SECRET: Semi-supervised Clinical Trial Document Similarity Search

Trisha Das[1], Afrah Shafquat[2], Mandis Beigi[2], Jacob Aptekar[2], Jimeng Sun[1],

[1]University of Illinois Urbana-Champaign, [2]Medidata Solutions,

Abstract

Clinical trials are vital for evaluation of safety and efficacy of new treatments. However, clinical trials are resource-intensive, time-consuming and expensive to conduct, where errors in trial design, reduced efficacy, and safety events can result in significant delays, financial losses, and damage to reputation. These risks underline the importance of informed and strategic decisions in trial design to mitigate these risks and improve the chances of a successful trial. Identifying similar historical trials is critical as these trials can provide an important reference for potential pitfalls and challenges including serious adverse events, dosage inaccuracies, recruitment difficulties, patient adherence issues, etc. Addressing these challenges in trial design can lead to development of more effective study protocols with optimized patient safety and trial efficiency. In this paper, we present a novel method to identify similar historical trials by summarizing clinical trial protocols and searching for similar trials based on a query trial's protocol. Our approach significantly outperforms all baselines, achieving up to a 78% improvement in recall@1 and a 53% improvement in precision@1 over the best baseline. We also show that our method outperforms all other baselines in partial trial similarity search and zero-shot patient-trial matching, highlighting its superior utility in these tasks. Our code is publicly available at https://github.com/trishad2/SECRET.

1 Introduction

Clinical trials are vital for advancing medical interventions. However, the success of these trials is largely influenced by the quality of trial design and risk mitigation strategies (Fogel, 2018). To improve the probability of trial success, similar historical trials are used as references to inform the design of future trials (Luo et al., 2024). Historical trials can be used to determine and optimize trial design including aspects like target population, eligibility criteria, mitigation strategies, dosage schedules, and anticipation of risks and adverse events. Identifying similar trials is not a trivial task and requires investigators to search and review numerous historical protocols —a process that is labor-intensive and error-prone, often involving the manual examination of thousands of studies (Luo et al., 2024). Given the importance of historical clinical trials in optimizing trial protocols (Wang et al., 2022), it is essential to develop faster, streamlined, and more efficient trial search methods.

While advancements in data mining have improved the efficiency of similar clinical trial retrieval, most efforts have focused largely on section-level retrieval rather than comprehensive protocol-to-protocol matching (Roy et al., 2019; Rybinski et al., 2021). Trial2Vec introduced an initial clinical trial search framework for unsupervised trial similarity search (Wang and Sun, 2022). GTSLNet is a recent supervised approach trained on a private dataset of clinical trials labeled by experts (Luo et al., 2024). The main sections of a sample clinical trial protocol are listed in Table 1. In this paper, we focus solely on clinical trial protocols and refer to them as documents and use the terms "document" and "protocol" interchangeably.

The main challenges of developing a method for clinical trial search are:

- **Challenge 1: Lack of publicly available labeled data** - A significant challenge is the lack of publicly available labeled data needed to train supervised methods. GTSLNet (Luo et al., 2024) improves over Trial2Vec (Wang and Sun, 2022) on their private labeled dataset indicating the need for supervised approaches to improve accuracy and effectiveness.
- **Challenge 2: Lengthy documents** - As trial documents often exceed 1,000 words (Wang and Sun, 2022), encoding long trial documents

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5278–5291

Title	STunning in Acute Myocardial Infarction - BAS (STAMI-BAS)
Description	The objective of this trial is to examine the effect of immediate versus late administration ...
Eligibility Criteria	Inclusion Criteria: 1. Patients with STEMI who undergo primary PCI... 2. Informed consent Exclusion Criteria: 1. Killip class ≥ 3 2.Chronic kidney disease with GFR <25 ml/min/1.73 m2
Outcome	Measurement of Global longitudinal strain (GLS, %)
Disease	Myocardial Infarction
Intervention	1. Bisoprolol Oral Tablet 2. Ramipril Oral Product 3. Dapagliflozin Oral Product
...	...

Table 1: An example of clinical trial document drawn from *ClinicalTrials.gov*.

by truncating or averaging the embeddings inevitably results in poor retrieval quality. Although Trial2Vec aims to solve the long document problem by encoding different sections of the document separately, some sections of these documents (e.g., *eligibility criteria, description*, etc.) can be larger than the context length of the encoder model used in Trial2Vec leading to truncation of potentially important information in these sections. Moreover, Trial2Vec combines *eligibility criteria, description*, etc. long sections into a combined section called *context* instead of separately encoding them. This highlights the need for a clinical trial search method that can address the long document problem (which persists in existing methods including Trial2Vec) while preventing loss of critical information.

- **Challenge 3: Lack of understanding of local context** - Two medical texts can have significant word-wise overlap while describing entirely different concepts, posing a challenge for similarity computation. Trial2Vec (Wang and Sun, 2022) aims to solve this by extracting medical entities using local contrastive learning. However, two sentences can have exactly the same medical entities but have very different meanings. For example: *"The patient was tested for **insulin** levels to diagnose their **diabetes**."* and *"The patient was prescribed **insulin** to manage their **diabetes**."* both have the same medical entities ***insulin*** and ***diabetes*** but have totally different meanings. Better approaches are needed to ensure that the encoder model has an improved understanding of the local context.

- **Challenge 4: Inefficient contrastive supervision** - Though existing methods like SimCSE and Trial2Vec provide contrastive supervision mechanisms to train models with the ability to differentiate between similar (positive) and non-similar (negative) trials, the methodologies used to define 'similar' and 'non-similar' leave room for improvement. Unsupervised methods such as SimCSE (Gao et al., 2021) use instance-level contrastive learning by generating positive trial document pairs (i.e., 'entailments') by using the same trial document input twice and treating all other trial document inputs as negatives (i.e., 'contradictions'). Trial2Vec (Wang and Sun, 2022) creates positive trial document pairs by omitting sections from the trial document (see different sections in Table 1). However, this approach may result in the loss of critical information across these pairs, causing the model to identify similar trials without accounting for the missing details.

We developed SECRET, a SEmi-supervised Clinical tRial protocol similariTy searching method, to address the above-mentioned challenges. Our approach minimizes the reliance on very large publicly available labeled datasets (Challenge 1) by using labeled trial similarity data from (Wang et al., 2025) and publicly available unlabeled data in a semi-supervised manner.[1] To address the long document problem (Challenge 2), we represent a clinical trial as a set of question-answer (Q/A) pairs (generated by humans and LLMs), which significantly reduces the length of trial documents. To better capture local semantic context (Challenge 3), we train our model contrastively at the Q/A level instead of the entity level. Since the generated Q/A pairs can vary significantly in meaning depending on the original sentences they were derived from, contrastive training in SECRET ensures that sentences containing the same medical entities but different semantic meanings are assigned distinct embeddings. To tackle inefficient contrastive supervision (Challenge 4), we employ a two-level contrastive approach:

1. **Local (Q/A Level):** Contrastive training at the Q/A level ensures that the model accurately captures local context. Positive samples for each Q/A pair are automatically selected (details in Section 3).

[1] https://clinicaltrials.gov

2. **Global (Trial Level):** Contrastive training at the trial level ensures the model embeds similar trials close together and dissimilar trials far apart in the embedding space. Positive samples are generated by removing a single Q/A pair from a large section for unlabeled data, and by using similar trials from labeled data. Both hard and soft negatives are utilized to further improve performance. In contrast to existing approaches that use the same trial or drop sections from the trial document to label positive samples (methods that may be too restrictive or may lose critical information needed for trial similarity), our approach minimizes the loss of critical information by summarizing in Q/A pairs while adding flexibility to the search.

This paper compares SECRET to baseline methods, showing it outperforms them in trial document similarity search while using less than a quarter of the training data required by Trial2Vec. It also achieved superior scores in query-to-trial matching and zero-shot patient-to-trial matching tasks. The paper discusses (i) related methods for clinical trial search in Section 2, (ii) detailed method details of SECRET in Section 3, (iii) experimental setting and results in Section 4, (iv) conclusion in Section 5, and (v) limitations and potential future directions in Section 6.

2 Related Work

2.1 Text and Document Retrieval

2.1.1 General Text

Dense retrieval methods using distributional word representations, such as Word2Vec (Mikolov, 2013) and GloVe (Pennington et al., 2014) became popular due to their superior performance in capturing semantic similarity compared to traditional methods such as TF-IDF (Salton and Buckley, 1988). In contrast, early information retrieval approaches relied heavily on manual feature engineering (Trotman et al., 2014; Yang et al., 2017). The rise of deep learning models, especially contextualized encoders like BERT (Devlin, 2018), has driven significant advancements in neural retrieval methods (Van Gysel et al., 2016; Dehghani et al., 2017; Yates et al., 2021).

In domains like clinical trials where access to labeled data is limited due to cost, privacy, and other reasons (Das et al., 2023, 2024), zero-shot learning models become a necessity. Although some approaches have attempted to improve retrieval quality by performing post-processing on pre-trained BERT embeddings (Li et al., 2021), their performance remains suboptimal without domain-specific training. While BERT-like models fine-tuned or pre-trained on biomedical documents and electronic health records, such as BioBERT (Lee et al., 2020) and Bio_ClinicalBERT (Alsentzer et al., 2019) exist, they are not trained for clinical trial retrieval, resulting in suboptimal performance.

2.1.2 Clinical Trial

Traditional clinical trial query engines use rule-based entity matching on trial metadata which relies heavily on databases hence limiting their flexibility (Tasneem et al., 2012; Tsatsaronis et al., 2012; Jiang and Weng, 2014; Park et al., 2020). Recent approaches utilize supervised neural ranking to match trial titles or relevant segments with user queries (Roy et al., 2019; Rybinski et al., 2021). However, these methods are limited to specific parts of the trial documents. PRISM (Gupta et al., 2024) is a recent patient-to-trial matching model that transforms the trial criteria from clinicaltrials.gov into simplified, independent questions, each with answers like "Yes," "No," or "NA." However, we utilize an LLM to generate Q/A pairs from eligibility criteria by extracting key information and the answers are not limited to "Yes," "No," or "NA." Trial2Vec (Wang and Sun, 2022) is a self-supervised method that encodes entire trial documents, enabling searches based on trial-level similarity and query-based searches. GT-SLNet (Luo et al., 2024), a supervised approach designed to identify similarity at the trial level, outperforms Trial2Vec and other unsupervised and self-supervised approaches in retrieval tasks, but requires large labeled datasets to avoid overfitting (Althnian et al., 2021). Self-supervised methods avoid manual annotation by using unlabeled data, though they typically deliver lower performance (Luo et al., 2024). To our knowledge, SECRET introduces the first semi-supervised approach for trial retrieval that balances between the drawbacks of supervised and unsupervised methods and shows improved performance in trial-level retrieval scores.

2.2 Text Contrastive Learning

Contrastive learning has recently become a widely discussed topic (Chen et al., 2020; Chen and He, 2021; Carlsson et al., 2021; Gao et al., 2021; Aberdam et al., 2021; Yang et al., 2022; Zhang et al.,

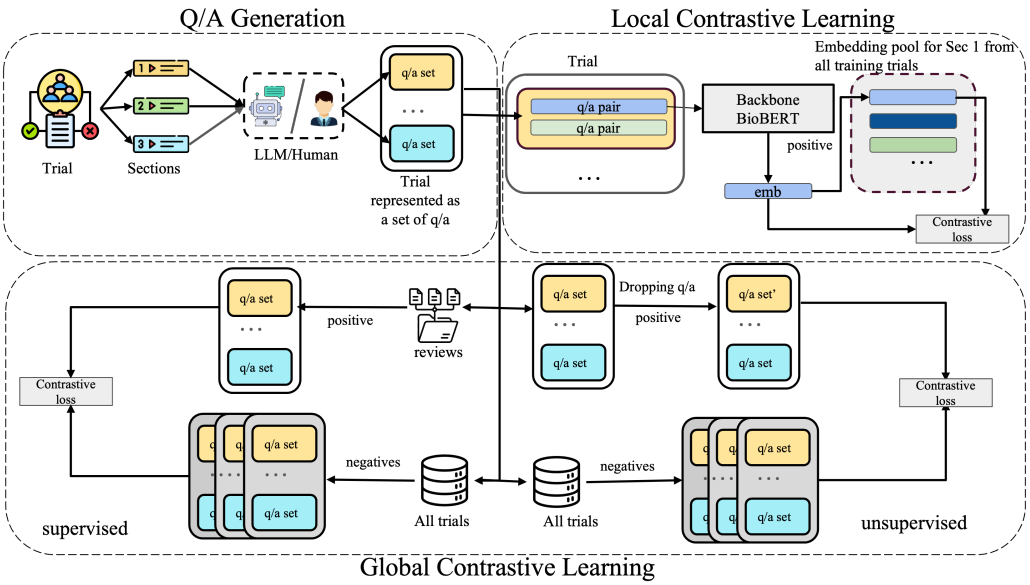


Figure 1: Overview of SECRET. SECRET consists of three main components: Q/A Generation, Local Contrastive Learning and Global Contrastive Learning. Without loss of generality, we illustrate Sec 1 in the Local Contrastive Learning block, applicable to all sections.

2020; Wang et al., 2020). This technique has been used for zero-shot retrieval capabilities (Wang et al., 2020; Zhang et al., 2020) and to enhance downstream NLP tasks, such as text classification (Pan et al., 2022; Chen et al., 2022). However, current approaches focus on improving sentence embeddings by manipulating text alone, making them less effective for handling long clinical trial-specific documents. Trial2Vec addresses this limitation by generating document embeddings, providing a better solution for such scenarios.

3 Proposed Method

In this section, we detail the architecture of SECRET (Figure 1). SECRET is built on three main components: (1) Q/A generation, (2) Local contrastive learning, and (3) Global contrastive learning. First, we generate key Q/A pairs for each trial using LLM. Next, we fine-tune the BioBERT (Lee et al., 2020) backbone encoder using local and global contrastive learning to generate the final embeddings.

3.1 Q/A Generation

We represent a clinical trial with a set of key Q/A pairs generated from different sections of a clinical trial document. The method assumes that two similar documents will have a similar set of key Q/A pairs. This helps reduce the length of the trial documents and better capture the local context (see 3.2). We utilize the Llama-3.1-8B-Instruct model to generate Q/A pairs from large sections (e.g. *eligibility criteria*, etc.) that can follow user instructions effectively. For smaller sections (e.g., *title, disease, intervention*, etc.), we use predefined questions generated by humans. An example of a clinical trial represented by a set of Q/A pairs (Figure 7) and the prompt used to generate them can be found in the Appendix A.1.

3.2 Local Contrastive Learning

To enhance SECRET's discriminative power and improve its understanding of local context, we perform contrastive training at the Q/A level. Previous methods such as Trial2Vec rely on entity similarity to compare sentences or documents which could lead to incorrect assumptions when sentences share entities but differ in meaning (Challenge 3). By focusing on Q/A pairs, SECRET effectively captures these nuances. This also helps in partial trial matching to find the best matching trial documents given the *title* of the query trial (see results in Section 4.5). Similar to Trial2Vec, BioBERT is used as the backbone encoder for SECRET.

We finetune the BioBERT backbone at the Q/A level, where positive and negative samples are automatically selected from the training pool. Given a training pair (Q_i, A_i), the most similar pair (Q_p, A_p) within the same section's Q/A pool is selected as the positive sample. Let v_i, v_i^+ denote the embeddings from the backbone (before finetuning) of the anchor and positive Q/A pairs, respectively.

$$v_i^+ = \underset{v_j \neq v_i}{\text{argmax}}, \psi(v_i, v_j), \quad v_j, v_i \in \mathcal{P}, \quad (1)$$

where $\mathcal{P}$ represents the pool of Q/A pairs from the same section as v_i across all trials. A similarity function $\psi(x, y)$ computes the cosine similarity between the embeddings of the Q/A pairs. All other pairs in the batch, excluding the selected positive pair, serve as negative samples. InfoNCE loss for local contrastive learning is defined as:

$$\mathcal{L}_{\text{l}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\psi(v_i,v_i^+)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\psi(v_i,v_j)/\tau\right)}. \quad (2)$$

In the equation, v_i and v_i^+ represent the embedding for the i-th anchor Q/A pair in the batch and the embedding for its corresponding positive Q/A pair associated with the query v_i, respectively. v_j refers to all Q/A pairs in the batch. τ refers to the temperature scaling parameter. N represents number of Q/A pairs in the batch. An example (anchor, positive) pair:

```
('What is the age range for eligible
children? 6-12 years',
'What is the age range for enrollment?
6-12 years')
```

Here, the anchor Q/A pair is from trial NCT00000113 and the positive Q/A pair is from trial NCT00006565.

3.3 Global Contrastive Learning

In global contrastive learning, each trial is represented as a set of Q/A pairs, where the entire set is used as input. The objective of global contrastive training is to learn trial-level embeddings such that positive trials with similar patterns in their Q/A pairs (e.g., similar trials or modified versions of the same trial) are pulled closer in the embedding space and negative trials (e.g., unrelated or hard negatives) with dissimilar patterns are pushed apart.

We utilize both labeled and unlabeled trials for contrastive training. For unlabeled trials, a positive sample for an anchor trial T_i is created by dropping one Q/A pair from a section with multiple Q/A pairs of that trial, resulting in T_i^+. Hard negatives are chosen as trials that share the same disease indication (similar approach as Trial2Vec) as T_i in the entire dataset, while other trials in the batch serve as additional negatives. A negative trial for anchor T_i is denoted as T_i^-. For labeled trials, the positive trial T_i^+ is explicitly selected based on ground truth (provided label), representing a known similar trial. Hard negatives and other negatives are constructed as in the unsupervised setting.

$$\mathcal{L}_{\text{paired}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\psi(z_i,z_i^+)/\tau\right)}{Z_{\text{p}}}, \quad (3)$$

$$Z_{\text{P}} = \exp\left(\psi(z_i,z_i^+)/\tau\right) + \exp\left(\psi(z_i,z_i^-)/\tau\right),$$

$$\mathcal{L}_{\text{in-batch}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\psi(z_i,z_i^+)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\psi(z_i,z_j)/\tau\right)}, \quad (4)$$

$$\mathcal{L}_{\text{g}} = \mathcal{L}_{\text{paired}} + \mathcal{L}_{\text{in-batch}}. \quad (5)$$

In the equations for global contrastive learning, the following notations are used. z_i represents the query trial embedding for the i-th trial, and z_i^+ is its positive trial's embedding. In the case of the paired loss (Eq. 3), z_i^- refers to an explicit negative trial embedding corresponding to z_i. τ is the temperature scaling parameter. Z_{p} is the normalization term. For in-batch loss (Eq. 4), the denominator is the sum over all trials in the batch, where z_j represents all trial embeddings in the batch. N denotes batch size. $\psi()$ computes cosine similarity.

Finally, we use cosine similarity on the trial embeddings to rank clinical trials given a query trial.

4 Experiments and Results

4.1 Datasets and Setup

To retrieve similar trial documents from full or partial query trials, we trained the model at the global level using around 10,000 labeled trials (Wang et al., 2025) and 60,000 unlabeled trials downloaded from https://aact.ctti-clinicaltrials.org. This is less than one-fourth of the training data used in Trial2Vec. All datasets are in English. For each review paper utilized in (Wang et al., 2025), a set of trials is identified as similar based on shared characteristics, such as diseases, interventions, population and outcomes, sourced from Cochrane Reviews.[2] From the remaining labeled trials, we prepared validation and test sets. Each query trial in these sets has 10 corresponding trials, labeled as 'relevant' (i.e., positive) or 'not relevant' (i.e., negative). Relevant trials were identified from review data, while negative trials were chosen from the same disease category but excluded from the training data and relevant set, with random sampling used if no such trials were available. The test set has 1,420 pairs, and the validation set has 2,000 pairs.

[2] https://www.cochranelibrary.com

	precision@1	recall@1	precision@2	recall@2	precision@5	recall@5	nDCG@5	MAP
TF-IDF	0.363 ± 0.073	0.244 ± 0.054	0.298 ± 0.048	0.388 ± 0.064	0.217 ± 0.023	0.687 ± 0.067	0.522 ± 0.055	0.501 ± 0.050
BM25	0.334 ± 0.071	0.223 ± 0.055	0.271 ± 0.050	0.350 ± 0.066	0.180 ± 0.026	0.567 ± 0.079	0.454 ± 0.065	0.471 ± 0.052
Word2Vec	0.293 ± 0.071	0.184 ± 0.047	0.266 ± 0.049	0.328 ± 0.061	0.191 ± 0.024	0.573 ± 0.067	0.435 ± 0.058	0.435 ± 0.049
BERT	0.241 ± 0.067	0.130 ± 0.040	0.235 ± 0.046	0.279 ± 0.064	0.197 ± 0.024	0.591 ± 0.064	0.415 ± 0.052	0.400 ± 0.043
BioBERT	0.347 ± 0.078	0.202 ± 0.053	0.275 ± 0.051	0.313 ± 0.068	0.202 ± 0.022	0.612 ± 0.061	0.462 ± 0.057	0.450 ± 0.051
Bio_ClinicalBERT	0.280 ± 0.071	0.169 ± 0.048	0.261 ± 0.050	0.317 ± 0.054	0.207 ± 0.024	0.644 ± 0.058	0.464 ± 0.050	0.437 ± 0.046
Longformer	0.253 ± 0.070	0.158 ± 0.049	0.272 ± 0.050	0.338 ± 0.065	0.198 ± 0.025	0.631 ± 0.067	0.447 ± 0.056	0.425 ± 0.048
Clinical-Longformer	0.206 ± 0.065	0.123 ± 0.042	0.211 ± 0.050	0.252 ± 0.066	0.168 ± 0.021	0.533 ± 0.065	0.369 ± 0.051	0.369 ± 0.045
IDCM	0.156 ± 0.054	0.102 ± 0.040	0.167 ± 0.039	0.206 ± 0.056	0.132 ± 0.019	0.391 ± 0.059	0.286 ± 0.048	0.324 ± 0.039
Trial2Vec	0.422 ± 0.078	0.263 ± 0.054	0.375 ± 0.060	0.458 ± 0.068	0.227 ± 0.029	0.689 ± 0.067	0.553 ± 0.064	0.539 ± 0.058
SECRET	**0.647 ± 0.077**	**0.467 ± 0.063**	**0.508 ± 0.046**	**0.682 ± 0.061**	**0.297 ± 0.023**	**0.924 ± 0.034**	**0.796 ± 0.042**	**0.754 ± 0.044**

Table 2: Performance evaluation of retrieval models for complete trial similarity search on the labeled test set. The table presents precision, recall, nDCG, and MAP metrics, reported as mean ± standard deviation, with the best values highlighted in **bold**.

	precision@1	recall@1	precision@2	recall@2	precision@5	recall@5	nDCG@5	MAP
TF-IDF	0.359 ± 0.078	0.252 ± 0.058	0.320 ± 0.055	0.410 ± 0.062	0.214 ± 0.023	0.664 ± 0.054	0.517 ± 0.054	0.505 ± 0.052
BM25	0.363 ± 0.074	0.248 ± 0.052	0.298 ± 0.050	0.390 ± 0.060	0.199 ± 0.024	0.627 ± 0.064	0.493 ± 0.053	0.491 ± 0.048
Word2Vec	0.308 ± 0.079	0.203 ± 0.055	0.237 ± 0.050	0.298 ± 0.065	0.190 ± 0.024	0.595 ± 0.072	0.447 ± 0.062	0.442 ± 0.053
BERT	0.233 ± 0.057	0.137 ± 0.037	0.204 ± 0.042	0.241 ± 0.048	0.179 ± 0.022	0.565 ± 0.064	0.392 ± 0.046	0.385 ± 0.036
BioBERT	0.288 ± 0.078	0.191 ± 0.059	0.244 ± 0.055	0.318 ± 0.079	0.192 ± 0.024	0.601 ± 0.074	0.444 ± 0.065	0.439 ± 0.056
Bio_ClinicalBERT	0.257 ± 0.083	0.172 ± 0.063	0.203 ± 0.047	0.276 ± 0.069	0.180 ± 0.023	0.578 ± 0.069	0.416 ± 0.059	0.413 ± 0.051
Longformer	0.235 ± 0.067	0.163 ± 0.049	0.199 ± 0.046	0.264 ± 0.059	0.156 ± 0.022	0.490 ± 0.067	0.364 ± 0.054	0.385 ± 0.044
Clinical-Longformer	0.238 ± 0.066	0.160 ± 0.045	0.221 ± 0.045	0.303 ± 0.056	0.181 ± 0.025	0.573 ± 0.063	0.413 ± 0.050	0.412 ± 0.041
IDCM	0.306 ± 0.073	0.213 ± 0.055	0.273 ± 0.049	0.369 ± 0.070	0.180 ± 0.024	0.584 ± 0.079	0.452 ± 0.064	0.462 ± 0.051
Trial2Vec	0.456 ± 0.088	0.322 ± 0.065	0.370 ± 0.057	0.482 ± 0.068	0.228 ± 0.027	0.717 ± 0.066	0.592 ± 0.062	0.579 ± 0.056
SECRET	**0.548 ± 0.083**	**0.390 ± 0.066**	**0.465 ± 0.044**	**0.623 ± 0.059**	**0.289 ± 0.023**	**0.902 ± 0.040**	**0.745 ± 0.044**	**0.696 ± 0.047**

Table 3: Performance evaluation of retrieval models for query-to-trial matching (partial trial similarity search) on the labeled test set. Metrics include precision, recall, nDCG, and MAP, reported as mean ± standard deviation. Best-performing results are highlighted in **bold**.

We also ran experiments to evaluate how SECRET performs zero-shot on patient-to-trial matching. For this task, we used 75 patients from the TREC2021 dataset.[3] The dataset has ground-truth labels that indicate each patient's best match to a trial. We prepared a test set where for each patient, there are 10 trials with ground truth labels (both relevant and not relevant trials) resulting in 731 unique test trials for this task.

We evaluated performance using precision@k, recall@k, nDCG@5 and MAP where k can be 1, 2, or 5. Details about the metrics are available in the Appendix A.3.

4.2 Implementation Details

We used a large language models (LLM) to generate Q/A for *eligibility criteria* which is the largest section among all sections we use. For all the other sections (e.g., *title, disease, intervention, keywords, outcome*), we use predefined questions. We conducted local contrastive learning for 10 epochs with a batch size of 32. Then, we fine-tuned the model at the trial level for an additional 10 epochs, varying the batch sizes (16 and 32) to identify the configuration that yielded the highest validation scores. The validation results, shown in Figure 4 in Appendix A, indicate that a batch size of 16 for SECRET achieved the best performance. For both local and global contrastive learning, we selected the best model based on validation scores across all epochs. For optimization, we used a learning rate of 2e-5 for local and 1e-6 for global contrastive training, employing the AdamW optimizer (Loshchilov, 2017). We leveraged mixed precision during training, which reduced the computational resources required and accelerated the training process. We set τ to the default value (0.1) used in the implementation of InfoNCE loss (Oord et al., 2018). [4] The experiments were carried out using 2 RTX 6000 GPUs. We bootstrapped 50 samples for 100 iterations, then calculated the average score and standard deviation.

4.3 Baselines

Due to the lack of ground truth labels for most training samples, we focus on unsupervised and self-supervised baselines for retrieval (similar to Trial2Vec). The baselines include TF-IDF (Salton and Buckley, 1988), BM25 (Trotman et al., 2014), Word2Vec (Mikolov, 2013), BERT (Devlin, 2018), BioBERT (Lee et al., 2020), Bio_ClinicalBERT

[3] http://www.trec-cds.org/2021.html

[4] https://pypi.org/project/info-nce-pytorch/

	precision@1	recall@1	precision@2	recall@2	precision@5	recall@5	nDCG@5	MAP
TF-IDF	0.494 ± 0.064	0.097 ± 0.012	0.529 ± 0.043	0.217 ± 0.020	0.536 ± 0.029	0.546 ± 0.033	0.541 ± 0.030	0.641 ± 0.023
BM25	0.527 ± 0.065	0.112 ± 0.016	0.530 ± 0.043	0.218 ± 0.019	0.493 ± 0.023	0.504 ± 0.023	0.514 ± 0.025	0.623 ± 0.020
Word2Vec	0.523 ± 0.070	0.102 ± 0.014	0.551 ± 0.056	0.218 ± 0.022	0.546 ± 0.028	0.548 ± 0.027	0.548 ± 0.032	0.642 ± 0.026
BERT	0.483 ± 0.066	0.095 ± 0.013	0.519 ± 0.047	0.218 ± 0.020	0.513 ± 0.032	0.520 ± 0.030	0.516 ± 0.033	0.619 ± 0.026
BioBERT	0.610 ± 0.067	0.125 ± 0.015	0.574 ± 0.044	0.244 ± 0.022	0.563 ± 0.022	0.582 ± 0.024	0.586 ± 0.024	0.659 ± 0.019
Bio_ClinicalBERT	0.565 ± 0.074	0.122 ± 0.020	0.601 ± 0.052	0.251 ± 0.024	0.552 ± 0.027	0.570 ± 0.027	0.575 ± 0.030	0.659 ± 0.026
Longformer	0.534 ± 0.079	0.106 ± 0.016	0.532 ± 0.049	0.210 ± 0.019	0.539 ± 0.025	0.546 ± 0.024	0.545 ± 0.028	0.628 ± 0.023
Clinical-Longformer	0.460 ± 0.069	0.095 ± 0.015	0.499 ± 0.047	0.200 ± 0.019	0.479 ± 0.029	0.487 ± 0.031	0.487 ± 0.031	0.598 ± 0.022
IDCM	0.353 ± 0.071	0.068 ± 0.014	0.443 ± 0.052	0.178 ± 0.021	0.516 ± 0.026	0.521 ± 0.024	0.493 ± 0.029	0.586 ± 0.023
Trial2Vec	0.608 ± 0.071	0.129 ± 0.019	0.598 ± 0.051	0.250 ± 0.022	0.602 ± 0.031	0.616 ± 0.026	0.618 ± 0.031	0.695 ± 0.025
SECRET	**0.710 ± 0.073**	**0.158 ± 0.021**	**0.682 ± 0.057**	**0.292 ± 0.027**	**0.627 ± 0.034**	**0.641 ± 0.031**	**0.666 ± 0.036**	**0.744 ± 0.029**

Table 4: Performance evaluation of retrieval models for patient-to-trial matching on a subset of the TREC2021 labeled test set. The table presents precision, recall, nDCG, and MAP metrics, reported as mean ± standard deviation, with the best values highlighted in **bold**.

(Alsentzer et al., 2019), Longformer (Beltagy et al., 2020), Clinical_Longformer (Li et al., 2022), IDCM (Hofstätter et al., 2021), and Trial2Vec (Wang and Sun, 2022). IDCM, Longformer, and Clinical_Longformer are designed for long documents, while TF-IDF, BM25, Word2Vec, BERT, and Longformer are general retrieval methods. BioBERT, Bio_ClinicalBERT, and Clinical_Longformer are tailored to the biomedical and clinical domains. Among all baselines, only Trial2Vec is specifically trained to retrieve clinical trials based on protocol similarity and we use their precomputed trial embeddings. For SECRET, we consider a trial to be a set of Q/A pairs from different sections. For other baselines, we concatenate the full text of these sections.

4.4 Complete Trial Similarity Search

Given the protocol of a query trial, we evaluate the performance across models to retrieve trials with similar protocols. As shown in Table 2, SECRET outperforms all baselines by a significant margin, achieving up to 78% improvement in recall@1 and 53% improvement in precision@1 over the best baseline. Improvements are also seen in other metrics, with SECRET surpassing the best baseline by around 30%-40%. The precision and recall gaps between SECRET and the baselines are larger when k is small. As k increases, precision@k decreases for all methods due to the increased chance of selecting dissimilar trials the more trials are selected. It is important to note that there is a limited number of positive pairs (1.32 trials on average) relative to the 10 candidate trials (Eq. 6). Recall@k improves with larger k because it allows retrieval of more relevant items, increasing the proportion of relevant items in the retrieved set.

4.5 Partial Trial Similarity Search

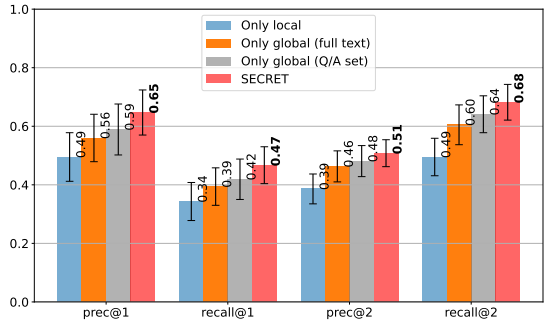
We evaluated performance across the models on the partial trial retrieval scenario, where users aim to find similar trials based on short or incomplete descriptions (partial attributes). We utilize *title* as a partial attribute. As shown in Table 3, SECRET outperforms all baselines by a substantial margin, achieving up to a 29% improvement in recall@2 over the best baseline. Furthermore, SECRET shows improvements in other metrics that surpass the best baseline by 20%-27%. The evaluation in Figure 5 in the Appendix A.6 shows that combining the title with additional sections consistently improves both precision@1 and recall@1 compared to the title-only approach. Combining the title with intervention achieves the highest scores.

4.6 Patient-to-trial Similarity Search

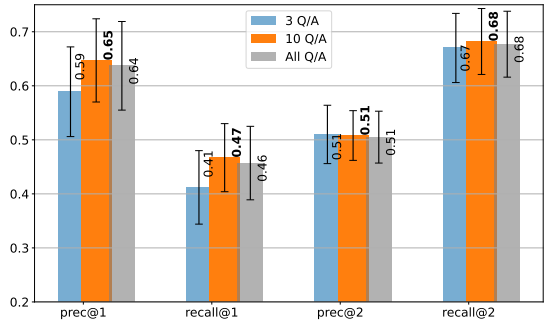
We evaluated the performance of SECRET in zero-shot patient-to-trial matching, where each patient is represented by a clinical note (patient summary). The patient and trial embeddings are generated using SECRET and cosine similarity between embeddings is used to rank the trials. We do not train SECRET for patient-to-trial matching. As shown in Table 4, SECRET outperforms all baselines. It outperforms the best baseline with improvements of up to 22% in recall and 17% in precision.

4.7 Ablation Studies

We conducted ablation studies by removing either local or global contrastive training from SECRET. As shown in Figure 2a, the *only local* approach yields the lowest performance, followed by two *only global* approaches, which improve over *only local* approach. Representing trials with Q/A pairs (*only global (Q/A set)*) achieved better scores than using the full text of clinical trial protocol sec-



(a) Ablation results by dropping different levels of contrastive learning.



(b) Comparison of performance metrics across different Q/A counts.

Figure 2: Ablation results

	Query Trial	Trial2Vec Result	SECRET Result
NCTID	NCT00061594	NCT03470103	NCT00433017
Title	A Study to Compare rhu-Fab V2 With Verteporfin ... Macular Degeneration.	A Study in Patients With Wet Age-related Macular Degeneration ...	Verteporfin Photodynamic Therapy ... Age-related Macular Degeneration.
Intervention	Ranibizumab	Eylea (Aflibercept, VEGF Trap-Eye, BAY86-5321)	Verteporfin Photodynamic Therapy, Ranibizumab
Disease	Macular Degeneration	Macular Degeneration	Macular Degeneration
Age	≥50 years	≥18 years	≥50 years

Table 5: Case study comparing the retrieval performance of the SECRET and Trial2Vec.

tions (*only global (full text)*). SECRET, which combines both local and global training, outperforms both methods. Local training focuses on finer, context-specific details, while global training captures broader contextual information. Taking into account both the finer details and the broader context, SECRET outperforms both individual approaches. A table showing results across all metrics can be found in the Appendix (Table 8).

We have also experimented with the impact of the number of Q/A pairs on retrieval scores. Although the system prompt for the LLM instructed to generate 3-10 question-answer pairs, we observed that the model produced more than 10 pairs at times (see Figure 3 in Appendix A). We selected top 3, 10, and all question-answer pairs generated by the LLM. We then used these selected pairs, along with predefined Q/A pairs for small sections, to separately train our model. Using *'10 Q/A pairs'* achieved the best overall performance, outperforming both the *'all Q/A pairs'* and *'3 Q/A pairs'* settings in precision@1 and recall@1, while achieving comparable scores in precision@2 and recall@2. This indicates the importance of selecting the optimal number of Q/A pairs where too many Q/A may obscure critical information with irrelevant details, while too few Q/A may result in missing essential information (Figure 2b).

Additional ablation results for using *predefined Q/A only* and *answers only* can be found in Table 8 in the Appendix, which show that LLM-generated Q/A from large sections and full Q/A pairs (rather than just answers) are necessary for achieving better scores. We also evaluated BERT and BioBERT encoders to assess the impact of using Q/A pairs for representing trials, finding consistent improvements over full text (Figures 6, 7 in Appendix A.7).

4.8 Case Study

We conducted a qualitative analysis of the similarity search results (Table 5). The top-1 relevant clinical trial retrieved by SECRET is more closely aligned with the query trial and matches the ground truth label. In contrast, the top-1 trial retrieved by Trial2Vec shares the same disease as the query trial but differs in other details such as interventions and age. Due to the enhanced local context understanding provided by contrastive learning at the question-answer level in SECRET, the **Age** attribute (a common eligibility criterion) in the query trial and the top-1 trial retrieved by SECRET is identical. An additional case study can be found in the

Appendix (Table 6).

5 Conclusion

Efficient retrieval of similar clinical trials is critical for optimal design and evaluation of clinical trials, yet the process remains labor-intensive, inefficient and time-consuming. To address this, we developed SECRET, a semi-supervised clinical trial document similarity search method that reduces reliance on large labeled datasets, resolves the long document problem by a novel representation of trials as question-answer (Q/A) pairs, and captures both local and global semantic contexts through Q/A-level and trial-level contrastive training. Our approach outperforms existing baselines, including Trial2Vec, while requiring significantly less training data. SECRET achieves superior results in complete trial search, partial trial search, and zero-shot patient-to-trial matching tasks. In summary, SECRET offers an efficient solution for clinical trial retrieval, setting a new standard in the field of long document retrieval.

6 Limitations

The current investigation limited evaluations of trial protocols to the following sections: *title, disease, intervention, keywords, outcome* and *eligibility criteria*. We excluded *description and study design* sections from consideration, which are also lengthy components of the protocol. Although, the exclusion was due to resource constraints in using large language models (LLMs), we hypothesized that the included sections would be sufficient to differentiate between trials. Other related trial documents, such as informed consent forms and adverse event reports, were not included in these experiments. To avoid the complexity of parsing PDFs and the limited accessibility of these documents for many trials, we focused exclusively on trial protocols. Also, SECRET 's performance depends on LLM-generated questions and other factors like the choice of LLM, system prompts, etc. Despite these limitations, our approach outperforms the baseline methods. Future work may explore the utility of addition of trial metadata (e.g., forms, fields collected), other related documents related to clinical trials and medical domain knowledge for trial similarity search. We also want to investigate the importance of different sections in the trial search and improve our method on that basis.

References

Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R Manmatha, and Pietro Perona. 2021. Sequence-to-sequence contrastive learning for text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15302–15312.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou Elwafa, and Heba Kurdi. 2021. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Applied Sciences*, 11(2):796.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Fredrik Carlsson, Evangelia Gogoulou, Erik Ylipää, Amaru Cuba Gyllensten, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations, 2021*.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2022. Contrastnet: A contrastive learning framework for few-shot text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10492–10500.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Trisha Das, Zifeng Wang, Afrah Shafquat, Mandis Beigi, Jason Mezey, and Jimeng Sun. 2024. Synrl: Aligning synthetic clinical trial data with human-preferred clinical endpoints using reinforcement learning. *arXiv preprint arXiv:2411.07317*.

Trisha Das, Zifeng Wang, and Jimeng Sun. 2023. Twin: Personalized clinical trial digital twin generation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 402–413.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 65–74.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

David B Fogel. 2018. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemporary clinical trials communications*, 11:156–164.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Shashi Gupta, Aditya Basu, Mauro Nievas, Jerrin Thomas, Nathan Wolfrath, Adhitya Ramamurthi, Bradley Taylor, Anai N Kothari, Regina Schwind, Therica M Miller, et al. 2024. Prism: Patient records interpretation for semantic clinical trial matching system using large language models. *npj Digital Medicine*, 7(1):305.

Sebastian Hofstätter, Bhaskar Mitra, Hamed Zamani, Nick Craswell, and Allan Hanbury. 2021. Intra-document cascading: Learning to select passages for neural document ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1349–1358.

Silis Y Jiang and Chunhua Weng. 2014. Cross-system evaluation of clinical trial search engines. *AMIA Summits on Translational Science Proceedings*, 2014:223.

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.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Junyu Luo, Cheng Qian, Lucas Glass, and Fenglong Ma. 2024. Clinical trial retrieval via multi-grained similarity learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2950–2954.

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11130–11138.

Junseok Park, Seongkuk Park, Kwangmin Kim, Woochang Hwang, Sunyong Yoo, Gwan-su Yi, and Doheon Lee. 2020. An interactive retrieval system for clinical trial studies with context-dependent protocol elements. *PloS one*, 15(9):e0238290.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Soumyadeep Roy, Koustav Rudra, Nikhil Agrawal, Shamik Sural, and Niloy Ganguly. 2019. Towards an aspect-based ranking model for clinical trial search. In *International Conference on Computational Data and Social Networks*, pages 209–222. Springer.

Maciej Rybinski, Sarvnaz Karimi, and Aleney Khoo. 2021. Science2cure: A clinical trial search prototype. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2620–2624.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Asba Tasneem, Laura Aberle, Hari Ananth, Swati Chakraborty, Karen Chiswell, Brian J McCourt, and Ricardo Pietrobon. 2012. The database for aggregate analysis of clinicaltrials. gov (aact) and subsequent regrouping by clinical specialty. *PloS one*, 7(3):e33677.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, pages 58–65.

George Tsatsaronis, Konstantinos Mourtzoukos, Vassiliki Andronikou, Tassos Tagaris, Iraklis Varlamis, Michael Schroeder, Theodora Varvarigou, Dimitris Koutsouris, and Nikolaos Matskanis. 2012. Ponte: a context-aware approach for automated clinical trial protocol design. In *proceedings of the 6th International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases in conjunction with VLDB*.

Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2016. Learning latent vector spaces for product search. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 165–174.

Shuohang Wang, Yuwei Fang, Siqi Sun, Zhe Gan, Yu Cheng, Jing Jiang, and Jingjing Liu. 2020. Cross-thought for sentence encoder pre-training. *arXiv preprint arXiv:2010.03652*.

Zifeng Wang, Lang Cao, Qiao Jin, Joey Chan, Nicholas Wan, Behdad Afzali, Hyun-Jin Cho, Chang-In Choi, Mehdi Emamverdi, Manjot K. Gill, Sun-Hyung Kim, Yijia Li, Yi Liu, Hanley Ong, Justin Rousseau, Irfan Sheikh, Jenny J. Wei, Ziyang Xu, Christopher M. Zallek, Kyungsang Kim, Yifan Peng, Zhiyong Lu, and Jimeng Sun. 2025. A foundation model for human-ai collaboration in medical literature mining. *Preprint*, arXiv:2501.16255.

Zifeng Wang, Chufan Gao, Lucas M Glass, and Jimeng Sun. 2022. Artificial intelligence for in silico clinical trials: A review. *arXiv preprint arXiv:2209.09023*.

Zifeng Wang and Jimeng Sun. 2022. Trial2vec: Zero-shot clinical trial document similarity search using self-supervision. *arXiv preprint arXiv:2206.14719*.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. 2022. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1253–1256.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.

Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. 2020. An unsupervised sentence embedding method by mutual information maximization. *arXiv preprint arXiv:2009.12061*.

A Appendix

A.1 Details of Q/A Generated by LLM

We used LLMs to process the *eligibility criteria* section, which is typically a lengthy section (compared to other sections like *title, disease, intervention, outcome,* etc.) and includes both inclusion and exclusion criteria of a trial. The average number of words in the protocol is 312.77 whereas, the average number of words in the set of Q/A pairs is 254.05. Since the length of this section can vary across trials, we prompted the LLM to generate 3-10 Q/A pairs based on the section's length. However, the LLM occasionally generated more than 10 pairs. Figure 3 presents a histogram that illustrates the distribution of the number of Q/A pairs generated by the LLM.

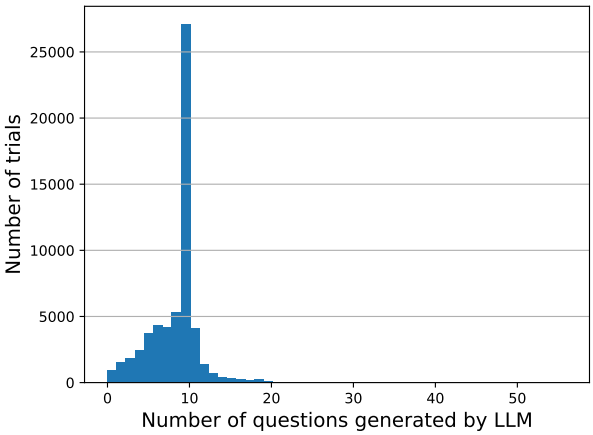


Figure 3: Distribution of Q/A Count per Trial via LLM

The prompt we used to generate Q/A pairs is as follows:

"You are an expert at creating key questions from a medical text and extracting the answers from the text. Extract 3-10 Q/A pairs without repetitions of key entities in the Q/As. Avoid general questions like 'What are the exclusion criteria?' Make sure that an answer is no more than 5 tokens/words. Output only json-formated Q/A pairs like this: {'Question': 'question1', 'Answer': 'answer1'} {'Question': 'question2' , 'Answer': 'answer2'}
...
Input:"

A.2 Hyperparameter Tuning Results

We performed hyperparameter tuning for the number of epochs and batch size. After each training epoch, the model was evaluated on the validation set and the best-performing model was saved. For global contrastive learning, we experimented with two batch sizes (16 and 32). Figure 4 illustrates the validation recall scores for these batch sizes. Since the validation scores for batch size 16 were higher than those for batch size 32, we selected 16 as the optimal batch size.

A.3 Metrics

These are the metrics we used for evaluation in current work:

1. **precision@**k measures how many of the top

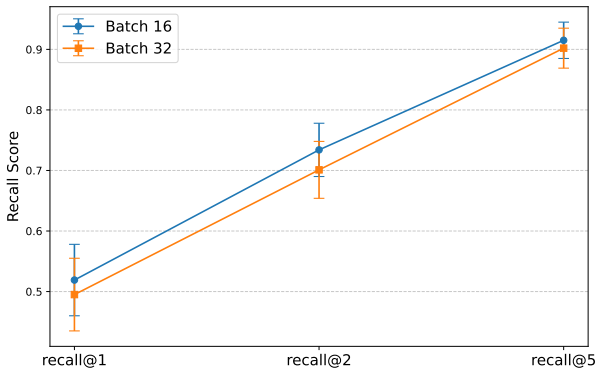


Figure 4: Effect of batch size on validation scores.

k items in a ranked list are relevant to a query.

$$\text{precision@k} = \frac{\text{\# of relevant items in top } k}{k}, \tag{6}$$

2. **recall@**k measures the proportion of relevant items that are successfully retrieved in the top k items.

$$\text{recall@k} = \frac{\text{\# of relevant items in top } k}{\text{total number of relevant items}}. \tag{7}$$

3. **nDCG** (normalized discounted cumulative gain) is a ranking evaluation metric that measures the quality of a ranked list by considering both the relevance of items and their positions in the list.

4. **MAP** (mean average precision) is the mean of the average precision (AP) scores in all queries. AP averages the precision scores at all ranks where a relevant item is retrieved for a single query.

A.4 Additional Case Study

We conducted another case study showing the superiority of SECRET over the best baseline Trial2Vec. Similar to case study 1 (Table 5), we can see in case study 2 (Table 6) that the top-1 trial retrieved by SECRET is more closely aligned with the query trial compared to the trial retrieved by Trial2Vec. Specifically, some important eligibility criteria, such as cancer stage and weight loss, match between the query trial and the retrieved trial by SECRET.

A.5 Example Trial Represented as a set of Q/A pairs

Table 7 shows an example trial (NCT06095622) after we represent it using Q/A pairs. The current investigation limited evaluations of trial protocols to the following sections: *title, disease, intervention, keywords, outcome* and *eligibility criteria*. By *outcome* in the paper, we mean only *primary outcome measures*.

A.6 Additional Experiments on Partial Retrieval

Combining the title with additional sections (*disease, intervention, keywords, outcome, eligibility criteria*) consistently improves both precision@1 and recall@1 compared to using the title alone. Among these combinations, integrating the title with the intervention section yields the highest precision@1 and recall@1.

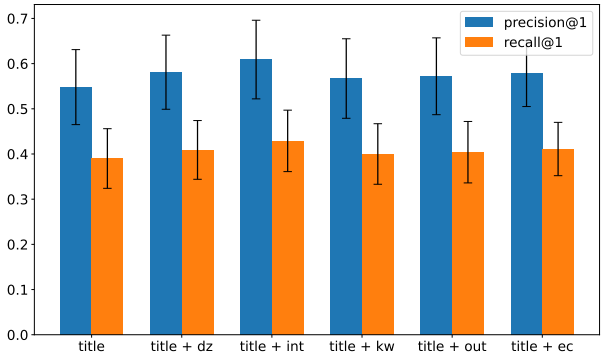


Figure 5: Performance of SECRET on the partial retrieval scenarios. We use different sections with title of the trial as queries to retrieve similar trials, including keyword kw, intervention int, disease dz, outcome out, eligibility criteria ec.

A.7 Effect of Representing Trials with Q/A Pairs

We performed some experiments to show the utility of using Q/A pairs as a way of representing the trial protocols. If we use whole sections of the trial protocols, some parts might get truncated depending on the length of the trial protocol and the method used to get embeddings. We show for BERT that representing trial protocols using Q/A pairs lead to significant improvement on retrieval scores. In Figure 6, BERT_q_a means that the trials are represented using Q/A pairs instead of the whole protocol. Similarly, in Figure 7, BioBERT_q_a means that the trials are represented using Q/A pairs instead of the whole protocol. We experimented with 5, 10 and all Q/A generated by LLMs.

	Query Trial	Trial2Vec	SECRET
NCTID	NCT01494558	NCT03800134	NCT00533949
Title	Study of Etoposide, Cisplatin, and Radiotherapy Versus Paclitaxel, Carboplatin and Radiotherapy ...	A Study of Neoadjuvant/Adjuvant Durvalumab for the Treatment of Patients With Resectable ...	High-Dose or Standard-Dose Radiation Therapy and Chemotherapy With or Without Cetuximab ...
Intervention	Chemoradiotherapy Regimen between PC (paclitaxel 45mg/m2 weekly over 1hour and carbplatin AUC =2mg/mL/min over 30min weekly) and PE (etoposide 50mg/m2 d1-5, 29-33 and cisplatin 50mg/m2 d1,8,29 and 36 29-33)	Drug: Durvalumab, Other: Placebo, Drug: Carboplatin, Drug: Cisplatin, Drug: Pemetrexed, Drug: Paclitaxel, Drug: Gemcitabine, Procedure: Surgery	Biological: Cetuximab, Drug: Carboplatin, Drug: Paclitaxel, Radiation: 60 Gy RT, Radiation: 74 Gy RT
Disease	Non-Small Cell Lung Cancer	Non-Small Cell Lung Cancer	Non-Small Cell Lung Cancer
Important Criteria	stage IIIA/IIIB NSCLC	Stage IIA to Stage IIIB	stage IIIA/IIIB NSCLC
	lose weight <10%	-	lose weight <10%

Table 6: Case study comparing the retrieval performance of the SECRET and Trial2Vec.

Question	**Answer**
What is the age requirement for participants?	18 years
What type of diabetes is required for participation?	Type-2
What is the dietary requirement for participants?	Chickpea rice pulao
What type of diet is not allowed for participants?	Vegan or keto
What are the drugs used?	Fenugreek Seeds and Indian Rennet
What is the disease treated in this trial?	Glucose Metabolism Disorders (Including Diabetes Mellitus)
What are the keywords?	Chickpea pulao
What is the title of the trial?	Formulation and Assessment of Chickpea Pulao Using Fenugreek Seeds and Indian Rennet for Improving Blood Glycemic Levels
What are the outcome measurements?	Improvement in blood glucose levels, Increase or decrease in postprandial glucose levels in mg/dL, 21 days

Table 7: Trial NCT06095622 represented as a set of Q/A pairs.

	precision@1	**recall@1**	**precision@2**	**recall@2**	**precision@5**	**recall@5**	**nDCG@5**	**MAP**
Only local	0.495 ± 0.083	0.343 ± 0.065	0.386 ± 0.051	0.495 ± 0.064	0.249 ± 0.025	0.767 ± 0.057	0.627 ± 0.057	0.603 ± 0.055
Only trial (full text)	0.560 ± 0.081	0.394 ± 0.064	0.463 ± 0.053	0.605 ± 0.068	0.275 ± 0.023	0.852 ± 0.050	0.716 ± 0.053	0.682 ± 0.054
Only global (Q/A set)	0.589 ± 0.087	0.419 ± 0.069	0.481 ± 0.053	0.641 ± 0.063	0.278 ± 0.027	0.871 ± 0.051	0.739 ± 0.055	0.703 ± 0.054
Predefined Q/A only	0.558 ± 0.084	0.401 ± 0.070	0.488 ± 0.056	0.644 ± 0.063	0.280 ± 0.027	0.877 ± 0.049	0.735 ± 0.051	0.701 ± 0.053
Answers only	0.607 ± 0.080	0.438 ± 0.067	0.456 ± 0.055	0.604 ± 0.070	0.287 ± 0.023	0.890 ± 0.043	0.751 ± 0.048	0.710 ± 0.051
SECRET	$\mathbf{0.647 \pm 0.077}$	$\mathbf{0.467 \pm 0.063}$	$\mathbf{0.508 \pm 0.046}$	$\mathbf{0.682 \pm 0.061}$	$\mathbf{0.297 \pm 0.023}$	$\mathbf{0.924 \pm 0.034}$	$\mathbf{0.796 \pm 0.042}$	$\mathbf{0.754 \pm 0.044}$

Table 8: Results of ablation study. The table presents precision, recall, nDCG, and MAP metrics, reported as mean $\pm$ standard deviation, with the best values highlighted in **bold**.

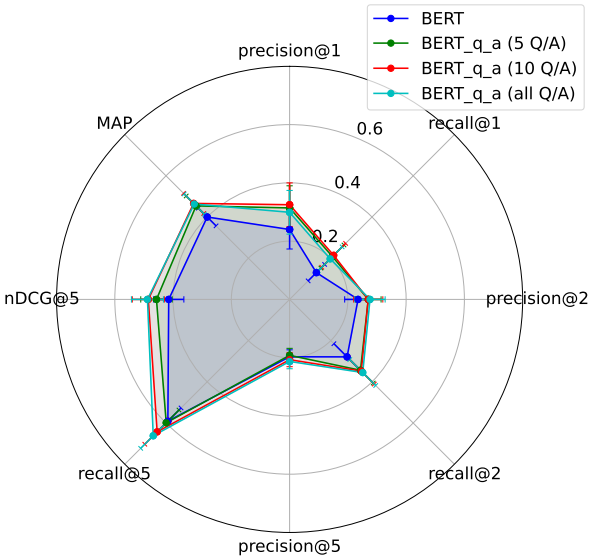


Figure 6: Performance evaluation of BERT and BERT_q_a models at different question-answer pair thresholds (5, 10, and all pairs) across various evaluation metrics: precision@1, recall@1, precision@2, recall@2, precision@5, recall@5, nDCG@5, and MAP.

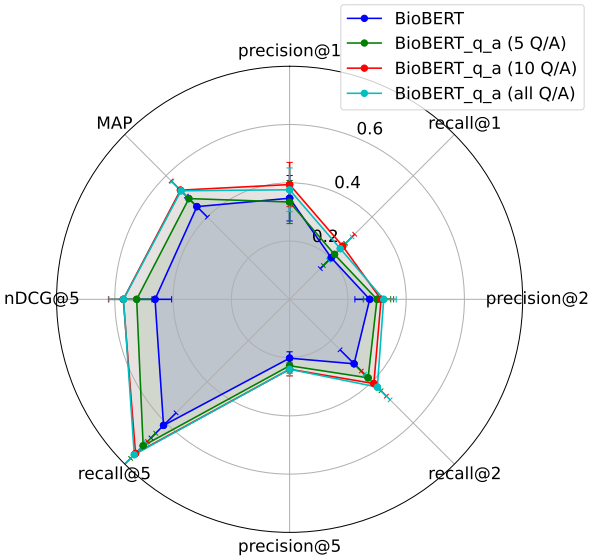


Figure 7: Performance evaluation of backbone BioBERT and BioBERT_q_a models at different question-answer pair thresholds (5, 10, and all pairs) across various evaluation metrics: precision@1, recall@1, precision@2, recall@2, precision@5, recall@5, nDCG@5, and MAP.