DIAL-RE: DUAL-REGION ACTIVE LEARNING WITH EVIDENCE-GUIDED RELATION VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Incomplete relation annotations create a difficult acquisition problem: uncertain entity pairs can refine a classifier's boundary, while high-confidence unannotated pairs may reveal missed relations. We present DIAL-RE, an iterative framework that allocates queries to both regions and uses an LLM to check candidate relations against textual evidence. Accepted supervision augments a biomedical relation extractor across training rounds. We evaluate the framework on BioRED and CirrhoRED, a liver-cirrhosis corpus containing 800 documents with relation-annotation agreement of $\kappa = 0.9325$. On BioRED, the Full configuration achieves 69.85 gold-entity (GE) micro-F1, improving over the shared encoder baseline by 3.03 percentage points. On CirrhoRED, it achieves 71.46 GE micro-F1, a gain of 2.62 points. The corresponding end-to-end scores are 59.47 on BioRED and 63.73 on CirrhoRED. Performance differences among configurations vary across datasets, and the iteration trajectory peaks before the final round. Evidence-level analyses illustrate the role of relation polarity and span extraction in supervision admission.

1 INTRODUCTION

Biomedical relation extraction links entities mentioned in scientific documents, providing structured information about diseases, chemicals, genes, and other biological entities. Domain-specific encoders have improved this task substantially (Gu et al., 2021; Lee et al., 2020). Nevertheless, the supervision available to a relation classifier is often incomplete. A document can contain many candidate entity pairs, only a subset of which receives a relation annotation. Treating every remaining pair as a reliable negative can reinforce errors when valid relations have been missed.

This creates a tension in selecting additional supervision. Pairs near a model's decision boundary may help distinguish competing labels, but a highly confident unannotated pair can also be informative: it may expose disagreement between the training annotations and the relation expressed in the text. An acquisition rule focused only on uncertainty gives such pairs little priority. Conversely, collecting only confident predictions can reinforce model errors. A useful training loop must address both which candidates to inspect and which inspection outcomes to admit.

Augmentation also changes which pairs are informative. Newly admitted supervision updates the extractor, altering its predictions on the remaining candidates. An iterative acquisition policy can respond to this changing model state, while evidence verification determines which query outcomes enter subsequent training.

We introduce DIAL-RE, a dynamic interactive active-learning framework that combines complementary candidate acquisition with evidence-guided supervision admission. It allocates queries between moderate-confidence pairs and confident unannotated predictions, then uses an LLM to assess the selected pairs against supporting text spans. Accepted relations and selected hard negatives augment the training data for the next round.

The LLM serves as a verification component during augmentation; the trained discriminative model performs relation prediction at evaluation time. Each accepted relation is accompanied by a supporting text span, making the generated supervision directly inspectable.

We study the framework on BioRED (Luo et al., 2022) and **CirrhoRED**, a document-level corpus focused on liver cirrhosis. The Full configuration improves both gold-entity (GE) and end-to-end

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

(E2E) scores relative to a shared PubMedBERT baseline. The difference between Active and Random is larger on CirrhoRED, whereas the difference between Full and Active is larger on BioRED. We analyze these dataset-dependent patterns through configuration comparisons, evidence-level cases, and iteration trajectories.

Our contributions are threefold:

- A dual-region acquisition and evidence-verification framework that couples candidate selection with iterative training-data augmentation.
- An empirical study showing GE and E2E improvements on two biomedical corpora, together with dataset-dependent configuration gains and a non-monotonic response to augmentation depth.
- CirrhoRED, an 800-document liver-cirrhosis relation corpus with four relation types and entity- and relation-annotation agreement measurements.

2 Related Work

Biomedical and document-level relation extraction. Document-level RE requires combining information across mentions and sentences, as formalized by DocRED (Yao et al., 2019). ATLOP addresses multi-label prediction and entity-specific context aggregation through adaptive thresholding and localized pooling (Zhou et al., 2021). In biomedical text, PubMedBERT and BioBERT provide domain-specific representations (Gu et al., 2021; Lee et al., 2020), while BioRED annotates heterogeneous entity and relation categories (Luo et al., 2022). BioREx integrates heterogeneous corpora (Lai et al., 2023); JTIS uses joint training with intermediate tasks for biomedical document-level extraction (Li et al., 2024); and REaMA specializes generative models through instruction tuning (Zhang et al., 2025). Our study holds the discriminative backbone fixed and examines how to acquire additional supervision.

Incomplete annotations and missing relations. Re-DocRED documents false negatives in a widely used benchmark and revises its annotations through iterative model-assisted human verification (Tan et al., 2022). This motivates inspecting unannotated pairs for potentially missed relations. SSR-PU treats incomplete document-level supervision through positive-unlabeled learning and a ranking objective (Wang et al., 2022). LogicST uses logical diagnoses to resolve pseudo-label conflicts (Fan et al., 2024). Our approach queries selected pairs and checks source evidence before admitting supervision.

Active acquisition. Active learning separates query selection from the cost and reliability of annotation (Zhang et al., 2022). In clinical relation extraction, Xie et al. (2021) combine random negative sampling with uncertainty sampling, and Liang et al. (2024) analyze performance, annotation, and computation trade-offs. Our acquisition policy explicitly includes confident unannotated pairs and links query outcomes to evidence-based supervision admission.

Evidence-guided and LLM-based extraction. DREEAM uses evidence supervision to guide attention and self-training to extend evidence retrieval (Ma et al., 2023a). ET-MIER uses entity types to guide mention identification and evidence retrieval (Li et al., 2025). These methods establish evidence as a modeling signal. Our verifier uses an extracted span to assess whether a candidate should enter the training pool. GPT-RE combines in-context demonstrations with label-induced reasoning (Wan et al., 2023), while Ma et al. (2023b) use LLMs to rerank difficult cases identified by smaller models. Long-context limitations further motivate inspecting the text used for a decision (Liu et al., 2024). In our framework, supporting spans provide a traceable link between generated labels and source text.

Confidence-based self-training. FixMatch illustrates confidence-thresholded pseudo-label admission in semi-supervised image learning (Sohn et al., 2020). Dataset Cartography uses training dynamics to characterize examples and diagnose dataset quality (Swayamdipta et al., 2020). Both connect model confidence with training-example selection. Our Full configuration uses confidence histories to revise pool labels and weights, as described in Appendix C.

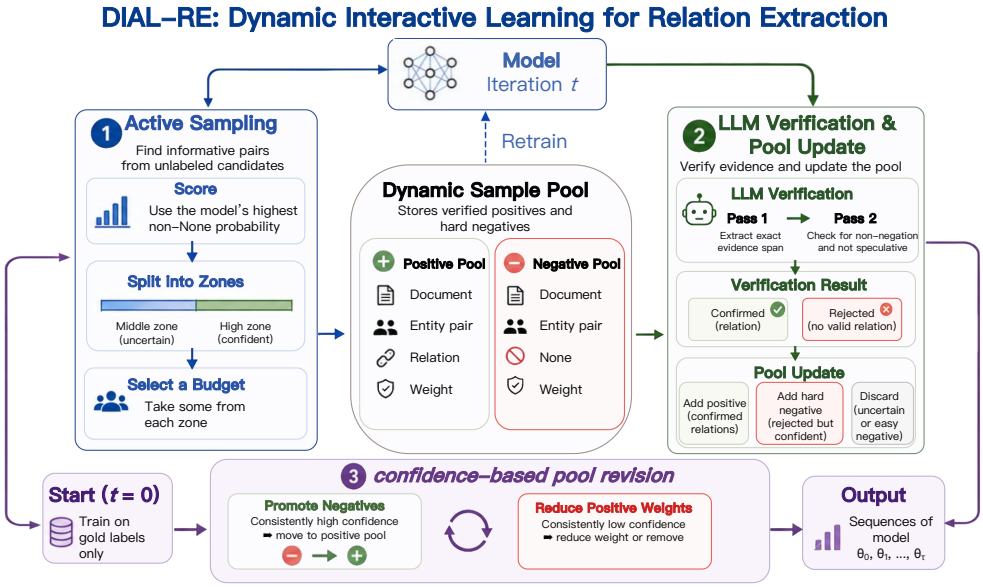


Figure 1: The DIAL-RE acquisition–verification loop. Queries produce admitted positives, selected hard negatives, or no pool entry. Updated training data change the next round's model scores. Full additionally applies an auxiliary confidence-based pool-revision operation.

3 METHOD

We start from a set of training documents with observed entity and relation annotations. For each document, the acquisition pool consists of entity pairs without an observed relation label. Our objective is to obtain additional supervision from these pairs and use it to improve the relation extractor. Figure 1 summarizes the process: score eligible pairs, query selected candidates, admit supervision, and retrain. The gold annotations remain part of the training objective throughout the loop.

3.1 CANDIDATE SCORING AND DUAL-REGION ACQUISITION

Let D be a training document, (e_i, e_j) a candidate entity pair, and $\mathcal{R}$ the set of target relations. At round t, the relation model assigns the acquisition score

$$s_t(i,j) = \max_{r \in \mathcal{R}} p_{\theta_t}(r \mid e_i, e_j, D), \qquad \texttt{None} \notin \mathcal{R}. \tag{1}$$

This maximum non-None score ranks candidates for acquisition. The unannotated candidate set $\mathcal{U}_t$ is drawn from training documents, excluding pairs already present in the augmentation pool.

We define two acquisition regions:

$$\mathcal{Z}_A^{(t)} = \{q \in \mathcal{U}_t : \beta_\ell \le s_t(q) \le \beta_h\}, \quad \mathcal{Z}_B^{(t)} = \{q \in \mathcal{U}_t : s_t(q) > \beta_h\}. \tag{2}$$

Zone A contains moderate-confidence pairs and provides exploratory queries. Zone B prioritizes high-confidence unannotated pairs for missing-label inspection. Pairs below β_ℓ are excluded from acquisition.

Using the maximum over target relations makes acquisition depend on the strongest relation prediction for each pair. Within Zone A, uniform selection spreads queries across the eligible candidates; within Zone B, ranking concentrates inspection on the strongest unannotated predictions.

For a round budget k_t, the intended Zone A allocation is $\lfloor \alpha k_t \rfloor$, sampled uniformly; the remaining allocation is drawn from Zone B in decreasing score order. The query budget decreases geometrically:

$$k_t = \max\{k_{\min}, \lfloor \gamma_q k_{t-1} \rfloor\}. \tag{3}$$

We track queried candidates, admitted labels, and LLM requests as separate quantities. Appendix A gives the experimental configuration.

3.2 Evidence verification and supervision admission

For each selected pair, the verifier receives the source document and target entities, and returns a relation with its supporting text span or a rejection:

$$V(i, j, D) = \begin{cases} (r, s) & \text{relation confirmed with supporting span } s, \\ \varnothing & \text{relation not confirmed.} \end{cases} \tag{4}$$

The verifier uses entity types, candidate relations, and existing document annotations to interpret the selected pair in context. Its first pass identifies supporting text; its second checks entity coverage, relation direction, and assertion status. A single prompt instructs the LLM to perform both passes within one call. Appendix B gives the complete prompt structure.

The instructions require evidence involving both entities and ask the verifier to avoid speculative or negated assertions. A negative-correlation statement such as "reduced levels" expresses relation polarity, whereas a sentence denying the relation expresses relation absence.

The pool contains admitted positives $\mathcal{P}_t^+$ and hard negatives $\mathcal{P}_t^-$. Admission follows

$$a_t(q) = \begin{cases} +r & V(q) = (r, s), \\ - & V(q) = \varnothing \ \wedge \ s_t(q) > \tau_{\text{neg}}, \\ \text{withhold} & \text{otherwise.} \end{cases} \tag{5}$$

The hard-negative rule targets disagreement between a confident model prediction and verifier rejection. The region boundaries control candidate selection, while τ_{neg} controls negative admission. This asymmetric rule concentrates negative augmentation on predictions that the model currently favors.

Each admitted entry records the document, entity pair, assigned label, supporting evidence when available, and a supervision weight. Positives add relation annotations to source documents. Hard negatives are represented as two-entity pseudo-documents with an empty relation set. Candidates withheld by the admission rule contribute no augmentation entry.

3.3 Iterative augmentation

The initial model is trained on the gold set $\mathcal{D}$. Subsequent rounds combine gold supervision with the accumulated augmentation pool:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{gold}}(\theta) + \sum_{e \in \mathcal{P}_t^+} w_e \ell(e, r_e; \theta) + \sum_{e \in \mathcal{P}_t^-} w_e \ell(e, \texttt{None}; \theta), \tag{6}$$

where ℓ is the instance-level focal loss. The resulting model supplies scores for the next query round.

In Equation 6, w_e scales the training contribution of an augmented example alongside gold supervision. Admitted entries remain available in later rounds, while updated scores guide acquisition from the remaining candidates. Full maintains post-training confidence histories to revise existing labels and weights, allowing supervision to evolve after its initial admission.

Active combines dual-region acquisition, evidence verification, and iterative augmentation. **Full** additionally applies confidence-based pool revision, including negative-to-positive promotion and positive-weight reduction. We evaluate Full as the complete configuration and Active as the configuration without pool revision. Appendix C describes the revision rules.

4 Experimental Setup

We organize the evaluation around overall extraction performance, differences among augmentation configurations, and the behavior of repeated acquisition. Aggregate GE/E2E results establish the benchmark comparison; iteration and per-class analyses describe where the configurations behave differently. Evidence-level cases then connect these patterns to concrete admission decisions.

Table 1: Dataset statistics. Entity and relation counts refer to stored annotations; Ent./doc. denotes the average number of entity annotations per document.

Dataset	Split	Documents	Entities	Relations	Ent./doc.
CirrhoRED	Train	640	11,984	3,068	18.7
	Dev	80	1,522	394	19.0
	Test	80	1,511	395	18.9
BioRED	Total	600	20,419	6,503	34.0

4.1 DATASETS

BioRED. We use the 600-document BioRED benchmark with its official train/development/test partition (Luo et al., 2022). Its heterogeneous biomedical entities and eight relation types provide a setting in which acquisition must accommodate different relation semantics and frequencies. The corpus totals used for the main study are shown in Table 1.

CirrhoRED. CirrhoRED contains 800 PubMed abstracts collected using the MeSH terms *Liver Cirrhosis*, *Hepatic Fibrosis*, and *Hepatocellular Carcinoma*. It covers POSITIVE CORRELATION, NEGATIVE CORRELATION, ASSOCIATION, and INTERACTION. Documents are partitioned in an 80/10/10 ratio. Relation-annotation agreement is $\kappa = 0.9325$ and entity-annotation agreement is $\kappa = 0.9837$; the corresponding agreement precision, recall, and F1 values appear in Appendix D.

4.2 CONFIGURATIONS AND TRAINING

All internal configurations use PubMedBERT-base-uncased with a biaffine relation head (Dozat & Manning, 2016). The maximum input length is 512 tokens; the head hidden dimension is 128. Training uses focal loss with exponent $\gamma_f = 2.0$ (Lin et al., 2017). The acquisition schedule runs for $T = 5$ iterations, with initial budget $k_0 = 200$, decay $\gamma_q = 0.7$, and minimum budget $k_{\min} = 30$. The intended allocation is $\alpha = 0.8$ to Zone A and 0.2 to Zone B; $\tau_{\text{neg}} = 0.65$. The oracle is Qwen2.5-7B-Instruct (Qwen Team, 2024) with temperature 0.0. Full invokes its auxiliary pool-revision procedure every two iterations.

The shared baseline uses the encoder and relation head without iterative augmentation. Random uses uniform random candidate acquisition. Active uses the dual-region acquisition and verification loop, and Full adds the auxiliary revision procedure. Hardware and model-size details appear in Appendix A.

Experiments run on a single NVIDIA A100 with 80 GB of memory. The encoder contains 110 M parameters and the relation head adds approximately 0.2 M, for 110.2 M trainable parameters. Training takes under one hour per round, with approximately 5–6 GPU hours for a five-iteration run including oracle calls. LLM computation is incurred during augmentation; evaluation uses the trained discriminative extractor.

4.3 EVALUATION

We evaluate relation extraction using micro-F1 under two settings. **Gold-entity (GE)** evaluation supplies gold entity spans and types. **End-to-end (E2E)** evaluation uses predicted entities and counts a relation as correct only when both entity spans, entity types, and the relation label match the gold annotation. All internal configurations share the same RE backbone and NER component. We report GE and E2E separately because they impose different prediction requirements.

We report internal results in Table 2 as mean $\pm$ standard deviation over three random seeds. Figure 2 presents the round-wise trajectory separately. Published BioRED results appear in a separate reference block with their original point estimates.

Table 2: Micro-F1 (%). Results for internal configurations are reported as mean ± SD across three random seeds; published references retain their original point estimates. Dashes indicate unreported results. Full includes auxiliary pool revision.

	BioRED		CirrhoRED	
Method	GE	E2E	GE	E2E
Published reference values				
BioRED baseline (Luo et al., 2022)	58.90	—	—	—
BioREx (Lai et al., 2023)	65.17	—	—	—
REaMA (Zhang et al., 2025)	68.32	—	—	—
Internal configurations				
Baseline	66.82 ± 1.58	56.32 ± 1.49	68.84 ± 1.33	60.21 ± 1.27
Random	66.77 ± 0.49	56.28 ± 0.52	65.94 ± 1.46	57.83 ± 1.38
Active	67.06 ± 0.67	56.71 ± 0.61	70.74 ± 0.80	62.11 ± 0.74
Full	$\mathbf{69.85 \pm 0.44}$	$\mathbf{59.47 \pm 0.46}$	$\mathbf{71.46 \pm 0.70}$	$\mathbf{63.73 \pm 0.36}$

5 Results and Analysis

5.1 Benchmark results and configuration contrasts

Overall performance. Full achieves the highest internal GE and E2E scores on both datasets (Table 2). Relative to Baseline, Full improves GE and E2E micro-F1 by 3.03 and 3.15 percentage points on BioRED, and by 2.62 and 3.52 percentage points on CirrhoRED, respectively. The gains carry through to extraction with predicted entities. Across all four evaluation columns, Full also has lower across-seed standard deviations than Baseline. Its BioRED GE score is 1.53 points above the published REaMA reference value.

Configuration differences. The relative gains have different profiles across datasets. On BioRED, Active improves over Baseline by 0.24 GE points, followed by a further 2.79-point difference between Full and Active. On CirrhoRED, the corresponding differences are 1.90 and 0.72 points. The larger step therefore occurs with the Full configuration on BioRED and with Active on CirrhoRED. Random is close to Baseline on BioRED and lower on CirrhoRED; Active exceeds it by 0.29 and 4.80 GE points, respectively. This pattern motivates examining the configurations at both the dataset and relation-type levels.

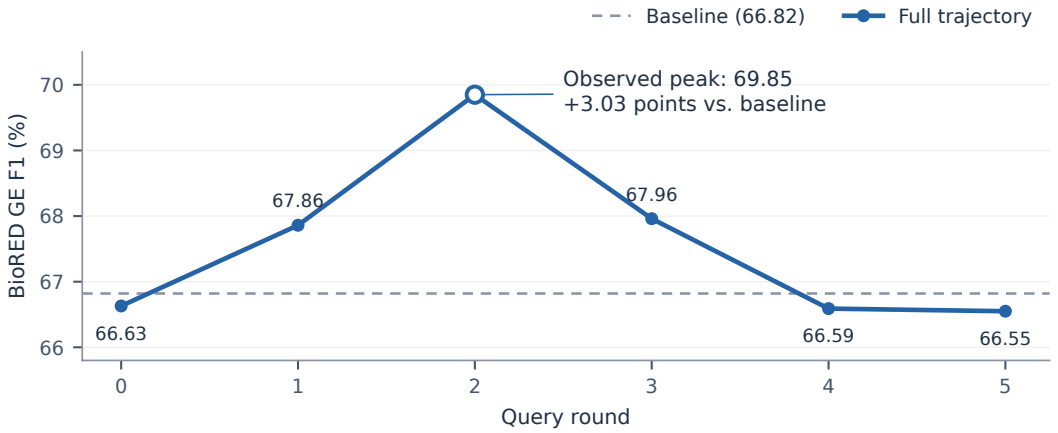


Figure 2: BioRED GE F1 across Full's augmentation rounds. The trajectory peaks at round 2 and declines in later rounds. The dashed line marks the three-seed mean of the internal baseline.

Extraction with predicted entities. The ordering Full, Active, Baseline, and Random is preserved under E2E evaluation. Entity prediction lowers the absolute scores: for Full, the GE–E2E gap is 10.38 points on BioRED and 7.73 on CirrhoRED. Because the configurations share the NER component, these results compare augmentation strategies within a common entity-prediction pipeline. Appendix E presents the GE/E2E profiles with their across-seed variation.

5.2 ITERATION BEHAVIOR

Figure 2 shows a non-monotonic response to additional augmentation. The plotted score rises from 66.63 at round 0 to 69.85 at round 2, then falls to 66.55 at round 5. The round-0 point and the separately aggregated baseline of 66.82 are distinct quantities.

This trajectory makes augmentation depth an important part of the training procedure. Later rounds combine a smaller incoming query budget with supervision accumulated earlier in the run. The pool-update rules determine which older entries continue to shape the loss as new examples arrive. Evaluating the loop therefore requires tracking both newly admitted supervision and the changing composition of the training pool.

5.3 RELATION-SPECIFIC OUTCOMES

Figure 3 shows class-specific trade-offs underlying aggregate F1. On BioRED, Active achieves the largest Comparison gain and exceeds Full on Comparison and Cotreatment. Performance on Negative Correlation varies across configurations.

On CirrhoRED, Full improves Association, Positive Correlation, and Interaction F1 relative to the baseline, while the configurations trade off performance on Negative Correlation. These results reveal relation-specific differences alongside the aggregate gains.

These class-specific rankings complement the micro-F1 comparison. Micro-F1 aggregates prediction counts across types, so relations contribute according to their annotation and prediction counts. The per-class view identifies where Active is preferable and where Full gains, giving a more detailed account of the aggregate configuration differences.

6 EVIDENCE-LEVEL ANALYSIS

High-confidence unannotated candidates. The BioRED pair *naproxen–diclofenac* illustrates high-confidence acquisition followed by evidence-based admission. At round 5, its maximum non-None score is 0.9995, and the verifier accepts a COMPARISON relation with weight 0.8. The supporting excerpt is "naproxen 1.59 (1.31, 1.93), diclofenac 1.17 (0.99, 1.38)." Both drug names occur in a comparative numerical passage. The source document and span accompany the admitted label, retaining the context in which the comparison is made.

Protective language and verifier rejection. The pair *Bifidobacterium–hepatic decompensation* illustrates a verification difficulty involving relation polarity. The source describes the bacterium as protective against future hospitalization for hepatic decompensation, supporting a NEGATIVE CORRELATION relation. The verifier nevertheless rejects the candidate despite its acquisition score of 0.9988. Here, the relational meaning is carried by protective language: identifying the two entities and their co-occurrence must be followed by interpreting the direction of the association. Appendix C describes the subsequent pool revision for this case.

Together, these examples illustrate two decisions in the learning loop: which unannotated predictions to inspect, and how to interpret their textual evidence. Appendix F.2 provides additional examples from historical configuration H, covering speculative language, entity coverage, and disease-subtype relations.

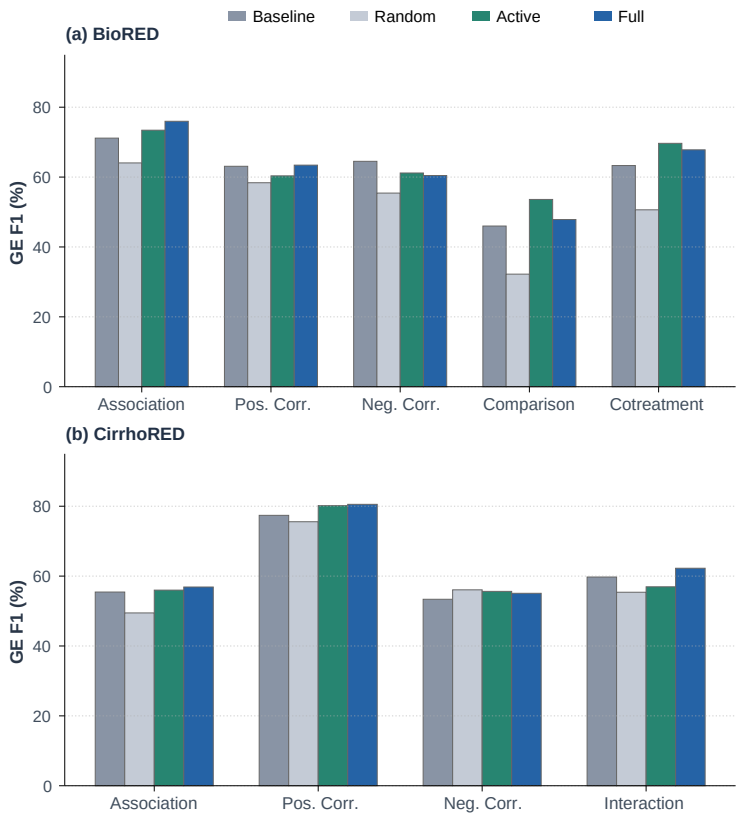


Figure 3: Per-class GE F1 for the analyzed relation types on BioRED and CirrhoRED. The configurations exhibit different performance patterns across relation types.

7 Limitations

Verification and domain coverage. Our evaluation covers two biomedical corpora and a fixed main-study verifier. Protective language, speculative statements, and entity aliases remain challenging for evidence extraction and relation verification. Future work will examine relation-conditioned prompts, broader biomedical domains, and expert assessment of admitted supervision. Evaluating different verifier sizes would also clarify the trade-off between supervision quality and augmentation cost.

Pool dynamics and component analysis. Performance varies across relation types and training rounds, with a decline after the peak of the BioRED trajectory. The current negative-pool revision centers on promotion through sustained high confidence. Confidence-based expiry and negative-side reweighting are promising extensions for managing older entries as the model changes. Further experiments will compare individual acquisition, verification, and revision choices under matched budgets, together with per-round measurements of pool composition and supervision quality.

8 Conclusion

We present DIAL-RE, a dual-region acquisition and evidence-verification framework for iterative biomedical relation augmentation. The Full configuration improves GE and E2E performance over a shared encoder baseline on BioRED and CirrhoRED, while configuration and iteration analyses reveal dataset-dependent trade-offs. CirrhoRED adds a focused resource for liver-cirrhosis relations. The results motivate further study of how acquisition and evidence quality jointly determine the value of automatically generated supervision.

AI USE STATEMENT

We used generative AI to assist manuscript organization and editing, figure-generation code, experimental-artifact inspection, and numerical consistency checks. An LLM also serves as the experimental verification component described in Section 3.2. The authors take responsibility for the final text, figures, claims, and artifacts.

ETHICS STATEMENT

This study uses biomedical literature for relation-extraction research. We provide textual evidence to support inspection of generated relations. Corpus distribution and reuse are subject to the conditions governing the original sources.

REPRODUCIBILITY STATEMENT

Section 3 specifies acquisition and supervision admission, and Section 4 describes the experimental settings and evaluation. The appendices detail the verification prompt, pool-revision rules, annotation agreement, and supplementary analyses. Accompanying materials include figure-generation scripts and data-source mappings.

REFERENCES

Timothy Dozat and Christopher D Manning. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*, 2016.

Shengda Fan, Yanting Wang, Shasha Mo, and Jianwei Niu. LogicST: A logical self-training framework for document-level relation extraction with incomplete annotations. In *Proceedings of EMNLP 2024*, pp. 5496–5510, 2024.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. doi: 10.1145/3458754.

Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Qingyu Chen, and Zhiyong Lu. BioREx: improving biomedical relation extraction by leveraging heterogeneous datasets. *Journal of Biomedical Informatics*, 146:104487, 2023.

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

Jiru Li, Dinghao Pan, Zhihao Yang, Yuanyuan Sun, Hongfei Lin, and Jian Wang. JTIS: enhancing biomedical document-level relation extraction through joint training with intermediate steps. *Database*, 2024:baae125, 2024. URL https://doi.org/10.1093/database/baae125.

Xin Li, Huangming Xu, Fu Zhang, and Jingwei Cheng. ET-MIER: Entity type-guided key mention identification and evidence retrieval for document-level relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17701–17714, 2025. URL https://aclanthology.org/2025.findings-emnlp.961/.

Siting Liang, Pablo Valdunciel Sánchez, and Daniel Sonntag. Optimizing relation extraction in medical texts through active learning: A comparative analysis of trade-offs. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pp. 23–34, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282, 2022. doi: 10.1093/bib/bbac282.

Youmi Ma, An Wang, and Naoaki Okazaki. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1971–1983, 2023a. doi: 10.18653/v1/2023.eacl-main.145. URL https://aclanthology.org/2023.eacl-main.145/.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10572–10601, 2023b.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115v1.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP 2020*, pp. 9275–9293, 2020. URL https://aclanthology.org/2020.emnlp-main.746/.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting DocRED—addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8472–8487, 2022. doi: 10.18653/v1/2022.emnlp-main.580. URL https://aclanthology.org/2022.emnlp-main.580/.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of EMNLP 2023*, pp. 3534–3547, 2023.

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proceedings of EMNLP 2022*, pp. 4123–4135, 2022. URL https://aclanthology.org/2022.emnlp-main.276/.

Weixin Xie, Limei Wang, Qi Cheng, Xueying Wang, Ying Wang, Hongyuan Bi, Bo He, and Weixing Feng. Integrated random negative sampling and uncertainty sampling in active learning improve clinical drug safety drug–drug interaction information retrieval. *Frontiers in Pharmacology*, 11: 582470, 2021.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 764–777, 2019. doi: 10.18653/v1/P19-1074. URL https://aclanthology.org/P19-1074/.

Yidan Zhang, Junlin Yu, Guobo Li, Zhenan He, and Gary G. Yen. REaMA: Building biomedical relation extraction specialized large language models through instruction tuning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(12):20258–20272, 2025. doi: 10.1109/TNNLS.2025.3596257.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. A survey of active learning for natural language processing. In *Proceedings of EMNLP 2022*, pp. 6166–6190, 2022. URL https://aclanthology.org/2022.emnlp-main.414/.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14612–14620, 2021. doi: 10.1609/aaai.v35i16.17717. URL https://ojs.aaai.org/index.php/AAAI/article/view/17717.

A Configuration and Protocol Details

Table 3 summarizes the experimental settings. The focal-loss exponent γ_f controls training-loss modulation, while γ_q controls query-budget decay.

Table 3: Main experimental configuration.

Item	Setting
Encoder / relation head	PubMedBERT-base-uncased / biaffine, hidden dimension 128
Maximum input length	512 tokens
Focal-loss exponent	$\gamma_f = 2.0$
Query schedule	$T = 5$, initial $k_0 = 200$, $\gamma_q = 0.7$, $k_{\min} = 30$
Zone allocation	$\alpha = 0.8$ to A; remaining allocation to B
Negative-admission threshold	$\tau_{\text{neg}} = 0.65$
LLM and decoding	Qwen2.5-7B-Instruct; temperature 0.0
Full auxiliary revision interval	Every 2 iterations
Trainable parameters	Encoder 110 M; head approximately 0.2 M; total 110.2 M
Hardware / reported runtime	One NVIDIA A100 (80 GB); under 1 hour per training round; approximately 5–6 GPU hours per five-iteration run including oracle calls

B Verification Prompt

We use the following two-pass verification structure within a single LLM call. Inputs include the source document, target entities, candidate relation set, and existing document annotations.

> **Inputs.** Source document; two target entity mentions and semantic types; relations already annotated for the document; candidate relation set.
>
> **Instruction.** Determine whether the text explicitly states a relation between the two entities. Co-occurrence alone is insufficient. Do not infer beyond the source text.
>
> **Pass 1: evidence extraction.** For each candidate relation, identify the shortest verbatim substring that explicitly expresses it between the entities. The span must contain both entity mentions. If no such span exists, return `NO_EVIDENCE`. Reject spans expressing negation or speculation.
>
> **Pass 2: verification.** Check that both entities occur in the span, that the relation is stated affirmatively and non-conditionally, that its direction matches the candidate, and that the span is not merely background, citation, or hypothesis. Return Yes only if these conditions hold; otherwise return No with an explanation.
>
> **Output.** Return the relation and supporting span for a confirmed candidate, or no confirmed relation otherwise.

C Auxiliary Pool Revision in Full

This appendix describes the confidence-based pool-revision procedure used in Full. The procedure revises negative labels and positive-entry weights using confidence histories.

Promotion. For each negative pool entry, confidence is recorded after training in each augmentation round. Full evaluates the promotion rule every two rounds. At a scheduled evaluation, an entry is eligible for promotion if its three most recent post-training confidence scores exceed τ^+ and it satisfies the additional minimum-confidence requirement. The entry is assigned the current model's most probable relation. Confidence histories track persistent model disagreement with the assigned negative label.

Positive-weight reduction. A positive entry whose confidence stays below a relation-specific threshold τ_r^- for ν_r iterations has its weight updated by

$$w_e \leftarrow \max(w_e - \delta_w, 0). \tag{7}$$

Entries with weights below $w_{\min}$ can be removed, subject to a minimum-count constraint for each relation type. Frequent and rare types use different thresholds or windows.

Protective-language case. The *Bifidobacterium–hepatic decompensation* case records rejection at $t = 2$ with score 0.9988, a confidence history $\langle 0.9956, 0.9954, 0.9963 \rangle$, and promotion to Negative Correlation at $t = 4$. The source sentence is: "Bifidobacteriaceae Bifidobacterium in duodenal mucosa was protective of epithelial permeability and future hospitalization for hepatic decompensation." The sentence appears in document 423 of the historical corpus.

D Annotation Agreement and NER

Table 4: CirrhoRED annotation agreement on the 0–1 scale.

Annotation	Precision	Recall	F1	Cohen's κ
Entity	0.9676	0.9780	0.9727	0.9837
Relation	0.9194	0.8865	0.9026	0.9325

Table 5: NER performance on the 0–1 scale. Correctness requires matching entity span and type. All internal configurations share the NER component.

Dataset	Precision	Recall	F1
BioRED	0.8785	0.9103	0.8941
CirrhoRED	0.7718	0.8795	0.8221

E Configuration Profiles

The GE–E2E gap varies with dataset and configuration. For Full it is 10.38 points on BioRED and 7.73 points on CirrhoRED. Figure 4 compares GE and E2E performance across configurations with the NER component held fixed.

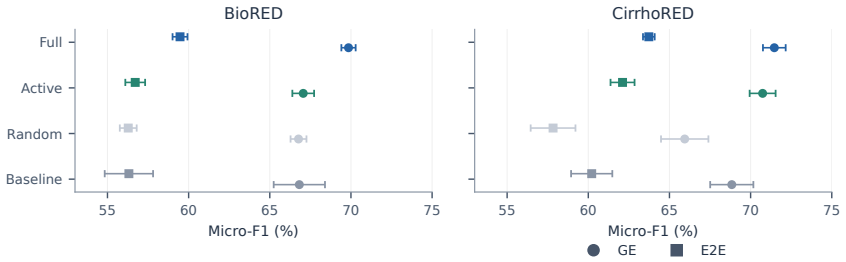


Figure 4: GE and E2E profiles for all internal configurations. Error bars indicate standard deviations across three random seeds.

F Historical Verification Analyses

We analyze acquisition and verification behavior using historical configuration H. This configuration uses an earlier corpus version and a different verifier; Table 6 summarizes its settings alongside the main experiment. Its confidence monitor operates in diagnostic-only mode.

Table 6: Settings for the main experiment and historical configuration H.

Dimension	Main study	Historical configuration H
LLM	Qwen2.5-7B-Instruct	Qwen3.5-35B-A3B
Query schedule	Five rounds, decaying budget	Four rounds of 100 candidates
Verification calls	Two instructed passes in one call	Separate evidence/verification calls; conditional no-relation query
Pool revision	Auxiliary revision in Full	Diagnostic-only; no pool mutation
CirrhoRED entities / relations	15,017 / 3,857	16,175 / 4,127
Evaluation	GE / E2E	Strict normalized concept-pair F1

F.1 Oracle decisions and request counts

Across three H runs, 1,200 selected candidates yield 245 admitted positives, 161 admitted negatives, and 794 uncertain outcomes. These correspond to 20.42%, 13.42%, and 66.17% of candidates, respectively. The total 1,877 requests comprise 1,200 first-pass, 281 second-pass, and 396 no-relation requests. Candidate count therefore differs from both supervision yield and request count.

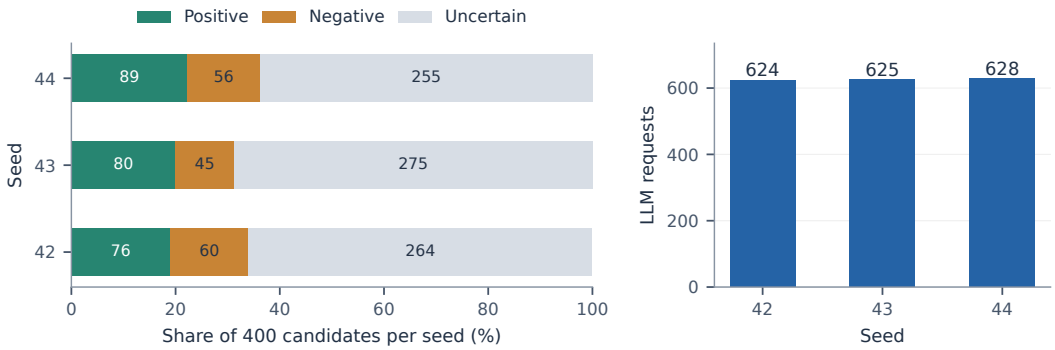


Figure 5: Oracle outcomes in historical configuration H. Segment labels show candidate counts; the horizontal scale shows percentages. The right panel shows LLM request counts.

F.2 EVIDENCE-LEVEL CASE LEDGER

Table 7 presents five automated decision paths from historical configuration H, with internal document IDs and corresponding source evidence.

Table 7: Examples of verification outcomes in historical configuration H. Quoted excerpts reproduce source text.

Pair / record	Source evidence	Verification outcome
NK cells / HBV-HCC Doc 851; seed 44; round 3	"HBV-HCC tumors have reduced levels of NK cells"	Negative correlation; second-pass Yes; positive admitted. The document also contains an NK cells–HCC annotation.
S. japonicum / hepatic fibrosis Doc 1004; seed 43; round 4	"Schistosoma japonicum eggs deposited in the liver tissue induce egg granuloma formation and liver fibrosis"	Positive correlation; second-pass Yes; positive admitted. The example involves entity aliases.
NASH / diabetes Doc 454; seed 42; round 1	NASH is introduced as a form of NAFLD; the explicit diabetes association is stated for NAFLD.	Score 0.6772; NO_EVIDENCE then NO_RELATION; negative admitted. The verifier rejects the candidate involving the narrower disease subtype.
Monocytes / endothelial cells Doc 712; seed 42; round 3	"adhered to large intrahepatic vessels through CD44 owing to increased endothelial cell adhesivity"	Proposed Interaction; extracted span omits the monocyte subject; entity-match failure; uncertain, no negative admitted.
CD4+ / CD8+ T cells Doc 434; seed 42; round 2	"HIV probably influences the immune activation of CD4+ and CD8+ T cells"	Association admitted despite speculative wording. The sentence describes HIV acting on both target cell types.

In document 352 (seed 43, round 1), the LLM reorders source sentences, failing the contiguous-span check. A subsequent query returns `RELATION_PRESENT`, leaving the candidate uncertain. This example illustrates how the additional query handles an evidence-extraction failure.

F.3 HISTORICAL DATASET QUALITY CONTROLS

For the historical corpus, we finalized 29 relation-label conflict decisions and reviewed normalization for 316 surfaces: 176 accepted, 85 rejected, and 55 unresolved. We checked entity offsets, relation endpoints, within-document IDs, and cross-split overlap using document IDs and exact-text hashes.