R@1K	nDCG	R@1K	R@1	R@10	R@1	R@10	R@1	R@10
StarVR												0.165	0.473
LanguageBind						0.324	0.846					0.133	0.830
(w)*No Expansion*	BM25	0.157	0.267	0.114	0.204	0.195	0.322	0.141	0.278	0.044	0.097	0.160	0.305
	DPR	0.088	0.334	0.146	0.482	0.153	0.532	0.042	0.120	0.036	0.089	0.051	0.148
	ColBERT	0.317	0.616	0.344	0.583	0.437	0.740	0.134	0.259	0.051	0.114	0.153	0.292
(x)Machine Translation	BM25	0.319^{w}	0.592^{w}	0.300^{w}	0.559^{w}	0.427^{w}	0.733^{w}	0.147	0.297^{w}	0.046	0.100^{w}	0.168^{w}	0.325^{w}
	DPR	0.166^{w}	0.500^{w}	0.198^{w}	0.513^{w}	0.236^{w}	0.629^{w}	0.043	0.117	0.037	0.092^{w}	0.052	0.148
	ColBERT	0.375^{w}	0.633^{w}	0.401^{w}	0.589	0.517^{w}	0.760^{w}	0.131	0.260	0.051	0.114	0.155	0.304^{w}
NFT (y)Llama Summary	BM25	0.360^{wxz}	0.646^{wxz}	0.351^{wxz}	0.606^{wxz}	0.492^{wxz}	0.788^{wxz}	0.156^{w}	**0.314**^{wx}	**0.054**^{wx}	**0.128**^{wx}	0.178^{w}	0.346^{wx}
	DPR	0.237^{wx}	0.575^{wxz}	0.249^{wxz}	0.554^{wx}	0.318^{wx}	0.708^{wxz}	0.059^{wx}	0.164^{wx}	0.034	0.099^{w}	0.067^{wx}	0.191^{wx}
	ColBERT	0.429^{wx}	0.675^{wx}	0.434^{wx}	0.616^{wx}	0.564^{wx}	0.795^{wx}	0.147^{wx}	0.282^{wx}	0.047	0.122	0.167^{wx}	**0.329**^{wx}
(z)FORTIFied Summary	BM25	0.350^{wx}	0.630^{wx}	0.333^{wx}	0.595^{wx}	0.475^{wx}	0.779^{wx}	**0.160**^{wx}	0.312^{wx}	0.052^{w}	0.123^{wx}	**0.180**^{wx}	0.356^{wxy}
	DPR	0.241^{wx}	0.564^{wx}	0.240^{wx}	0.547^{wx}	0.315^{wx}	0.699^{wx}	0.059^{wx}	0.159^{wx}	0.036	0.104^{wx}	0.059^{w}	0.183^{wx}
	ColBERT	**0.431**^{wx}	**0.688**^{wxy}	**0.435**^{wx}	**0.623**^{wxy}	**0.569**^{wx}	**0.805**^{wxy}	0.144^{x}	0.278^{wx}	0.053^{y}	0.123^{wx}	0.168^{wx}	0.319^{wx}

R@1000 are significant, indicating that the FORTIFY-expanded documents include more related terms to the expansion but are not more accurate than what zero-shot prompting the generative model can provide. When using BM25 and DPR to encode and index the FORTIFY-expanded documents, since they are not the predefined customer of the summarization model, the resulting retrieval metrics are only similar or slightly lower than NFT summaries, which also indicates that FORTIFY can effectively tailor the document expansion to the expressed preferences of the downstream retrieval model during fine-tuning.

Interestingly, although DPR significantly underperforms with respect to ColBERT, the improvement due to expansion with generative summaries is much larger for DPR than for ColBERT, which validates our initial intuition that it is possible to leverage the linguistic ability of a generative model to provide additional context and language structure for the downstream neural retrieval model to consume. Since DPR encodes the entire piece of text as a single dense vector, providing it with better-structured documents is more advantageous for DPR than ColBERT, which is capable of falling back to term matching through dense token embeddings. Without such expansion, DPR is even less effective than BM25 as shown in the *No Expansion* condition in Table 1. When using both OCR and ASR text, DPR improves 106% in nDCG@10 when expanding with FORTIFY summaries (0.153 to 0.315) while ColBERT *"only"* demonstrates a 30% improvement (0.437 to 0.569). Even compared against machine translation, which already processes and potentially denoises the raw and noisy text via a language model, DPR still improves 33% when using FORTIFied summaries while ColBERT *"only"* improves by 10%.

5.1 Out-of-Distribution Transfer

Zero-shot transferring FORTIFY to TextVR, which demonstrated a very different distribution both in videos and extracted text (presented in Table 3), the differences between zero-shot prompting and the FORTIFY-fine-tuned summarizer are small and not statistically significant. Since the distribution of the queries and the videos are significantly different from MultiVENT 2.0, on which the model was trained, the additional preference optimization through ORPO is not particularly helpful but also not harmful. Such robustness indicates the FORTIFY-fine-tuned model still retains its original language modeling capability to support generalization while providing more beneficial information when preferences of the downstream retrieval model were communicated during fine-tuning. Interestingly, expanding ASR and OCR text with FORTIFied summaries using BM25 is still 9% more effective in R@1 (0.165 to 0.180) than StarVR, proposed along with the introduction

Table 2: Retrieval effectiveness when concatenating multiple sources of text in MultiVENT 2.0 using ColBERT. nDCG values in the table uses a rank cutoff at 10. Checkmarks indicate inclusion of such source of text in the documents for ColBERT indexing.

Original Noisy Text		Machine Translation		FORTIFied Summary			
OCR	ASR	OCR	ASR	OCR	ASR	nDCG	R@1K
	✓					0.317	0.616
✓						0.344	0.583
✓	✓					0.437	0.740
✓	✓	✓	✓			0.517	0.760
✓	✓			✓	✓	0.569	0.805
✓	✓	✓	✓	✓	✓	0.578	0.797

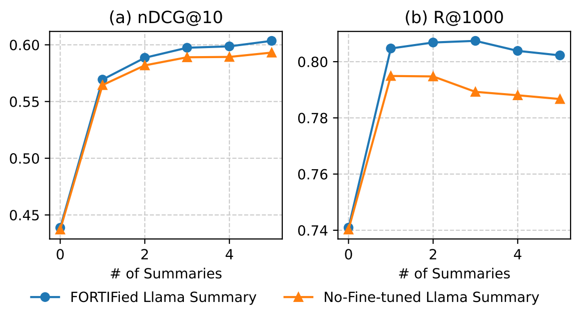

Figure 2: Effectiveness of concatenating multiple generated summaries on MultiVENT 2.0 using both OCR and ASR text.

of TextVR (Wu et al., 2025). Notably, StarVR involves a heavy video space-time encoder, as well as projection from a scene text encoder.

Note that since the amount of text extracted via ASR from the audio tracks of the videos in TextVR is scarce (only on average 185 characters per video), no expansion approach can expand the short text in any meaningful way, resulting in roughly the same effectiveness as forgoing document expansion.

5.2 Expansion with Multiple Summaries

Given the variability of generative models, we investigate generating multiple summaries using both the NFT Llama and FORTIFied models on MultiVENT 2.0. Illustrated in Figure 2, concatenating more summaries provides marginal improvements in both nDCG@10 and R@1000. However, such improvements quickly start to diminish as more summaries are added, as expected. Particularly in R@1000, expanding the noisy text with five summaries produces documents whose meanings begin to drift away from those of the original texts. This results in the promotion of more irrelevant videos to the top 1000 and thus decreases R@1000 when adding more than three summaries. Notably, FORTIFied summaries, despite still inducing a semantic drift, are still more effective than the NFT version, indicating that FORTIFY consistently instills the preference into the model, even when we are generating more summaries through randomized decoding.

nDCG@10, on the other hand, continues to improve when adding more summaries, indicating that summaries are still beneficial in terms of promoting relevant videos to the top of the ranked list. Such a trade-off between the top and the bottom of the ranked list is expected when expanding queries or documents and remains an issue for neural models such as ColBERT (Wang et al., 2023).

Finally, we also investigate expanding the noisy documents with their machine translation and FORTIFied summaries. Presented in Table 2, the final retrieval effectiveness increases as we introduce more expansion to the documents. Although expansion with machine translation is less effective than FORTIFied summaries, the two expansion approaches provide complementary information to the retrieval model. Thus, combining both approaches by concatenation results in a statistically significant improvement in nDCG@10 over just using the FORTIFied summaries (0.569 to 0.578). As before, such elaborated expansion also promotes more irrelevant videos, resulting in a slightly lower R@1000.

6 Conclusion and Future Work

In this paper, we proposed a generative model fine-tuning approach FORTIFY for document expansion. FORTIFY tailors a generative model to a specific kind of noisy document and a downstream retrieval model through ORPO, a preference optimization approach. We showed that models fine-tuned with FORTIFY provide more effective expansion summaries than an out-of-the-box Llama model. The resulting FORTIFied Llama model also demonstrates robustness to documents and retrieval models beyond the ones predefined during ORPO fine-tuning.

Beyond the success of FORTIFY on noisy text, we would like to explore it on other general ad hoc retrieval tasks to tailor the retrieval to a specific domain, corpus, or even user. Given the flexibility of preference optimization, we believe FORTIFY can be adapted to arbitrary retrieval model preference.

References

AI@Meta. 2024. Llama 3 model card.

Michael Arrigo, Stephanie Strassel, Nolan King, Thao Tran, and Lisa Mason. 2022. CAMIO: A corpus for OCR in multiple languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1209–1216, Marseille, France. European Language Resources Association.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, and 16 others. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Tianchi Cai, Xierui Song, Jiyan Jiang, Fei Teng, Jinjie GU, and Guannan Zhang. 2024. Unified language model alignment with demonstration and point-wise human preference.

Meng Cao, Haoran Tang, Jinfa Huang, Peng Jin, Can Zhang, Ruyang Liu, Long Chen, Xiaodan Liang, Li Yuan, and Ge Li. 2024. RAP: Efficient text-video retrieval with sparse-and-correlated adapter. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7160–7174, Bangkok, Thailand. Association for Computational Linguistics.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Preprint*, arXiv:2307.15217.

Mohna Chakraborty, Adithya Kulkarni, and Qi Li. 2023. Zero-shot approach to overcome perturbation sensitivity of prompts. *Preprint*, arXiv:2305.15689.

Andreas Chari, Sean MacAvaney, and Iadh Ounis. 2023. On the effects of regional spelling conventions in retrieval models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2220–2224, New York, NY, USA. Association for Computing Machinery.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 72842–72866, New Orleans, Louisiana, USA. Curran Associates, Inc.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, Seattle, Washington, USA. IEEE.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the trec 2019 deep learning track. *Preprint*, arXiv:2003.07820.

Lucas Lima de Oliveira, Danny Suarez Vargas, Antônio Marcelo Azevedo Alexandre, Fábio Corrêa Cordeiro, Diogo da Silva Magalhães Gomes, Max de Castro Rodrigues, Regis Kruel Romeu, and Viviane Pereira Moreira. 2023. Evaluating and mitigating the impact of OCR errors on information retrieval. *International Journal on Digital Libraries*, 24(1):45–62.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Mohit Dua, Akanksha, and Shelza Dua. 2023. Noise robust automatic speech recognition: review and analysis. *International Journal of Speech Technology*, 26(2):475–519.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models.

Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. 2024. What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering. *Preprint*, arXiv:2406.12334.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Vienna, Austria. JMLR.org.

David Etter, Cameron Carpenter, and Nolan King. 2023. A hybrid model for multilingual ocr. In *Document Analysis and Recognition - ICDAR 2023*, pages 467–483, Cham. Springer Nature Switzerland.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. 2022. Tevatron: An efficient and flexible toolkit for dense retrieval. *ArXiv*, abs/2203.05765.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. *Preprint*, arXiv:2012.15723.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Vitor Jeronymo, Roberto Lotufo, and Rodrigo Nogueira. 2023. Neuralmind-unicamp at 2022 trec neuclir: Large boring rerankers for cross-lingual retrieval.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

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

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *ArXiv*, abs/2403.01255.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, Venice, Italy. IEEE.

Reno Kriz, Kate Sanders, David Etter, Kenton Murray, Cameron Carpenter, Kelly Van Ochten, Hannah Recknor, Jimena Guallar-Blasco, Alexander Martin, Ronald Colaianni, and 1 others. 2024. Multivent 2.0: A massive multilingual benchmark for event-centric video retrieval.

Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, Kai Chen, Wayne Zhang, and Dahua Lin. 2021. Mmocr: A comprehensive toolbox for text detection, recognition and understanding. *Preprint*, arXiv:2108.06543.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2023a. Overview of the trec 2022 neuclir track.

Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. 2024. Overview of the trec 2023 neuclir track.

Dawn Lawrie, James Mayfield, Douglas W. Oard, Eugene Yang, Suraj Nair, and Petra Galuščáková. 2023b. Hc3: A suite of test collections for clir evaluation over informal text. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2880–2889, New York, NY, USA. Association for Computing Machinery.

Dawn Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. 2023c. Neural approaches to multilingual information retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, page 521–536, Berlin, Heidelberg. Springer-Verlag.

Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. 2014. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777.

Minghan Li, Sheng-Chieh Lin, Barlas Oguz, Asish Ghoshal, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. CITADEL: Conditional token interaction via dynamic lexical routing for efficient and effective multi-vector retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11891–11907, Toronto, Canada. Association for Computational Linguistics.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2022. Trocr: Transformer-based optical character recognition with pre-trained models. *Preprint*, arXiv:2109.10282.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. 2025. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):708–724.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomput.*, 508(C):293–304.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. 2023. Can generative large language models perform asr error correction?

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. 2021. Pyterrier: Declarative experimentation in python from bm25 to dense retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4526–4533, New York, NY, USA. Association for Computing Machinery.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, New Orleans, Louisiana, USA. Neurips.

Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396. Springer.

Thong Nguyen, Sean MacAvaney, and Andrew Yates. 2023. A unified framework for learned sparse retrieval. In *European Conference on Information Retrieval*, pages 101–116. Springer.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017, Bangkok, Thailand. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M. de Melo, Benjamin Van Durme, and Rama

Chellappa. 2025. Video-colbert: Contextualized late interaction for text-to-video retrieval. *Preprint*, arXiv:2503.19009.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, and 1 others. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022. Plaid: an efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1747–1756.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Ray Smith, Daria Antonova, and Dar-Shyang Lee. 2009. Adapting the tesseract open source ocr engine for multilingual ocr. In *Proceedings of the international workshop on multilingual OCR*, pages 1–8.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2023. Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web*, 17(1):1–39.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. 2024. Disentangled representation learning. *Preprint*, arXiv:2211.11695.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, and 1 others. 2020. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings.

Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. 2025. A large cross-modal video retrieval dataset with reading comprehension. *Pattern Recognition*, 157:110818.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Brian Yan, Matthew Wiesner, Ondřej Klejch, Preethi Jyothi, and Shinji Watanabe. 2023. Towards zero-shot code-switched speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Eugene Yang, Dawn Lawrie, and James Mayfield. 2024a. Distillation for multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2368–2373.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W. Oard, and Scott Miller. 2024b. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *Proceedings of the 46th European Conference on Information Retrieval (ECIR)*.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W Oard, and Scott Miller. 2024c. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *European Conference on Information Retrieval*, pages 50–65. Springer.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, and 1 others. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Appendix

A Out-of-Domain Transfer

To evaluate FORTIFY on a completely different domain, we again zero-shot transfer the MultiVENT-FORTIFied model to generate summaries for the academic abstracts in the NeuCLIR Technical Document collection. Presented in Table 4, FORTIFied summaries still provide additional information to the original document despite not being noisy, resulting in a 43% improvement in nDCG@20 on the 2023 topics and 35% on 2024. However, the NFT Llama summary, in this case, is slightly more effective since it was trained to accomplish a wide range of tasks under a wide range of conditions.

Such differences are expected as our MultiVENT-FORTIFied model has moved from a general-purpose model to a more task-specific one. As we move further away from the original training setup, which assumes noisy, fragmented text with ColBERT being the retrieval model, the model becomes less capable of generating retrieval model-favored summaries, especially when using BM25. With that said, FORTIFY can be tailored to any domain as long as the retrieval model preference can be collected. We leave the exploration of FORTIFY to general ad hoc retrieval to future work.

B Prompts

In this section, we provide prompts that we optimize FORTIFY for. Figure 3 presents the primary prompt that we use, while Figure 4 presents the OCR-focused prompt.

C Examples

In this section we two examples of the noisy extracted text. Documents composed principally of noisy text are often difficult to retrieve (de Oliveira et al., 2023). In term frequency approaches such as BM25, performance is harmed when there are typographical errors, text recognition errors (substitution of visually similar characters), speech transcription errors (substitution of letters pronounced similarly), or other character-level errors. For instance, if we were to search for for the noisy document in Figure 5, we might not be successful if our query is “Rolling Stones” - note that neither of these words appear in the document, despite the fact that the document is very clearly the lyrics to Jumpin’ Jack Flash, albeit with significant text recognition errors. If we now produce a summary using a general-purpose generative text model, the summary not only corrects the character-level errors in the original document, but it elaborates on the content further, and finally includes a list of useful keywords and phrases.

Table 3: Dataset Statistics. Note that all three collections are multilingual. The average character counts treat all scripts (Latin, CJK, Perso-Arabic, Cyrillic, etc.) identically.

	MultiVENT 2.0 Test Set			
	Videos			
	w/OCR	w/ASR	Total	Queries
Count	105,026	109,488	109,800	2,546
Avg. # of Chars	529	1,092	–	42

	TextVR Test Set			
	Videos			
	w/OCR	w/ASR	Total	Queries
Count	2,726	2,249	2,727	2,727
Avg. # of Chars	441	185	–	73

	NeuCLIR Technical		
		Queries	
	Documents	2023	2024
Count	395,927	41	106
Avg. # of Chars	206	131	131

Table 4: Zero-shot cross-domain transfer of the MultiVENT-FORTIFied model (training on MultiVENT 2.0 training set) to the NeuCLIR Technical Document task with topics from 2023 and 2024. nDCG in this table uses a rank cutoff at 20. Rows in light gray indicate retrieval methods relying on features other than text.

Expansion Approach	Retrieval Model	**2023** nDCG	**2023** R@1K	**2024** nDCG	**2024** R@1K
English-Chinese ColBERT-X		0.339	0.783	0.338	0.796
(w) *No Expansion*	BM25	0.054	0.128	0.049	0.106
	ColBERT	0.277	0.736	0.256	0.687
(x) Machine Translation	BM25	0.239^{w}	0.588^{w}	0.240^{w}	0.588^{w}
	ColBERT	0.330^{w}	0.788^{w}	0.326^{w}	0.763^{w}
(y) NFT-Llama Summary	BM25	0.330^{wxz}	0.803^{wxz}	0.336^{wxz}	0.726^{wx}
	ColBERT	$\mathbf{0.404}^{wx}$	$\mathbf{0.838}^{w}$	$\mathbf{0.356}^{w}$	$\mathbf{0.783}^{w}$
(z) FORTIFied Summary	BM25	0.286^{w}	0.733^{wx}	0.305^{wx}	0.694^{wx}
	ColBERT	0.395^{w}	0.813^{w}	0.349^{w}	$\mathbf{0.783}^{w}$

System Prompt: You are tasked with summarizing text. This summary will be used for the task of information retrieval. As such, it is of paramount importance that the summary include as many relevant keywords, synonyms, and related words as possible. Feel free to restate the text in multiple ways. A longer document is better than a shorter one. The more keywords you include, the better the summary will be for the task of information retrieval. The summary can be in the style of a description or a synopsis.

User Prompt: Please produce a summary of the following text content that is suited for the task of information retrieval, focusing on including keywords, synonyms and related words: `<noisy text>`

Figure 3: Prompt for Llama-3-8B-Instruct for generating summaries. This prompt is used to generate both No-fine-tune (NFT) Llama summaries and the FORTIFY summaries.

System Prompt: You are tasked with summarizing the OCR text extracted from a video. This summary will be used for the task of information retrieval. As such, it is of paramount importance that the summary include as many relevant keywords, synonyms, and related words as possible. Feel free to restate the text in multiple ways. A longer document is better than a shorter one. The more keywords you include, the better the summary will be for the task of information retrieval. The summary can be in the style of a video description or a plot synopsis or summary. For instance, given the following OCR text:
Ptička is the name of the song. How to make a whistle The bird. How to make a whistle How to make a whistle The bird.
Produce a summary like so:
A tutorial or a series of instructions on how to make a whistle (a musical instrument) that mimics or imitates the sound or call produced by a bird. The tutorial mentions a song called "Ptička", which is the Russian word for "bird". The whistle may be used to replicate birdsong.

User Prompt: Please produce a summary of the following text content that is suited for the task of information retrieval, focusing on including keywords, synonyms and related words: `<noisy text>`

Figure 4: OCR-focused one-shot prompt used for generating summaries. This prompt is only used for generating training summaries for FORTIFY.

Noisy text: Kolling Sones. umping Jack Hash. One Two! But it's all right now, in fact, it's a gas! And Thowled . . .

Machine translation: Kolling Sones. umping Jack Hash. One Two! But it's all right now, in fact, it's a gas! And Thowled . . .

NFT summary: The song "Jumpin' Jack Flash" by The Rolling Stones is a rock and roll classic . . . Synonyms: tumultuous, difficult, abusive, rebellious, defiant, call toaction [sic]. Related words: rock music, classic, childhood experiences, abuse of power, individualism, resilience in the face of adversity.

FORTIFIed summary: The song "Jumping jack flash" is a popular rock song by the British rock band The Rolling stones, released in 1968 . . . the song is often interpreted as a commentary on the struggles and hardships faced by young people.

Figure 5: An example of some extracted text with significant character-level errors and the outputs produced by the three document expansion strategies we compare (machine translation, Llama summarization without fine-tuning, and FORTIFY). Red indicates a character-level error. Yellow indicates unusual or disjointed structure. Green indicates a spelling correction. Blue indicates useful keywords and phrases not in the original text.

NOISY TEXT: t80 /ts Po4. BO5jCa(OM → Cas (PO)SMD. 3OS+ 2KOH → Kz SOu + H O. SO, +2NgOH → Na2SD, + HO. HeSO4. Ma (e D)=. t80. /ts Po4) PDS +3Ca (OHL. → CasPO%),0. 3O,+ 2K0H → ke SOu + HO. SO, + . . .
MACHINE TRANSLATION: t80 /ts Po4. BO5jCa(OM → Cas (PO)SMD. 3OS+ 2KOH → Kz SOu + H O. SO, +2NgOH → Na2SD, + HO. HeSO4. Ma (e D)=. t80. /ts Po4) PDS +3Ca (OHL. → CasPO%),0. 3O,+ 2K0H → ke SOu + HO. SO, + . . .
NFT SUMMARY: The text describes a series of chemical reactions involving calcium, phosphorus, and sulfur compounds. The reactions involve the formation of various calcium phosphates and sulfates, including calcium phosphate (Ca3(PO4)2), calcium sulfate (CaSO4), and sodium sulfate (Na2SO3). The reactions also involve the use of various reagents, including potassium hydroxide (KOH), sodium hydroxIDE [sic] (NaOH) . . . Overall, the text provides a detailed description of various chemical reactions and the compounds involved.
FORTIFY-GENERATED SUMMARY: A comprehensive guide to the synthesis of calcium phosphate (Ca3(PO4)2) is presented, providing a detailed overview of the chemical reactions involved. The process involves the reaction of calcium oxide (CaO) with phosphoric acid (H3PO4), as well as the addition of sodium hydroxide (NaOH) and other chemicals. The resulting compound, calcium phosphate, is a key component in the production of ceramic materials, fertilizers, . . . Keywords: calcium phosphate synthesis, ceramic materials production, fertilizer production, pharmaceuticals, pH levels, catalysts, sodium sulfate, sodium nitrite, chemical reactions.

Figure 6: An example of some extracted text with significant structural problems and the outputs produced by the three document expansion strategies we compare (machine translation, Llama-generated summaries without fine-tuning, and FORTIFY). Highlights mean the same as in Figure 5. Note that this document's overall structure is highly problematic as well.