GRAFT: Gated Retrieval-Augmented Fine-Tuning for Relation Extraction

Yuhang Jiang and Ramakanth Kavuluru
Division of Biomedical Informatics, Department of Internal Medicine
Department of Computer Science
University of Kentucky, Lexington, KY USA
yj38@iu.edu,ramakanth.kavuluru@uky.edu

Abstract

Even in the era of large language models (LLMs), biomedical relation extraction (RE) still plays a major role in timely creation of knowledge graphs that further guide biomedical knowledge discovery. The main task in RE is to extract a relation "as expressed" in an input text. At times, crucial definitional information or other auxiliary information about the entities involved may be missing from the input text. Augmenting it from other external textual sources appears helpful on the surface but can be harmful too, as these sources can overwhelm the signal in the original input, leading to false positives or false negatives. To counter this, we leverage a pre-trained biomedical text retriever to augment original inputs with additional instance-specific snippets. This is done through a gating mechanism that allows the retrieved snippets to enhance but not overwhelm the signal from the original input. We evaluate our approach on three standard biomedical relation extraction datasets (CDR, BioRED, and ChemProt) and show consistent improvements (up to 10 F1 points) compared with strong supervised baselines involving both encoder and decoder models. All our code and the datasets used are available for reuse: https://github.com/bionlproc/GRAFT-RE.

1 Introduction

Biomedical relation extraction (RE) is a core task in biomedical natural language processing (Huang et al., 2024), considering relations among biomedical entities drive etiology, pathology, treatment, and recovery in human diseases. Relations span the translational science spectrum from biological links (e.g., protein-protein interactions) to translational connections (e.g., gene-disease associations) and clinical information (e.g., drug-disease treatment relations). The goal in RE is to infer and classify the relation between a pair of input entities, typically from a fixed set of relation types. Relations extracted lead to knowledge graphs (KGs), which further enhance biomedical knowledge discovery via reasoning over KGs (Raghavan et al., 2021). Although LLMs have become general purpose NLP tools, KGs are still deemed essential to reduce hallucinations and ground reasoning in discovery tasks (Xiong et al., 2025; Xu et al., 2025). Hence, overall, RE remains an important primitive in BioNLP.

Consider the sentence: "We conclude that tamoxifen therapy is more effective for early stage breast cancer patients." From this sentence an RE model is expected to extract the relation (tamoxifen, TREATS, breast cancer) where tamoxifen is the *subject* entity, breast cancer is the *object* entity, and TREATS is the relation type or more formally the *predicate*. RE by definition implies that the only input provided is the sentence (or document) from which the relation is to be extracted. This means, we are **not** looking at general (global) knowledge about the entities mentioned in the input, but more interested in relations as explicitly stated (or implied) in the input (local assertions). But a given input often may not include background information about the entities needed to tease out a potential relation. This could be in the form of definitions (or glosses) of the entities in the input. This could also be in the form of hyponyms of entities involved. For the TREATS example we chose earlier, the definition of tamoxifen in the Medical Subject Heading (MeSH) vocabulary includes: "Tamoxifen acts as an anti-estrogen (inhibiting agent) in the mammary tissue, but as an estrogen (stimulating agent) in cholesterol metabolism, bone density, and cell proliferation in the endometrium." In the MeSH hierarchy we can also see that tamoxifen is a "benzylidene compound." In addition to the original text input to the RE, it might be beneficial to augment it with this extra contextual information. The contextual information can also be arbitrary pieces of text from the broad scientific literature that may enhance RE. Our current effort investi-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 920–931

gates how to do this in a way that is both *effective* and *efficient*. Neither of these objectives is easy to meet because the extra text added as additional context to the input can confuse the model and drown the signal from the original input text. So in the end, the performance may actually worsen if the context contradicts what is stated in the input. The augmented context also increases the length of the input thus increasing computational cost.

Another concern in using additional context in RE may arise if the context already contains a directly stated positive relation, which is discussed in a complex and nuanced manner in the original input. In this case, a natural question to ask is if this constitutes cheating, considering the main goal is to extract relations as expressed in the input; the context might already contain the right answer. However, over the past decade encoder LMs already ingest billions of tokens of biomedical free text during pre-training (e.g., BioBERT (Lee et al., 2020) and BiomedBERT (Gu et al., 2021)). Furthermore, autoregressive LLMs are also pre-trained on PubMed abstracts and full text articles in PMC (Yang et al., 2022; Sallinen et al., 2025; Bolton et al., 2024). Given most modern RE efforts are almost exclusively based on encoders or LLM backbones, besides the input text that is explicitly provided, the models have entire PubMed/PMC at their disposal (via LM model weights), in an implicit manner. We propose to simply make this explicit along the lines of retrieval-augmented generation (RAG), already popular when using LLMs for question answering (QA) (Lewis et al., 2020; Xiong et al., 2024). The testing is always done against gold answers derived from the original input and hence the context's influence is strictly measured against what is available in the input.

In this paper, we present **G**ated **R**etrieval-**A**ugmented **F**ine-**T**uning (GRAFT) for RE, a parametrized approach that incorporates external retrieved context as part of the input for RE. GRAFT can seamlessly integrate with both small encoder models and autoregressive LLMs. It relies solely on a small encoder based, zero-shot retriever for biomedical RE (Jin et al., 2023), requiring no supervised data for retrieval. The documents retrieved are processed in parallel, thus reducing computational overhead significantly. Additionally, we incorporate a dynamic weighting (gating) layer that amplifies useful information from the context, while filtering out less relevant content. We evaluate GRAFT on three widely used biomedical RE datasets and show promising improvements across all of them. While this paper is focused on biomedicine, the method is agnostic of the domain and can be employed as long as there is a way to retrieve general Web text using well known models for the general domain (Izacard et al., 2022; Wang et al., 2022; Ni et al., 2022). The code and datasets used for this paper are available: `https://github.com/bionlproc/GRAFT-RE`.

2 Related Work

Several efforts on retrieval-augmented biomedical tasks focus on biomedical QA (Xiong et al., 2024; Sohn et al., 2024; Jeong et al., 2024; Das et al., 2024). Specifically, Xiong et al. (2024) benchmark a spectrum of retrieval–generator pairings across BioASQ, MedQA and PubMedQA revealing that naive coupling of powerful generators with off-the-shelf BM25 retrieval leaves substantial head-room that can be exploited by hybrid dense-sparse retrievers and cross-encoder rerankers. Sohn et al. (2024) extend this line of work with RAG^2, where rationale-guided query reformulation and a lightweight snippet filter reduce noise propagation and boost answer F1 by up to 6 points on three medical QA benchmarks.

Regarding other related efforts for RAG in biomedicine, Lopez et al. (2025) present CLEAR (Clinical Entity Augmented Retrieval), a task-aware RAG pipeline that first detects clinically salient entities in a potentially long clinical note, then restricts retrieval to context windows surrounding those entities, thereby trimming the context fed to the generator. This use case is different from GRAFT's use of external context to augment an input text for RE. BiomedRAG (Li et al., 2025) retrieves chunk-based documents from an external database and concatenates them directly into the LLM input, using a custom chunk scorer trained with a perplexity-driven supervision signal. It addresses noise at the retriever level (via better chunk selection), whereas GRAFT addresses it at the model level (gated fusion). Furthermore, GRAFT supports smaller encoder models for RE while BiomedRAG is an LLM-only method.

There are two related efforts in general domain RE. First, RetrievalRE (Chen et al., 2022) uses a non-parametric kNN retriever to select examples from the training data to be used in an interpolated manner along with a conventional parametric model's predictions. Hence, there is no

use of external knowledge as in GRAFT. Next, RAG4RE (Efeoglu and Paschke, 2025b) and its fine-tuned extension (Efeoglu and Paschke, 2025a) concatenate the retrieved context directly into the prompt and rely on LLMs, without any mechanism to control the influence of retrieved documents. In contrast, GRAFT processes retrieved documents in parallel, employs a learned gating mechanism to dynamically modulate retrieval influence, and is effective with small encoder models too. To summarize, BiomedRAG (Li et al., 2025) and RAG4RE (Efeoglu and Paschke, 2025a) differ from GRAFT in two aspects: (1) both target only LLM backbones, whereas GRAFT applies equally to small encoder models and causal LLMs. (2) GRAFT processes retrieved documents in parallel rather than concatenating them, decoupling per-document context length from the number of retrieved snippets.

Finally, we note that augmenting RE with auxiliary textual or structured information is a recurring theme in our prior work: Jiang and Kavuluru (2025) use instance-adapted predicate descriptions to enrich relation representations in a dual-encoder setup, while Jain et al. (2024a) cast document-level RE as context-guided link prediction and Jain et al. (2024b) incorporate structured knowledge for cross-document RE. GRAFT extends this direction by drawing on dynamically retrieved unstructured text rather than fixed predicate definitions or curated knowledge bases.

3 The GRAFT Method

3.1 Retriever model

To eliminate the need for retriever training, we optimize our entire system using MedCPT, a zero-shot biomedical IR model (Jin et al., 2023) trained on PubMed search logs by researchers at the US National Library of Medicine. In biomedical RE, the primary source of information lies in the entities themselves; entity traits can offer critical insights that inform the eventual relationship. Therefore, the retrieved information about the entities serves as ancillary context, aiding the prediction model in determining the relationship. We prepare two kinds of queries for MedCPT retrieval:

1. **Entity-wise query** to search for documents that describe or characterize an entity. We use the query *"What is the <entity_type> <entity_string>?"*. This is launched for both entities involved.
2. **Entity-pairwise query** to retrieve documents that discuss both entities and may reveal interactions between them. This query is formatted as: *"What is the interaction type between the <entity1_type> <entity1> and the <entity2_type> <entity2>?"*.

For biomedical RE, PubMed is an ideal choice for fetching snippets for augmentation and MedCPT comes pre-loaded with it. Additionally, we incorporate Wikipedia due to its broad complementary coverage of various biomedical concepts. Thus, the resulting documents retrieved constitute two components, the entity-wise documents and pairwise documents, both featuring PubMed abstracts and Wikipedia articles.

3.2 Zero-shot RE

To assess zero-shot performance in conjunction with the retrieved documents, we merge these documents with the input text for RE as part of the prompt. We test this strategy with the `Llama-3.2-Instruct-3B` model (Grattafiori et al., 2024), which allowed us to assess how effective the retrieved MedCPT documents are without training data. To ensure consistent outputs, we prompt the model to produce responses in a predefined JSON format, as shown in Figure 6 (of the Appendix). We present the results in Figure 1 for two datasets: chemical-disease relations (CDR (Li et al., 2016)) and chemical-protein relations (ChemProt (Krallinger et al., 2017)). We notice that the zero-shot prediction results remain suboptimal. The retrieved documents clearly hurt the performance as they likely introduced excessive noise when used in this simple way to augment the original input. The absence of clearly defined filtering mechanisms to distinguish between genuinely useful background information and irrelevant context may have caused the model to overlook or misinterpret cues essential for RE. Thus, simply appending external context can worsen results.

3.3 Fine-tuning GRAFT

To address the issues with naive augmentation, we propose using a fine-tuned model rather than a purely zero-shot approach. Although frontier LLMs can alleviate certain issues in zero-shot settings (Brokman et al., 2025), fine-tuning a smaller model offers a more cost-effective option. But several challenges remain unresolved: (1) incorporating extensive external documents into the trans-

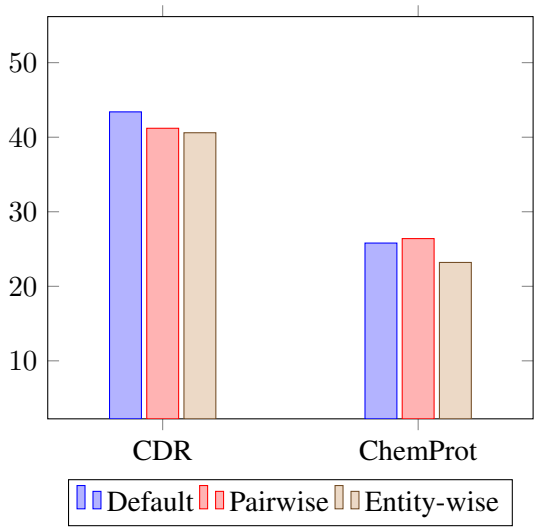


Figure 1: Zero-shot performance (in F1 scores) on CDR and ChemProt datasets using different types of retrieved documents. "Default" indicates that no external documents were applied. Up to 6 snippets were tried and the top score is reported.

former architecture's self-attention mechanism remains computationally expensive, incurring $O(n^2)$ computational complexity, and (2) failing to discard irrelevant documents still introduces additional noise into the model. The relation could be easily determined from the original context in some simple cases; the retrieved documents for these instances may lead to errors.

Therefore, instead of appending documents, we opt to input documents in parallel for two reasons: (1) we assume each document carries its own distinct information and should be processed independently, and (2) this reduces the computational overhead that would arise from cross-document interactions. As a result, the generative capability of language models is replaced by a BERT-like classification approach. Furthermore, inputting documents in parallel drastically reduces context length limitations, enabling the method to be applied to smaller models as well.

Specifically, let $\mathcal{M}$ be any transformer based model, x the input relation text, and $D = \{d_1, d_2, \ldots, d_n\}$ the set of retrieved documents. The model encodes the input text and each document independently:

$$h_x = \mathcal{M}_t(x), \quad h_i = \mathcal{M}_d(d_i), \ \forall i \in \{1, \ldots, n\}.$$

We form the representation by concatenating token-wise the text and document embeddings, then pass it through a lightweight, two-layer self-attention module to enable cross-interaction:

$$H_x = Self\,Attention([h_x \parallel h_1 \parallel h_2 \parallel \ldots \parallel h_n]).$$

Furthermore, we design a gating mechanism that allows less attention to unimportant documents or when the input text is very informative, needing less external context interference. It employs an MLP layer as the scorer to evaluate the original input text and assign a dynamic weight for the retrieved documents. We define this weight as

$$g_x = \sigma\left(MLP(h_x)\right) \in (0, 1).$$

Before we discuss how this weight g_x is incorporated into the training and inference process, we first ought to cover how the relation is represented. For bi-directional models, we explicitly mark each entity span with unique boundary tokens so the model can localize and differentiate the arguments of the relation. For example, the entities `acetaminophen` and `liver injury` are marked as: "`The analgesic [E1] acetaminophen [/E1] may induce [E2] liver injury [/E2] in susceptible patients.`" Let h_x^{E1} and h_x^{E2} be the contextualized output of `[E1]` and `[E2]`, the relation representation $Rep(h_x)$ in bi-directional model is defined as

$$Rep(h_x) = LayerNorm([h_x^{E1} \parallel h_x^{E2}]).$$

In causal (autoregressive) models, where each token attends solely to preceding tokens, we enclose the entities in boundary markers and append an additional `[EOT]` token at the end of the sequence, for the same example: "`The analgesic [E1] acetaminophen [/E1] may induce [E2] liver injury [/E2] in susceptible patients. [EOT].`" Let $h_x^{[EOT]}$ be the contextualized output of `[EOT]`, the $Rep(h_x)$ in causal model form is defined as

$$Rep(h_x) = LayerNorm(h_x^{[EOT]}).$$

With this setup for $Rep(h_x)$, we derive final relation labels via

$$\hat{h}_x = \mathbf{W_1}(Rep(h_x))$$

where $\mathbf{W_1}$ is the projection layer such that $\hat{h}_x$ is an r-dimensional vector with r being the total number of possible relations (label cardinality). Without the gating mechanism, the loss function is

$$\mathcal{L}_{\text{orig}} = -\sum_{j=1}^{r} y_j \log(\hat{h}_x^j).$$

As H_x is essentially h_x with retrieved documents concatenated, the notion of $Rep(\cdot)$ is applicable to H_x too. Therefore, the aggregated relation representation controlled by the weight g_x is

$$\hat{H}_x = (1-g_x)\mathbf{W_1}(Rep(h_x)) + g_x\mathbf{W_2}(Rep(H_x))$$

where $\mathbf{W_2}$ is a projection matrix just as $\mathbf{W_1}$. Now the aggregate loss function

$$\mathcal{L}_{\text{agg}} = -\sum_{j=1}^{r} y_j \log(\hat{H}_x^j),$$

and the full loss incorporates both the original and the aggregate losses (involving retrieved docs) into

$$\mathcal{L} = \mathcal{L}_{\text{agg}} + \mathcal{L}_{\text{orig}}.$$

During inference, the output logits ought to be informed by the augmentation of the retrieved documents. Hence, the aggregated form $\hat{H}_x$ is used to predict the relation.

4 Experiments

4.1 Datasets

We conduct the experiments on three commonly used biomedical relation extraction datasets: CDR, ChemProt and BioRED. Table 1 displays the resulting processed statistics of the datasets.

- **CDR** (Li et al., 2016) is annotated at the entity level for chemical-disease inducement relations of PubMed abstracts. Each entity is normalized to a unique identifier (e.g., *lidocaine* → *D008012*). Therefore, one identifier could sometimes refer to multiple entity strings. Because there are only two entity types: chemical and disease, and one relation type (inducement), the task simplifies to binary classification. To create a more difficult and realistic evaluation setting, we sever the linkage between mentions that share the same identifier, splitting them into separate relation instances. This adjustment ensures that each entity mention is handled independently, making the task more challenging than the standard setup that operates on normalized entities.

- **BioRED** (Luo et al., 2022) dataset is built on top of the original CDR abstracts, but it greatly expands both the entity and relation inventories, incorporating six entity types (chemical, disease, gene/protein, ...) and eight relation categories. Like CDR, BioRED includes normalized entity annotations, so we process it in the same manner as CDR dataset.

- **ChemProt** (Krallinger et al., 2017) captures interactions between chemicals and genes or proteins, collected from PubMed abstracts and spanning five relation types. Because the annotations are at the entity mention level, we apply no additional preprocessing.

Dataset	# Pred.	# Train	# Valid	# Test	Nor.
CDR	1	8441	8340	8502	Yes
BioRED	8	37286	11407	11880	Yes
ChemProt	5	17991	11328	15474	No

Table 1: Statistics of datasets used (columns 3–5 are numbers of instances/relations used in our experiments).

4.2 Retrieval corpora

Corpus	#Snippets	Ave. L	Domain	Entity-wise	Pairwise
PubMed	23.9M	296	Biomed.	✓	✓
Wikipedia	29.9M	162	General	✓	

Table 2: Statistics of the corpora; **Entity-wise** and **Pairwise** denote whether the corpus is applied with entity-wise query or pairwise query.

For retrieval, we employ two corpora, PubMed abstracts and Wikipedia articles, which together provide broad and diverse coverage of biomedical knowledge. PubMed serves as our biomedical corpus, whereas Wikipedia represents the general-domain corpus. As described in Section 3.1, we generate two retrieval queries for every entity pair: an entity-wise query, executed on both corpora, and an entity-pairwise query, applied only to PubMed abstracts due to efficiency considerations. The details of each corpus are in Table 2.

Dataset	#Pairs	#Mentions	%PubMed	%Wiki
CDR	21853	5384	69.4%	30.6%
BioRED	58998	6048	66.4%	33.6%
ChemProt	28449	12310	60.3%	39.7%

Table 3: Statistics on the retrieved snippets for all three datasets, combining the training, validation, and test splits; **#Pairs** and **#Mentions** denote the counts of unique pairs and unique mentions, respectively.

We query with each entity's surface form, so when multiple mentions share the same identifier

Model	*CDR*			*BioRED*			*ChemProt*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BiomedBERT	57.6	67.9	62.3	52.9	53.2	53.1	81.1	75.4	78.2
GRAFT w/ BiomedBERT (ours)	58.6	77.4	**66.7** ↑4.4	55.8	53.7	**54.7** ↑1.6	80.5	78.8	**79.6** ↑1.4
BERT-based									
Llama-3.2-3B-Instruct	69.6	62.0	65.6	54.2	51.6	52.9	73.4	59.9	66.0
w/ Retrieval	67.6	58.7	62.8	54.2	50.2	52.1	73.7	55.6	63.4
Llama-3.1-8B-Instruct	62.7	76.8	69.0	58.8	54.2	56.4	77.2	66.4	**71.4**
w/ Retrieval	67.4	71.2	69.2	59.2	55.1	57.1	77.2	65.4	70.8
GRAFT w/ Llama-3.2-3B-Instruct (ours)	67.9	74.7	**71.2** ↑5.6	64.7	62.7	**63.7** ↑10.8	76.6	65.6	70.6 ↑4.6
Causal LMs									

Table 4: Experimental results of GRAFT compared to corresponding baseline models. Our method consistently outperforms the corresponding baselines and surpasses the larger 8B model on CDR and BioRED. Improvements are shown with ↑, representing gains over the baseline using the same transformer model.

(e.g. *Parkinson's disease* and *PD*), every mention is submitted to the retriever separately. Following Xiong et al. (2024), we chunk each corpus into short snippets to enhance retrieval granularity, reduce noise, and ensure that the retriever returns passages focused on the target entities. These snippets are ranked by MedCPT according to the relevance scores it assigns; for efficient fine tuning, we select the top one to five snippets and feed them into GRAFT. Statistics of retrieved snippets are displayed in Table 3. Although Wikipedia holds more snippets overall, entity-wise queries return substantially more snippets from PubMed, reflecting our emphasis on biomedical tasks. Since the number of unique mentions is much smaller than that of unique pairs across all three datasets, using entity-wise queries for retrieval is considerably more efficient.

4.3 Baseline models

As outlined in Section 3.3, we evaluate our framework in two architectural settings: (i) bidirectional encoder models and (ii) left-to-right (causal) language models. For the causal-LM track, we adopt the recently released Llama-instruct models: `Llama-3.2-Instruct-3B` and `Llama-3.1-Instruct-8B`. For the encoder-based setting, we select BiomedBERT (Gu et al., 2021), a transformer encoder pre-trained exclusively on PubMed abstracts. We fine-tune BiomedBERT following Zhong and Chen (2021), a commonly used fine-tuning approach for RE tasks.

In baseline experiments, we fine-tune BiomedBERT using only the original text; its input-length limit of 512 tokens prevents appending any retrieved documents. (Note that this restriction applies only to naive augmentation; GRAFT allows for retrieval augmentation because of parallel processing of individual snippets.) For Llama models, we conduct two separate fine-tuning regimes: one on the raw input and one on document-augmented input, both optimized with the standard language modeling objective:

$$\mathcal{L}_{\mathrm{LM}}(\theta) = -\sum_{t=1}^{T} \log P_\theta\big(x_t \mid x_{<t}\big).$$

We employ the instruction template from Figure 6, consistent with the zero-shot configuration. We provide all model choices in Table 5.

Model	Size	Context Length	Domain
BiomedBERT	110M	512	Biomed.
Llama-3.2-Instruct	3B	128k	General
Llama-3.1-Instruct	8B	128k	General

Table 5: Model selection of our baseline experiments.

4.4 Main results

We present our results in two model categories: bidirectional encoders (e.g., BiomedBERT) and causal LMs. As shown in Table 4, GRAFT fine-tuned on BiomedBERT outperforms standard BiomedBERT fine-tuning, yielding F1 improvements of 1.4 % to 4.4 %. This demonstrates that even encoder models with limited context length

Model	*CDR*			*BioRED*			*ChemProt*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
BiomedBERT *(baseline)*	57.6	67.9	62.3	52.9	53.2	53.1	81.1	75.4	78.2
GRAFT-Entity-wise	58.6	77.4	**66.7**	55.8	53.7	**54.7**	80.5	78.8	**79.6**
GRAFT-Pairwise	58.0	71.1	63.9	-	-	-	81.5	77.4	79.4
BERT-based									
Llama-3.2-3B-Instruct *(baseline)*	69.6	62.0	65.6	54.2	51.6	52.9	73.4	59.9	66.0
GRAFT-Entity-wise	67.9	74.7	**71.2**	64.7	62.7	**63.7**	76.6	65.6	**70.6**
GRAFT-Pairwise	62.7	76.8	69.0	-	-	-	77.4	64.8	**70.6**
Causal LMs									

Table 6: Results comparing different retrieval query criteria; BioRED results for pairwise snippets are excluded due to the vast number of unique entity pairs.

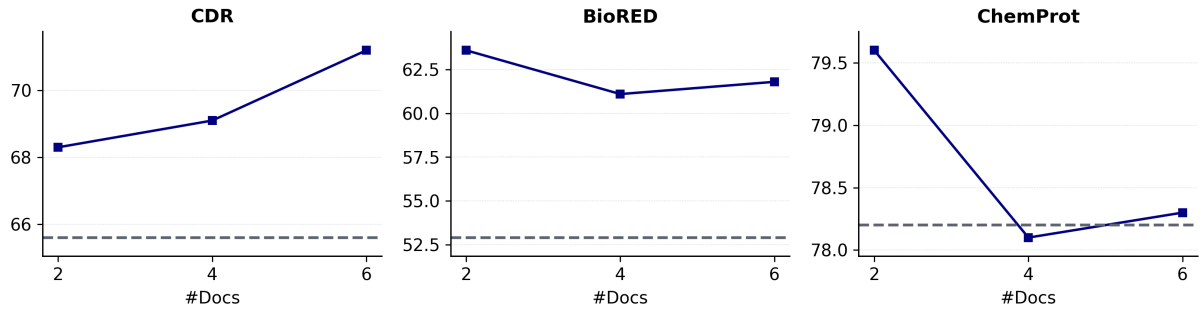


Figure 2: F1-score comparison based on different numbers of entity-wise documents retrieved.

can benefit from retrieval augmentation when fine-tuned with GRAFT. For causal LMs, we also evaluate a retrieval-augmented variant of vanilla fine-tuning to leverage their extended context capacity. GRAFT-tuned Llama-3.2-3B-Instruct outperforms the baseline on all three datasets, with F1 gains of 5.6%–10.8%. Additionally, GRAFT with the 3B-parameter Llama model surpasses a larger 8B-parameter Llama on the CDR and BioRED datasets by 2.0%–6.6% F1.

The magnitude of improvement varies across datasets and tracks the difficulty of the underlying task. ChemProt has a high baseline F1 in the BiomedBERT setting (78.2 vs. 62.3 for CDR and 53.1 for BioRED), leaving relatively less headroom for improvement; the corresponding 1.4 F1 gain with GRAFT likely reflects this ceiling effect. Conversely, BioRED, the most challenging dataset with eight relation types and the lowest baseline F1 (52.9 for Llama-3B, 53.1 for BiomedBERT), shows the largest gain (10.8 F1 with Llama-3B).

We observed an interesting phenomenon: standard fine-tuning of the 3B-parameter Llama model with retrieved documents actually degrades performance compared to fine-tuning without document augmentation. However, with the larger 8B-parameter model, this performance drop is alleviated, suggesting that the additional complexity from retrieved documents is more challenging for smaller LMs to manage. GRAFT enables smaller language models to benefit from retrieval augmentation, which is an advantage that, with naive concatenation, is only realized by larger architectures.

We note that the *w/Retrieval* rows of Table 4 are akin to the core RAG4RE design (Efeoglu and Paschke, 2025a): retrieved snippets are concatenated into the prompt and the model is fine-tuned with the standard language modeling objective, without any gating or noise-control mechanism. The gap between these rows and GRAFT rows isolates the contribution of gated parallel encoding rather than retrieval itself.

We also present results for various document-retrieval schemes in Table 6. As BioRED contains over 50K unique relation pairs, we did not evaluate pairwise retrieval on it. Across the other two

datasets, entity-wise queries consistently achieve higher performance for both BERT-based models and causal LMs. Since each dataset has fewer unique mentions than unique pairs, using entity-wise queries is inherently more efficient.

5 Discussion

5.1 Varying number of snippets

We vary the number of retrieved snippets/documents and evaluate how this affects performance. The results are shown in Figure 2. To ensure the efficient fine-tuning, we experiment with retrieving between 2 and 6 snippets, finding that all three datasets deliver strong results with just two retrieved snippets. Increasing the number of retrieved snippets led to performance declines on BioRED and ChemProt, implying that the inclusion of too much information can degrade performance. From a computational angle, limiting retrieval snippets also keeps both memory footprint and latency low. With fewer snippets retrieved, the model can more accurately assign attention, reducing the risk of overfitting to noisy context.

5.2 Ablation on corpus domains

We found that the choice of corpus domain has an impact on model performance. To quantify this, we performed ablation experiments, omitting one corpus at a time, and present the results in Figure 3. For BioRED and CDR, the mixed-corpus setup consistently achieved the highest average F1. In particular, augmenting PubMed with Wikipedia improved CDR by 3% F1 and BioRED by 2% F1, demonstrating that general-domain text can effectively complement biomedical literature. Conversely, relying solely on one domain proved insufficient for these tasks, confirming that a hybrid corpus strategy yields the most robust results. One reason could be because Wikipedia has general introductory exposition on many biomedical topics and hence could offer a clearer signal that is easy to process semantically compared with complicated jargon-loaded long sentences typically seen in PubMed abstracts. With ChemProt, which targets chemical–gene/protein interactions, we observe a slightly different pattern: PubMed-only retrieval delivers near-optimal performance. Adding Wikipedia incurs a 1% F1 drop. The results indicate that for datasets like ChemProt, where interactions are narrowly defined, domain-specific literature is more essential.

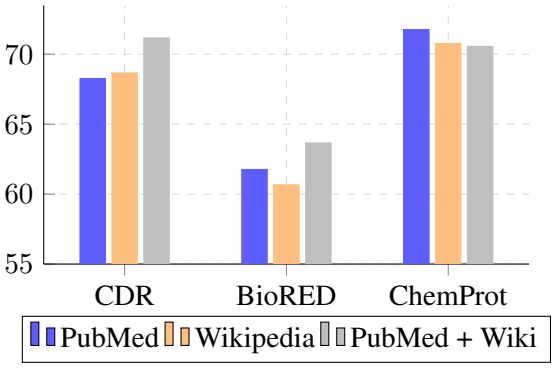


Figure 3: Model performance across different corpora.

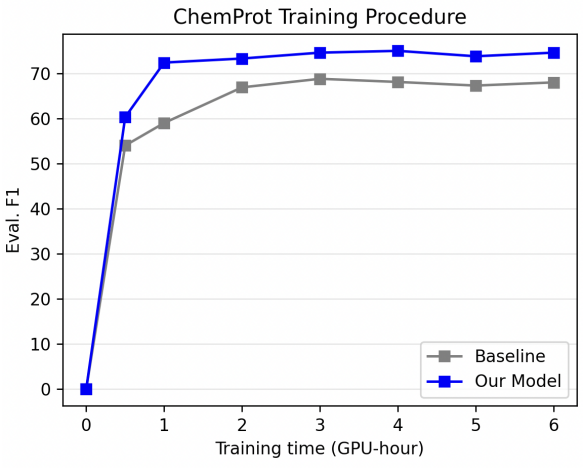
5.3 Computational efficiency

We report validation F1 scores on the ChemProt dataset across multiple epochs with our approach, comparing with the 3B-parameter Llama baseline model. As shown in Figure 4a, our model has reduced computational demands, and it converges more rapidly than the baseline, which relies on the standard language modeling objective. During inference, GRAFT requires only a single forward pass, making it faster than next-token generation methods, which require multiple forward passes. Meanwhile, when processing batched document inputs, GRAFT only requires a context window equal to the length of the longest individual document or text segment, whereas the baseline approach requires a context window large enough to accommodate all concatenated inputs. Consequently, GRAFT delivers substantially faster processing speeds than the baseline, which is shown in Figure 4b.

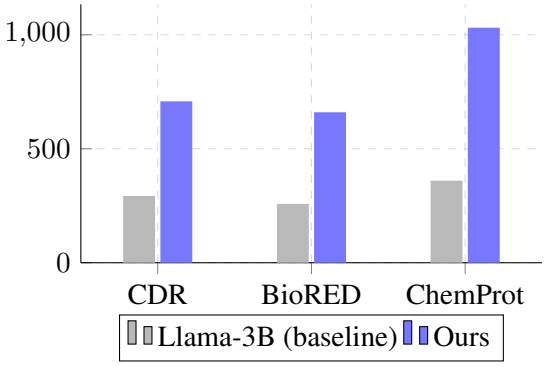
5.4 Case study: how much does prediction rely on retrieved information?

A natural concern with retrieval augmentation in RE is that retrieved snippets, drawn from corpora reflecting consensus knowledge about entity pairs, may bias the model toward predicting commonly attested relations rather than the relation as expressed in the input. This is especially relevant when the input asserts a novel, qualified, or context-specific relation that diverges from retrieved snippets. In Section 1, we posited that this is unlikely to happen since both training and testing are conducted against gold relations derived from the input.

Here to examine whether our approach addresses this concern, we look at the dynamic gate weights g_x across the entire CDR test set, shown in Figure 5 (y-axis: position in sorted order per g_x). Across



(a) Validation F1 over training epochs.



(b) Inference throughput (instances /min).

Figure 4: Computational performance: (a) learning curve and (b) processing speed.

all instances, g_x never exceeds 0.4. Thus, the gated combination consistently down-weights the retrieval branch. Despite this bounded influence, GRAFT improves CDR F1 from 62.3 to 66.7—a 4.4 F1 gain—indicating that retrieval supplies useful background cues without overwhelming local assertions in the input. This behavior is learned in GRAFT: the gate is trained jointly with the classifier and attenuates retrieval when the input is already informative.

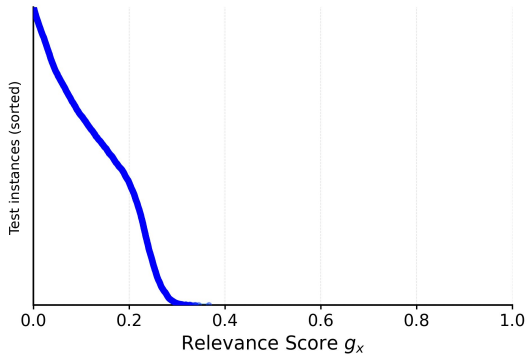


Figure 5: The gate g_x plotted (sorted) over instances of the CDR test set.

6 Conclusion

In this paper, we demonstrated that a training-free, zero-shot retrieval system can be effectively combined with task-aligned fine-tuning to address the challenges of lack of context in biomedical RE. Our proposed GRAFT framework seamlessly integrates a lightweight retriever (MedCPT) with both small encoder models (BiomedBERT) and causal language models (Llama), leveraging dynamic weighting to filter noise and amplify task-relevant information. Through extensive experiments on CDR, BioRED, and ChemProt, we showed that retrieval augmentation yields consistent F1 improvement, ranging from 1.4% to 4.4% for BiomedBERT and up to 10.8% for Llama-3B, while maintaining computational efficiency.

7 Limitations

Despite GRAFT's promise, our effort has some limitations and room for improvement. First, we evaluate on three biomedical RE datasets drawn from PubMed abstracts; generalization to clinical text or general-domain RE remains to be explored (although the method is agnostic to text type). Second, we use MedCPT as a frozen, zero-shot retriever; jointly optimizing the retriever for the RE objective could yield further improvements. Third, we tacitly assume adequate coverage of the target entities in the retrieval corpora; performance on rare or novel entities with sparse literature warrants further investigation. Finally, we use a fixed number of retrieved snippets per instance; adaptive retrieval strategies that dynamically adjust retrieval volume could offer additional gains.

Acknowledgment

This work is supported by the U.S. National Library of Medicine through grant R01LM013240. The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. National Institutes of Health.

References

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Aviv Brokman, Xuguang Ai, Yuhang Jiang, Shashank Gupta, and Ramakanth Kavuluru. 2025. A benchmark for end-to-end zero-shot biomedical relation extraction with llms: experiments with openai models. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 44–55.

Xiang Chen, Lei Li, Ningyu Zhang, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Relation extraction as open-book examination: Retrieval-enhanced prompt tuning. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2448.

Sudeshna Das, Yao Ge, Yuting Guo, Swati Rajwal, JaMor Hairston, Jeanne Powell, Drew Walker, Snigdha Peddireddy, Sahithi Lakamana, Selen Bozkurt, et al. 2024. Two-layer retrieval augmented generation framework for low-resource medical question-answering: proof of concept using reddit data. *arXiv preprint arXiv:2405.19519*.

Sefika Efeoglu and Adrian Paschke. 2025a. Fine-tuning large language models for relation extraction within a retrieval-augmented generation framework. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 1–7.

Sefika Efeoglu and Adrian Paschke. 2025b. Retrieval-augmented generation-based relation extraction. *Semantic Web*, 16(5):22104968251385519.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783. https://arxiv.org/abs/2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Ming-Siang Huang, Jen-Chieh Han, Pei-Yen Lin, Yu-Ting You, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2024. Surveying biomedical relation extraction: a critical examination of current datasets and the proposal of a new resource. *Briefings in Bioinformatics*, 25(3):bbae132.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru, and Kuldeep Singh. 2024a. Revisiting document-level relation extraction with context-guided link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18327–18335.

Monika Jain, Raghava Mutharaju, Kuldeep Singh, and Ramakanth Kavuluru. 2024b. Knowledge-driven cross-document relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3787–3797.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Yuhang Jiang and Ramakanth Kavuluru. 2025. Relation extraction with instance-adapted predicate descriptions. In *Proceedings of the AMIA Annual Symposium*, pages 546–555.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2025. Biomedrag: A retrieval augmented large language model for biomedicine. *Journal of Biomedical Informatics*, 162:104769.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, et al. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, et al. 2021. emrKBQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73.

Alexandre Sallinen, Antoni-Joan Solergibert, Michael Zhang, Guillaume Boyé Boyé, Maud Dupont-Roc, Xavier Theimer-Lienhard, Etienne Boisson, Bastien Bernath, Hichem Hadhri, Antoine Tran, Tahseen Rabbani, Trevor Brokowski, Meditron Medical Doctor Working Group, Tim G. J. Rudner, and Mary-Anne Hartley. 2025. Llama-3-meditron: An open-weight suite of medical LLMs based on llama-3.1. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2024. Rationale-guided retrieval augmented generation for medical question answering. *arXiv preprint arXiv:2411.00300*.

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

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Guangzhi Xiong, Eric Xie, Corey Williams, et al. 2025. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 7849–7857.

Ran Xu, Patrick Jiang, Linhao Luo, Cao Xiao, Adam Cross, Shirui Pan, Jimeng Sun, and Carl Yang. 2025. A survey on unifying large language models and knowledge graphs for biomedicine and healthcare. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6195–6205.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 50–61.

Appendix

Relation Extraction Template for LMs

Model Input: You are a helpful medical expert, and your task is to extract the relation between the given entity pairs using the relevant documents. The documents serve as a reference only; the primary focus for relation extraction should be on the input text. Organize your output in a json formatted as Dict{{"answer_choice": str(A/B/C/...)}}. Your responses will be used for research purposes only, so please have a definite answer.
Here are the relevant documents:
<Documents>
User: Here is the input text for relation extraction:
<Text>
What is the relationship between the subject <SUBJ> and the object <OBJ>?
Here are the potential choices: A. ... B. ... C. ... D. ... X. ...
Please generate your output in json.
Assistant:
Model Output: {"answer_choice": "X"}

Figure 6: An example prompt for RE with documents provided used for zero-shot methods.