Logical Consistency is Vital: Neural-Symbolic Information Retrieval for Negative-Constraint Queries

Ganlin Xu[1♦*] **Zhoujia Zhang**[1♦*] **Wangyi Mei**[1♦] **Jiaqing Liang**[1■†] **Weijia Lu**[2▲]
Xiaodong Zhang[2▲] **Zhifei Yang**[2▲] **Xiaofeng Ma**[2▲] **Yanghua Xiao**[3■] **Deqing Yang**[1■†]
[1]School of Data Science, Fudan University, Shanghai, China
[2]United Automotive Electronic Systems, Shanghai, China
[3]College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China
♦{glxu24, zhangzj24, wymei24}@m.fudan.edu.cn
■{liangjiaqing, shawyh, yangdeqing}@fudan.edu.cn
▲{alfredwjlu, xiaodong.zhang.chn, yangzhifei006, maxf0124}@gmail.com

Abstract

Information retrieval plays a crucial role in resource localization. Current dense retrievers retrieve the relevant documents within a corpus via embedding similarities, which compute similarities between dense vectors mainly depending on word co-occurrence between queries and documents, but overlook the real query intents. Thus, they often retrieve numerous irrelevant documents. Particularly in the scenarios of complex queries such as *negative-constraint queries*, their retrieval performance could be catastrophic. To address the issue, we propose a neuro-symbolic information retrieval method, namely **NS-IR**, that leverages first-order logic (FOL) to optimize the embeddings of naive natural language by considering the *logical consistency* between queries and documents. Specifically, we introduce two novel techniques, *logic alignment* and *connective constraint*, to rerank candidate documents, thereby enhancing retrieval relevance. Furthermore, we construct a new dataset **NegConstraint** including negative-constraint queries to evaluate our NS-IR's performance on such complex IR scenarios. Our extensive experiments demonstrate that NS-IR not only achieves superior zero-shot retrieval performance on web search and low-resource retrieval tasks, but also performs better on negative-constraint queries. Our scource code and dataset are available at https://github.com/xgl-git/NS-IR-main.

1 Introduction

Information retrieval (IR) tasks aim at obtaining relevant information from large-scale data collection, such as documents and databases. Dense retrieval (Karpukhin et al., 2020a) is an advanced information retrieval technique focusing on semantic embedding similarities between texts. It has been widely adopted in many applications such as search engines (Li et al., 2022), question answering (Zhao et al., 2021) and retrieval-augmented generation (RAG) systems (Huly et al., 2024), offering significant improvements in IR.

Query: In what city was Yao Ming born?

Negtive document: ***Similarity score*:0.72**
Ming Yao is a retired Chinese professional basketball player who is widely regarded as one of the greatest basketball players from Chin...Ming Yao played as a center and had a dominant presence on the court...Ming Yao played for the Rockets for the entirety of his NBA career, from 2002 to 2011. During his time in the NBA, Ming Yao was an 8-time NBA All-Star and was inducted into the Naismith Memorial Basketball Hall of Fame in 2016.

Positive document: ***Similarity score*:0.67**
Shanghai is a bustling metropolis located on China's eastern coast, along the Yangtze River Delta… ***Shanghai has produced many famous historical and modern celebrities, such as Shi Hu, Ailing Zhang, Ming Yao***. Known for its impressive skyline dominated by modern skyscrapers like the iconic Oriental Pearl Tower and the Shanghai Tower, the city is a striking blend of the old and the new.

Figure 1: An illustration of BGE-based retrieval. The word marked in green is the co-occurrence word between the query and documents.

The embeddings (representations) generated by dense retrievers (such as BGE (Xiao et al., 2024)) focus on overall semantic similarity, which is capable of handling semantically similar words compared to sparse retrieval (such as BM25 (Robertson et al., 2009)) that uses keyword matching. However, dense retrieval still relies on superficial word co-occurrence between queries and documents. As illustrated in Figure 1, the negative document has a bigger score than the positive document just because the query's keyword "Ming Yao" occurs in the former more frequently. Thus, the approach fails to understand the real query intent, thereby retrieving irrelevant documents (Wu et al., 2024; Fang et al., 2024).

Notably, the retrieval approaches based on word co-occurrence have to face some challenges on

*Equal contribution.
† Co-corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 1828–1847

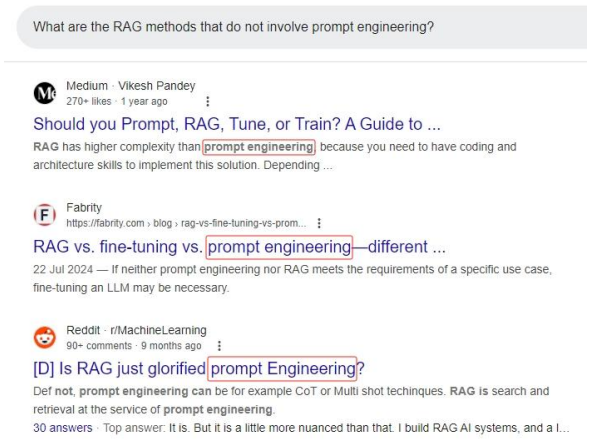

Figure 2: A retrieval example of Google search engine.

complex queries, particularly involving *negative-constraint queries*, due to the neglect of logical consistency. As shown in Figure 2[1], for a general query "What are the RAG methods that do not involve prompt engineering?", the documents returned by a keyword-based search engine often contain the excluded keyword "prompt engineering", which are not logically consistent with the query intent[2]. Therefore, understanding real query intents requires ensuring logical consistency between queries and documents besides semantic similarity. First-order logic (FOL), as a formal logical system, offers clear logical semantics and expresses complex relations in natural language (Barwise, 1977). For instance, the FOL of the aforementioned query is "RAGMethod(x) $\wedge$ $\neg$InvolvesPromptEngineering(x)", which clearly expresses the negative semantics through the logical connective '$\neg$'.

In light of this, through investigating the potential impact of FOL on complex logical queries, we propose a **N**eural-**S**ymbolic **I**nformation **R**etrieval method (NS-IR) to rerank the candidate documents returned by dense retrievers based on FOL, for more accurate retrieval results. Specifically, NS-IR first retrieves a set of ordered candidate documents using a dense retriever, then employs large language models (such as GPT-4o) to convert both the query and documents into FOL. To incorporate logical consistency and optimize embeddings of naive natural language (NL), we propose two independent techniques: 1) *Logic alignment*: To incorporate overall logic semantics in FOL into NL representations, we measure distribution differences between NL and FOL embeddings using optimal transport (Redko et al., 2019) and update the embeddings of queries and documents respectively. 2) *Connective constraint*: To reflect the role of partial words in FOL on logical consistency, we leverage different words in FOL (especially connectives) to render different attentions to words in NL, generating better embeddings with logical semantics. We leverage these two techniques to recalculate similarity scores and rerank candidate documents.

To evaluate the performance of NS-IR in negative-constraint query settings, we have constructed and released a dataset, namely **NegConstraint**, which contains three types of negative-constraint queries and was sourced from the Wikipedia dump (Karpukhin et al., 2020a). The experiments show our NS-IR's superiority over some state-of-the-art (SOTA) baselines on NegConstraint, and its potential for handling complex logical queries.

The main contributions of this paper include:

1. To address typical complex logical queries in IR, i.e., *negative-constraint queries*, we propose NS-IR which combines the strengths of NL and FOL to synthesize semantic similarity and logical consistency.

2. We introduce two key techniques: logic alignment (Sec. 4.2) and connective constraint (Sec. 4.3), to optimize naive NL embeddings by FOL, and rerank candidate documents to improve retrieval performance.

3. We further release a new dataset NegConstraint (Sec. 5), which can be utilized as a benchmark for negative-constraint queries. Our extensive experimental results on public datasets and NegConstraint show that our NS-IR significantly outperforms the SOTA methods on vanilla and negative-constraint queries in zero-shot settings, respectively. Our work in this paper paves the way for future research on complex queries in IR.

2 Related Work

2.1 Dense Retrieval

Dense retrieval has gained significant attention in information retrieval due to its advantages over traditional sparse vector space models. Sparse models represent documents and queries as high-

[1]In the example, the search engine's retrieval is based on BM25 algorithm, but dense retrieval produces similar results in negative-constraint queries.

[2]Although some search techniques in search engines can use '-' to filter keywords (such as "What are the RAG methods -prompt engineering"), they are not familiar to ordinary users.

dimensional vectors with mostly zero values (Yang et al., 2017; Chen et al., 2017). Dense retrieval models encode queries and documents into dense and low-dimensional vectors (Karpukhin et al., 2020a; Cai et al., 2022), which capture semantic similarity instead of match of terms, thus significantly outperforming sparse approaches. Relevant studies mainly focus on improving training approach (Qu et al., 2020), distillation (Zhang et al., 2023) and pre-training (Shen et al., 2022) for retrieval.

Many studies adopt a transfer learning framework where dense retrieval models are trained on high-resource passage retrieval datasets such as MS-MARCO (Bajaj et al., 2018) and then evaluated on queries from new tasks. However, collecting such large-scale corpora is both time-consuming and labor-intensive. Recent work has introduced zero-shot dense retrieval settings, which eliminates the need for relevance labels between queries and documents (Gao et al., 2022). Our work follows the zero-shot unsupervised setup for all experiments.

2.2 Optimal Transport in NLP

Optimal transport (OT) has been employed in various NLP tasks, where alignment exists implicitly or explicitly. The typical applications of OT include evaluating the similarity between sentences and documents (Wang et al., 2022; Mysore et al., 2021; Lee et al., 2022) or aligning cross-domain representations across different modalities (Zhou et al., 2023; Qiu et al., 2023). The evaluation mechanism can be integrated as a penalty term into language generation models (Chen et al., 2019; Zhang et al., 2020; Li et al., 2020). Moreover, OT effectively handles imbalanced word alignment, including both explicit alignment and null alignment (Arase et al., 2023). Inspired by unsupervised word alignment (Arase et al., 2023; Huang et al., 2024), we utilize the alignment matrix to measure distribution differences between natural language and first-order logic, enabling natural language to better focus on the logical semantics inherent in first-order logic.

2.3 NL-FOL Translation

NL-FOL (Natural Language to First-Order Logic) translation has long been a challenge in both natural language processing (NLP) and formal logic research. Traditionally, NL-FOL translation has been approached through rule-based methods (Abzianidze, 2017). However, due to the inherent complexity of natural language, these methods struggle to scale to real-world applications. As a result, traditional logic-based reasoning techniques have lost popularity due to limited scalability and coverage.

The recent breakthroughs in LLMs have reignited interest in logic, bringing it back to the forefront of reasoning tasks. One promising strategy to leverage the power of LLMs is to translate NL statements, such as premises and conclusions in textual entailment tasks, into first-order logic (FOL) formulas via in-context learning. These symbolic representations can be passed to Symbolic Mathematical Theory (SMT) solvers (Olausson et al., 2023; Xu et al., 2024) or used to make veracity predictions and generate explanations (Wang and Shu, 2023). In the context, (Yang et al., 2024) presents a NL-FOL dataset MALLs of 28K diverse and verified sentence-level pairs collected from GPT4, and a translator LOGICLLAMA, a LLaMA2-7B/13B fine-tuned on MALLS for NL-FOL translation. In this paper, we use LLMs as translators to implement NL-FOL translation.

3 Preliminaries

3.1 Task Formulation

In this paper, we focus on the task of zero-shot document retrieval, of which the model captures the similarity between queries and documents without model training. Given a query q and the document set D containing multiple documents, the goal of retrievers is to retrieve document d that satisfies the user's real search intent. Dense retrieval uses encoders to map q and d into a pair of dense vectors, whose inner product is leveraged as a similarity function:

$$\text{sim}(q, d) = \langle E_q(q), E_d(d) \rangle. \quad (1)$$

In this paper, we use the BGE model as query encoder E_q and document encoder E_d, and the embeddings of CLS token (the first token of a sequence) denotes dense vectors $E_q(q)$ and $E_d(d)$. Besides, we obtain word embeddings of NL and FOL sequences via BGE, respectively. Let $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_m](\boldsymbol{h}_i \in \mathbb{R}^d, 1 \leq i \leq m)$ and $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n](\boldsymbol{z}_j \in \mathbb{R}^d, 1 \leq i \leq n)$ be the embeddings of NL-queries and FOL-queries, respectively. We obtain the embeddings of NL-documents and FOL-documents in the same way[3]. We provide

[3]In the paper, NL-query and FOL-query refer to queries in natural language and first-order logic, respectively. Similarly,

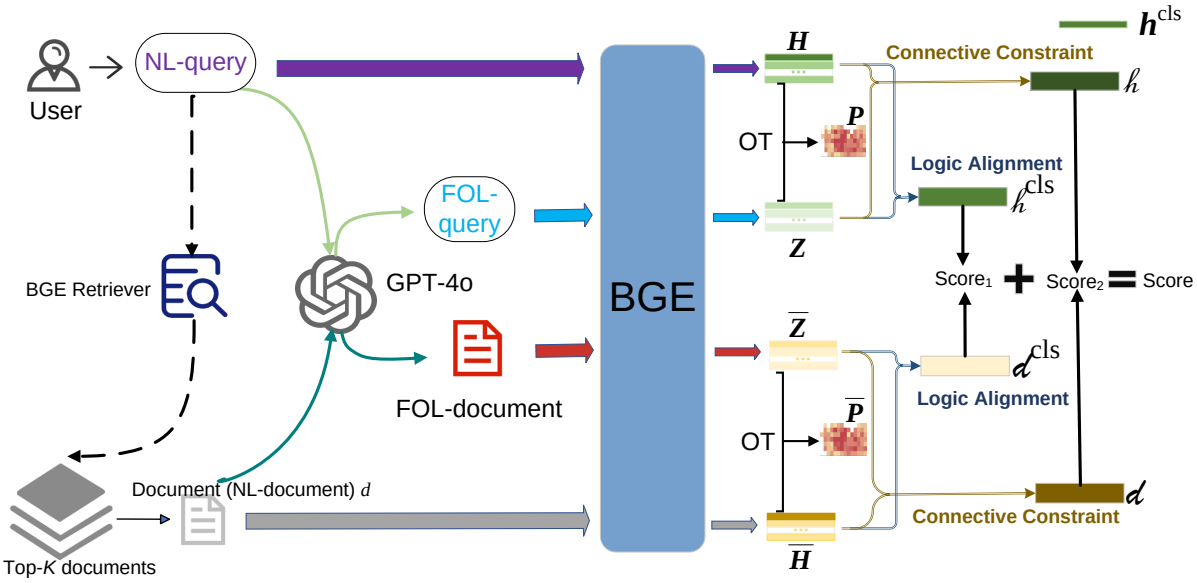

Figure 3: The pipeline of our proposed NS-IR. Dashed arrows represent the retrieval stage. In the figure, only one document d in top-K documents is encoded, but actually, all top-K documents are encoded together.

some examples of FOL in Appendix A.

3.2 Optimal Transport

Optimal transport (OT) seeks to find the most efficient way to transport one probability distribution μ (the source) to another ν (the target) while minimizing a predefined cost function (Redko et al., 2019). Formally, let μ and ν be probability measures on spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, and let function $c(\cdot,\cdot)$ represent the cost of transporting a unit of mass from points. The following explanation assumes that the source and target sentences $\boldsymbol{H}$ and $\boldsymbol{Z}$ and their word embeddings are at hand. A cost means a dissimilarity between $\boldsymbol{h}_i$ and $\boldsymbol{z}_j$ (NL and FOL word embeddings) computed by a distance metric $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, such as cosine distances. The cost matrix $\boldsymbol{C} \in \mathbb{R}_+^{m\times n}$ summaries the costs of any word pairs, that is, $\boldsymbol{C}_{i,j} = c(\boldsymbol{h}_i, \boldsymbol{z}_j)$. The OT problem identifies an alignment matrix P with which the sum of alignment costs is minimized under the cost matrix $\boldsymbol{C}$:

$$L_{\boldsymbol{C}(\mu,\nu)} := \min_{\boldsymbol{P}\in U(\mu,\nu)} \langle \boldsymbol{C}, \boldsymbol{P}\rangle, \qquad (2)$$

where $U(\mu,\nu) := \{\boldsymbol{P} \in \mathbb{R}_+^{m\times n} : \boldsymbol{P}\mathbb{1}_m = \boldsymbol{a}, \boldsymbol{P}^\top \mathbb{1}_n = \boldsymbol{b}\}$. $\boldsymbol{P}_{i,j} \neq 0$ if the i-th source word is aligned to the j-th target word, such that the aligned words have the smallest distance in the cost matrix $\boldsymbol{C}$. With this formulation, we can seek alignment matrix $\boldsymbol{P}$.

NL-documents and FOL-documents correspond to documents in natural language and first-order logic, respectively.

4 Methodology

4.1 Overview

The pipeline of our NS-IR is shown in Figure 3. To reduce the cost of NL-FOL translation, we first use BGE retriever to initially retrieve the top-K documents $D = \{d_1, \ldots, d_i, \ldots, d_K\}$ for a given query of NL (denoted as NL-query). Then, inspired by previous work (Olausson et al., 2023; Xu et al., 2024), we use the specific prompts detailed in Appendix B to let an LLM (GPT-4o) perform NL-FOL translation, so as to obtain the query and document of FOL (denoted as FOL-query and FOL-document respectively). Next, we employ BGE[4] to encode the NL-query, FOL-query, NL-document, and FOL-document to obtain corresponding embeddings. Finally, we introduce two independent techniques: logic alignment (Sec. 4.2) and connective constraint (Sec., 4.3) to recalculate the scores between queries and documents and rerank candidate documents[5].

4.2 Logic Alignment

To incorporate overall logic semantics in FOL into NL representations, we propose logic alignment based on optimal transport (OT) which is inspired by unsupervised word alignment (Arase

[4]Significantly, we use BGE twice for different purposes. The first time is to retrieve Top-K documents, and the second time is to encode NL and FOL.

[5]In this paper, if not specifically stated, queries and documents denote NL-queries and NL-documents, respectively.

Formulation	Query Example	#Query	#Pos.	#Neg.	#Irr.
A - a	Introduce Allen Ginsberg's works$_{(A)}$, but do not mention 'Howl'$_{(a)}$.	136	136	136	3000
(A - a) ∪ B	What themes are expressed in Allen Ginsberg's works$_{(A)}$ other than 'Howl'$_{(a)}$ and Edgar Allan Poe's works$_{(B)}$?	123	123	123	
(A - a) ∪ (B - b)	What themes do Allen Ginsberg's works$_{(A)}$ other than ' Howl'$_{(a)}$ and Edgar Allan Poe's works$_{(B)}$ other than 'The Raven'$_{(b)}$ express?	107	107	321	

Table 1: Examples of negative-constraint queries in our constructed dataset NegConstraint. #Query denotes the number of each type of query, #Pos., #Neg., and #Irr. denote the number of positive documents, and irrelevant documents, respectively.

et al., 2023; Huang et al., 2024). This approach measures the distribution differences between NL and FOL, and integrates word features of NL and FOL with context representation via the alignment matrix.

Specifically, for a given NL-query and FOL-query, we first use BGE to obtain their corresponding word embeddings $\boldsymbol{H}$ and $\boldsymbol{Z}$[6], respectively. Then, we compute a pairwise similarity between $\boldsymbol{H}$ and $\boldsymbol{Z}$ using cosine distance $\boldsymbol{C}_{i,j}$:

$$\boldsymbol{C}_{i,j} = 1 - \frac{\boldsymbol{h}_i^T \boldsymbol{z}_j}{\|\boldsymbol{h}_i\| \|\boldsymbol{z}_j\|}. \tag{3}$$

The OT can be formulated as:

$$\boldsymbol{P}^* = \underset{\boldsymbol{P} \in U(\boldsymbol{H}, \boldsymbol{Z})}{\operatorname{argmin}} \sum_{i,j} \boldsymbol{C}_{i,j} \boldsymbol{P}_{i,j}, \tag{4}$$

where alignment matrix $\boldsymbol{P}_{i,j} \in \mathbb{R}^{m \times n}$ indicates a likelihood of aligning $\boldsymbol{h}_i$ with $\boldsymbol{z}_j$. The optimization in Eq. 4 can be solved by linear programming (Bourgeois and Lassalle, 1971).

Finally, we integrate $\boldsymbol{H}$, $\boldsymbol{Z}$, $\boldsymbol{P}$ and $\boldsymbol{h}^{cls} \in \mathbb{R}^d$ ($\boldsymbol{h}^{cls} \in \mathbb{R}^d$ denote the embeding of special token CLS[7].) to obtain updated embedding $\hbar^{cls} \in \mathbb{R}^d$:

$$\hbar^{cls} = \boldsymbol{H}^T \cdot \boldsymbol{P} \cdot \boldsymbol{Z} \cdot \boldsymbol{h}^{cls}. \tag{5}$$

This approach synthesizes word distributions of FOL and NL, as well as context features, via alignment matrix $\boldsymbol{P}_{i,j}$ as the intermediary.

We use the same method to obtain the embedding $\mathscr{d}^{cls} \in \mathbb{R}^d$ of a document d in the document set D, and then calculate the similarity score between $\hbar^{cls}$ and $\mathscr{d}^{cls}$ as

$$\text{score}_1 = \text{sim}\left(\hbar^{cls} \cdot \mathscr{d}^{cls}\right). \tag{6}$$

[6] $\overline{\boldsymbol{H}}$ and $\overline{\boldsymbol{Z}}$ denote corresponding word embeddings of NL-documents and FOL-documents, respectively.

[7] In fact, the first token of queries is denoted as CLS, i.e., $\boldsymbol{h}^{cls} = \boldsymbol{h}_1$.

4.3 Connective Constraint

To precisely reflect the role of partial words in FOL on logical consistency, we also propose connective constraint that enables different words in FOL (especially connectives) to render different attentions to words in NL, thus generating better embeddings with logical semantics. Given a FOL-query sequence $t = \{t_1, ..., t_i, ..., t_n\}$, as well as NL-query word embeddings $\boldsymbol{H}$ and FOL-query word embeddings $\boldsymbol{Z}$, we integrate the embeddings of logical connectives into the attentions. That is, when calculating attention weights (Eq. 8) of FOL to NL and updated embeddings (Eq. 7), it takes the alignment between NL and FOL and the embeddings of logical connectives into account:

$$\hbar_j = \sum_{i=1}^{m} \alpha_{ji} \left(\boldsymbol{h}_i + \sigma_{ji} \boldsymbol{z}_j\right), \tag{7}$$

$$\alpha_{ji} = \frac{e^{\alpha'_{ji}}}{\sum_{i=1}^{m} e^{\alpha'_{ji}}}, \quad \alpha'_{ji} = \frac{\boldsymbol{z}_j (\boldsymbol{h}_i + \sigma_{ji} \boldsymbol{z}_j)^T}{\sqrt{d_k}}, \tag{8}$$

$$\sigma_{ji} = \begin{cases} 1, & t_j \neq \neg, t_j \in \mathcal{C}, \boldsymbol{P}_{ij} = 0 \\ -1, & t_j = \neg, \boldsymbol{P}_{ij} = 0 \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where $\boldsymbol{P}$ denotes the alignment matrix and logical connective set $\mathcal{C} = \{\neg, \rightarrow, \leftrightarrow, \wedge, \vee, \oplus\}$.

Our explanations for above equations are as follows. For the words that do not exhibit significant alignment between FOL and NL (where $\boldsymbol{P}_{i,j} = 0$), we incorporate logical connective embeddings to further enhance logical semantics. Additionally, for negative connective ($t_j = \neg$), the subtraction implies the negative-constraint semantics.

Finally, we perform mean pooling on $\hbar_j$ to obtain a query's embedding as

$$\hbar = \text{mean-pooling}(\hbar_j), \quad \hbar \in \mathbb{R}^d. \tag{10}$$

Similarly, we obtain the embedding of a document d from the document collection D, and then compute the similarity score between $\hbar$ and $\mathscr{d}$ as

$$\text{score}_2 = \text{sim}(\hbar \cdot \mathscr{d}). \tag{11}$$

For document d, the final recommended score is

$$\text{score} = \text{score}_1 + \text{score}_2. \tag{12}$$

Then, we rerank the top-K candidate documents based on the recommendation scores.

5 NegConstraint

To evaluate our NS-IR's performance specifically to negative-constraint queries, we have constructed a human-annotated dataset **NegConstraint**. Table 1 lists three formulations that represent three types of negative-constraints query. Let the letters (A, a, B, and b) denote entity sets where the lowercase letters represent the subsets of the sets denoted by the uppercase letters. Operator '-' indicates a negative-constraint condition, and '∪' denotes union operation. There are 136 queries corresponding to the first formulation, 123 queries corresponding to the second formulation, and 107 queries corresponding to the third formulation in NegConstraint. Each query is paired with one positive document and one or three negative documents as distractors. For example, for a query of formulation **A – a**, such as "Introduce Allen Ginsberg's works, but do not mention 'Howl"', there is a positive document that introduces the content about Allen Ginsberg's works but does not mention 'Howl'. Similarly, a negative document introduces Allen Ginsberg's and mentions 'Howl'. The documents stem from the Wikipedia dump (Karpukhin et al., 2020a), and the queries are generated by GPT-4o based on corresponding negative and positive documents where the prompts are presented in Appendix B. Besides, NegConstraint contains 3,000 irrelevant documents[8] that are irrelevant to all queries. More details about our data collection and snippets are provided in Appendix C.

6 Experiments

6.1 Datasets and Metrics

For **low-resource retrieval**, we use six diverse low-resource retrieval datasets from the BEIR benchmark (Thakur et al., 2021), including SciFact (fact-checking), ArguAna (argument retrieval), TREC-COVID (bio-medical IR), FiQA (financial QA), DBPedia (entity retrieval), and NFCorpus (medical IR). For this task, we report all compared methods' scores of the representative metric nDCG@10. For **web search**, we adopt widely-used web search dataset TREC Deep Learning 2019 (DL'19) (Craswell et al., 2020) and Deep Learning 2020 (DL'20) (Craswell et al., 2021) based on MS-MARCO (Bajaj et al., 2018). For these two datasets and our NegConstraint, we report the methods' scores of MAP (Mean Average Precision) and nDCG@10.

[8]Negative documents are essentially irrelevant documents, but easier to mislead retriever models.

6.2 Baselines

We first compare our NS-IR with some fully-supervised retrieval methods that are fine-tuned with extensive query-document relevance data (denoted as **w/ relevance judgment**), including DPR (Karpukhin et al., 2020b), ANCE (Xiong et al., 2020), and the fine-tuned Contriever (denoted as Contriever^{FT} (Izacard et al., 2021)). We also consider several zero-shot retrieval models not involving query-document relevance labels (denoted as **w/o relevance judgment**), including sparse retriever BM25 (Robertson et al., 2009), dense retriever BGE (Xiao et al., 2024), and Contriever (Izacard et al., 2021). For this type of baselines, we further consider the LLM-based retrieval models which rewrite queries with LLMs, including HyDE (Gao et al., 2023) and InteR (Feng et al., 2024).

6.3 Implementation Details

We employ "bge-large-en-v1.5" as embedding models. To make a fair comparison, we also reproduce results on HyDE and InteR where "bge-large-en-v1.5" is used as retriever models. We run all experiments on one Nvidia A800 80GB GPU. For NL-FOL translation, we use OpenAI API with a temperature of 0.5.

6.4 Main Results

In the following presentation, the techniques of logic alignment and connective constraint we proposed for NS-IR are abbreviated as LA and CC, respectively. In our comparison experiments, we adopt two variants of NS-IR which use LogicLLaMA (Yang et al., 2024) and GPT-4o to generate FOL, respectively.

Table 2 shows that, our NS-IR (GPT-4o) outperforms all baselines significantly on the tasks

Methods	SciFact	ArguAna	TREC-COVID	FiQA	DBPedia	NFCorpus	DL'19		DL'20	
w/ relevance judgment	nDCG@10						MAP	nDCG@10	MAP	nDCG@10
DPR	31.8	17.5	33.2	29.5	26.3	18.9	36.5	62.2	41.8	65.3
ANCE	50.7	41.5	65.4	30.0	28.1	23.7	37.1	64.5	40.8	64.6
Contriever^{FT}	67.7	44.6	59.6	32.9	41.3	32.8	41.7	62.1	43.6	63.2
w/o relevance judgment	nDCG@10						MAP	nDCG@10	MAP	nDCG@10
BM25♡	67.1	43.2	55.5	25.1	26.1	31.4	31.2	55.4	30.6	50.1
Contriever♡	55.0	44.5	12.5	12.4	29.2	26.0	22.8	37.5	24.3	42.5
HyDE♡	71.9	49.6	78.4	31.3	38.7	37.3	48.7	67.3	49.8	66.8
InterR♡	72.1	50.9	79.2	33.5	42.1	39.5	50.4	**69.7**	47.8	67.5
BGE♡*	71.3	48.4	75.3	30.6	38.9	35.4	46.9	64.4	45.7	63.4
BGE w/ LA♡*	72.6	53.2	78.3	33.7	42.8	38.1	48.9	67.5	47.5	68.9
BGE w/ CC♡*	73.3	52.3	77.6	35.6	43.2	37.7	49.1	66.8	48.5	66.6
NS-IR (LogicLLaMA)	73.7	51.1	78.8	35.5	42.6	38.8	49.8	67.9	48.9	68.1
NS-IR (GPT-4o)	**75.8**	**55.1**	**81.8**	**38.4**	**46.1**	**40.7**	**51.4**	68.4	**50.8**	**70.5**

Table 2: Performance of compared methods on the benchmarks of low-resource retrieval and web search. ♡ indicates the reported results were reproduced by us using the baselines' sourcecodes. We employ BGE as the embedding model in HyDE and InteR for fair comparison. * denotes the ablated variants of NS-IR which can be regarded as BGE w/ LA&CC.

Methods	**A - a**		**(A - a) ∪ B**		**(A - a) ∪ (B - b)**		Total	
	MAP	nDCG@10	MAP	nDCG@10	MAP	nDCG@10	MAP	nDCG@10
BM25♡	32.1	34.6	31.2	34.7	29.3	31.5	31.4	33.7
Contriever♡	34.8	36.6	32.1	33.3	30.9	32.7	31.8	35.7
HyDE♡	50.7	55.3	48.6	51.5	45.7	50.6	47.8	53.1
InterR♡	52.6	55.8	50.3	48.7	51.5	49.3	52.3	54.5
BGE♡*	37.9	40.5	34.8	36.8	33.7	34.8	36.3	40.8
BGE w/ LA♡*	42.1	45.6	41.2	44.6	39.9	42.5	40.8	47.6
BGE w/ CC♡*	48.9	50.7	46.6	48.5	43.9	48.2	47.8	46.9
NS-IR (LogicLLaMA)	53.2	54.6	51.6	50.2	49.6	**54.1**	50.7	55.2
NS-IR (GPT-4o)	**54.7**	**57.9**	**53.3**	**54.2**	**51.7**	53.7	**53.3**	**56.5**

Table 3: Performance comparisons of different methods on NegConstraint.

of low-resource retrieval and web search, including the SOTA model without relevance judgment InteR. Specifically, NS-IR (GPT-4o) obtains an average performance improvement of over 10% relative to the vanilla BGE. The inferiority of NS-IR (LogicLLaMA) compared to NS-IR (GPT-4o) is attributed to LogicLLaMA's weakness on NL-FOL translation.

For negative-constraint queries, we compare NS-IR and its ablated variants with the baselines without relevance judgment. Table 3 reports the compared methods' performance on three types of negative-constraint queries and whole queries in NegConstraint. The results reveal our method's superiority over the baselines on negative-constraint queries, which is achieved through synthesizing semantic similarity and logical consistency for handling complex logical queries. Although HyDE and InterR can partially eliminate the impact of negative constraints via hypothetical documents generated by LLMs, our proposed LA and CC are more effective.

The results in Tables 2 and 3 related to NS-IR's ablated variants (marked by *) also justify the effectiveness of employing either LA or CC. In particular, adopting CC improves NS-IR's performance more obviously than adopting LA on NegConstraint, suggesting that CC is more effective than LA in the scenarios of negative-constraint queries.

6.5 Effects of Different Dense Retrievers

To verify the effectiveness of different dense retrievers, we report the web search performance of HyDE, InteR, and NS-IR (GPT-4o) with different dense retrievers (bge-small, bge-base, and Contriver)[9] in Table 4. The results indicate that more powerful retriever models can facilitate accurate IR. NS-IR is consistently superior to HyDE and InteR with all retrievers. These results also indicate

[9]We replace dense retrievers in the process of retrieval and encoding as introduced in Sec. 4.1.

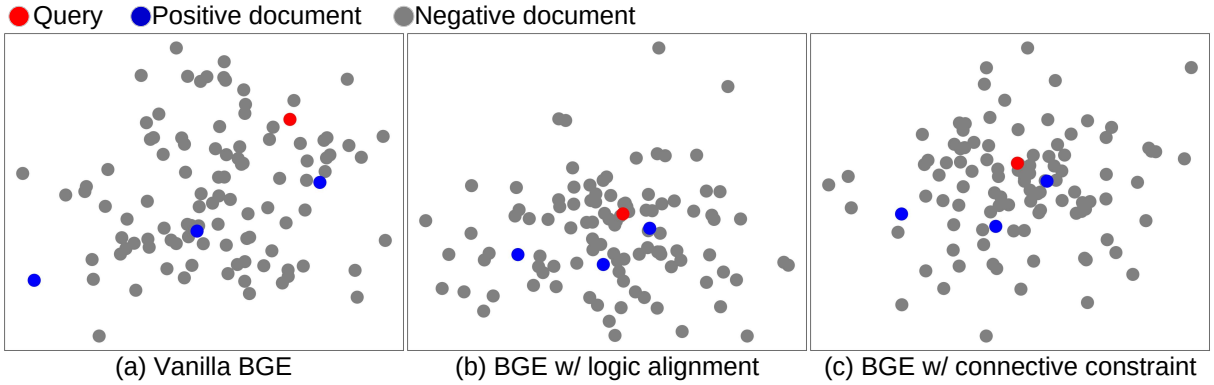

Figure 4: An example of query embedding visualization from TREC-COVID (better viewed in color): *What are the observed mutations in the SARS-CoV-2 genome and how often do the mutations occur?*

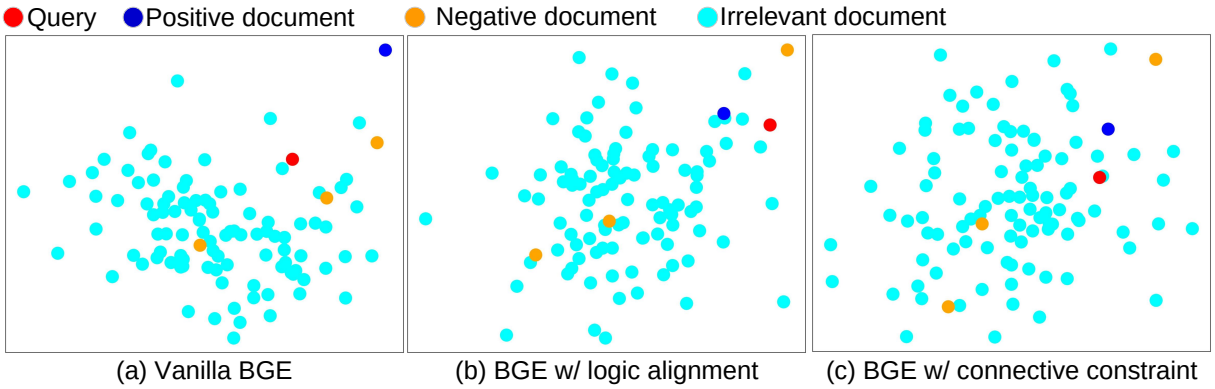

Figure 5: An example of query embedding visualization from NegConstraint (better viewed in color): *What are the similarities between Ginsberg's works (excluding 'Howl') and Poe's works (excluding 'The Raven')?*

Methods	DL'19		DL'20	
	MAP	nDCG@10	MAP	nDCG@10
bge-small	40.5	60.1	40.4	61.2
+ HyDE	42.7	61.4	41.8	62.7
+ InteR	43.6	63.4	42.8	63.2
+ NS-IR	**43.8**	**65.4**	**44.4**	**64.9**
bge-base	41.9	62.5	40.9	63.1
+ HyDE	42.4	64.4	42.6	64.5
+ InteR	45.9	65.3	45.5	66.7
+ NS-IR	**46.6**	**66.4**	**47.9**	**68.8**
Contriever	35.7	57.7	37.8	56.9
+ HyDE	37.5	59.3	39.9	58.9
+ InteR	38.6	58.9	39.8	61.7
+ NS-IR	**40.6**	**61.4**	**41.4**	**62.8**

Table 4: Web search performance of adopting different dense retriever models in HyDE, InteR and our NS-IR.

that the effectiveness of LA and CC on NS-IR's performance gains are model-agnostic.

6.6 Visualization on the Effects of Logic Alignment and Connective Constraint

We randomly pick two queries from TREC-COVID and our NegConstraint to visualize the effects of LA and CC. In Figures 4 and 5, we plot the embeddings generated by vanilla BGE, BGE w/ LA and BGE w/ CC in the embedding space using t-SNE, respectively. In Figure 4 of TREC-COVID, we can see that the query embeddings generated by BGE w/ LA and BGE w/ CC are closer to that of positive documents than the query embeddings generated by vanilla BGE. In Figure 5 of NegConstraint, the query embeddings generated by BGE w/ LA and BGE w/ CC are closer to that of positive documents and farther away from that of negative documents, compared to the query embeddings generated by vanilla BGE. These results demonstrate that LA anc CC are more effective on identifying positive documents.

6.7 Visualization on the Attention of Connective Constraint

As introduced in Sec. 4.2, CC enables the words in FOL (especially connectives) to assign different attentions to different words in NL. To verify the hypothesis, we examine the attention scores of logical negation $\neg$ in FOL to the words in NL.

Query: What are the similarities between Ginsberg's works (excluding 'Howl') and Poe's works (excluding 'The Raven')?

Positive Document: Ginsberg took part in decades of non-violent political protest against everything from the Vietnam War to the War on Drugs. His poem ""September on Jessore Road"", calling attention to the plight of Bangladeshi refugees, exemplifies what the literary critic Helen Vendler described as Ginsberg's tireless persistence…, though he died before it could be produced. by gambling, and the cost of secondary education for Poe. He attended the University of Virginia but left after a year due to lack of money. Poe quarreled with Allan over the funds for his education and enlisted in the Army in 1827 under an assumed name…Allan reached a temporary rapprochement. However, Poe later failed as an officer cadet at West Point, declaring…

Figure 6: Attention scores of logical connective ¬ in FOL to the words in NL. A deeper color indicates a bigger score (better viewed in color).

As shown in Figure 6, the deeper colors indicate the larger attention scores. We suppose that this operation emphasizes important entity tokens (such as 'Ginsberg' and 'Poe') and ignores entity tokens in negative-constraint conditions (such as 'Howl' and 'Raven'). That is, logical connective ¬ implies negative-constraint semantics. It reveals that our method tends to retrieve the documents without negative-constraint conditions mentioned in queries.

7 Effect of Parameter *K*

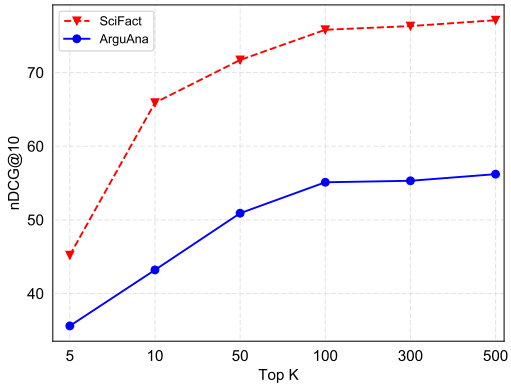

Figure 7: The performance of NS-IR on different Top *K* of SciFact and ArguAna.

We perform an additional study to investigate the impact of the number of retrieved documents (i.e., Top *K*) on the performance of NS-IR. Figure 7 illustrates nDCG@10 under different *K* on SciFact and ArguAna. Our observations revealed consistent patterns in both datasets: as Top *K* increased, performance showed a gradual improvement until *K* reached 100. Subsequently, the performance stabilized, indicating that increased *K* did not significantly enhance the results. This phenomenon can be attributed to the fact that the first 100 retrieved documents have already covered a significant amount of positive documents for a query. Additionally, the larger number of retrieval documents will extremely increase expenses for generating FOL. Therefore, we select 100 as Top *K* in this paper.

8 Conclusion

In this paper, we propose a novel IR method NS-IR, which integrates the strengths of NL and FOL and synthesizes semantic similarity and logical consistency. We specially propose two key techniques: logic alignment and connective constraint, to rerank the candidate documents. We also release a negative-constraint query dataset NegConstraint to evaluate our method. Extensive experiments on public IR benchmarks and NegConstraint show that, NS-IR significantly outperforms the existing IR approaches for general and negative-constraint queries under zero-shot settings, paving the way for future study on complex logical queries. Therefore, we will focus on more complex logical queries generated by set operations (such as union, intersection, difference, and complement) in the future.

9 Acknowledgments

The authors disclosed receipt of the following financial support for the research, authorship, and publication of this article: This research was supported by the AI Laboratory of United Automotive Electronic Systems (UAES) Co. (Grant no. 2025-3944) and the Chinese NSF Major Research Plan (No.92270121).

Limitations

We acknowledge that our method has several limitations. First, calling OpenAI API to perform NL-FOL translation will inevitably incur additional expenses to maintain high retrieval relevance. Second, to reduce the expenses of NL-FOL translation, we perform NL-FOL translation on the initially retrieved and limited documents, thus slightly reducing recall. Third, we use the same prompts for NL-FOL translation on all benchmarks, which may hinder further improvement. Therefore, these limitations are caused by NL-FOL translation. Although NL-FOL translation is not the main focus

of this paper, we argue that the limitations will be improved with further study in the era of LLMs.

References

Lasha Abzianidze. 2017. Langpro: Natural language theorem prover. *arXiv preprint arXiv:1708.09417*.

Yuki Arase, Han Bao, and Sho Yokoi. 2023. Unbalanced optimal transport for unbalanced word alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3966–3986, Toronto, Canada. Association for Computational Linguistics.

P Bajaj, D Campos, N Craswell, L Deng, J Gao, X Liu, R Majumder, A McNamara, B Mitra, T Nguyen, et al. 2018. A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Jon Barwise. 1977. An introduction to first-order logic. In *Studies in Logic and the Foundations of Mathematics*, volume 90, pages 5–46. Elsevier.

Francois Bourgeois and Jean-Claude Lassalle. 1971. An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*, 14(12):802–804.

ZeFeng Cai, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, Xin Alex Lin, Liang He, and Daxin Jiang. 2022. Hyper: Multitask hyper-prompted training enables large-scale retrieval generalization. In *The Eleventh International Conference on Learning Representations*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Improving sequence-to-sequence learning via optimal transport. *arXiv preprint arXiv:1901.06283*.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track. *Preprint*, arXiv:2102.07662.

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

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10028–10039, Bangkok, Thailand. Association for Computational Linguistics.

Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2024. Synergistic interplay between search and large language models for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9571–9583, Bangkok, Thailand. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Chenyang Huang, Abbas Ghaddar, Ivan Kobyzev, Mehdi Rezagholizadeh, Osmar Zaiane, and Boxing Chen. 2024. OTTAWA: Optimal TransporT adaptive word aligner for hallucination and omission translation errors detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6322–6334, Bangkok, Thailand. Association for Computational Linguistics.

Oz Huly, Idan Pogrebinsky, David Carmel, Oren Kurland, and Yoelle Maarek. 2024. Old ir methods meet rag. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2559–2563.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

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

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020b. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu. 2022. Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. *arXiv preprint arXiv:2202.13196*.

Dan Li, Vikrant Yadav, Zubair Afzal, and George Tsatsaronis. 2022. Unsupervised dense retrieval for scientific articles. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 313–321.

Jianqiao Li, Chunyuan Li, Guoyin Wang, Hao Fu, Yuhchen Lin, Liqun Chen, Yizhe Zhang, Chenyang Tao, Ruiyi Zhang, Wenlin Wang, Dinghan Shen, Qian Yang, and Lawrence Carin. 2020. Improving text generation with student-forcing optimal transport. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9144–9156, Online. Association for Computational Linguistics.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2021. Multi-vector models with textual guidance for fine-grained scientific document similarity. *arXiv preprint arXiv:2111.08366*.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. 2023. SCCS: Semantics-consistent cross-domain summarization via optimal transport alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1584–1601, Toronto, Canada. Association for Computational Linguistics.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. 2019. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on artificial intelligence and statistics*, pages 849–858. PMLR.

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

Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Xiaolong Huang, Binxing Jiao, Linjun Yang, and Daxin Jiang. 2022. Lexmae: Lexicon-bottlenecked pretraining for large-scale retrieval. *arXiv preprint arXiv:2208.14754*.

N Thakur, N Reimers, A Rücklé, A Srivastava, and I Beir Gurevych. 2021. A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.

Zihao Wang, Jiaheng Dou, and Yong Zhang. 2022. Unsupervised sentence textual similarity with compositional phrase semantics. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4976–4995.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *Preprint*, arXiv:2404.03302.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. *Preprint*, arXiv:2309.07597.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand. Association for Computational Linguistics.

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.

Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2024. Harnessing the power of large language models for natural language to first-order logic translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6942–6959, Bangkok, Thailand. Association for Computational Linguistics.

Kai Zhang, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, Binxing Jiao, and Daxin Jiang. 2023. Led: Lexicon-enlightened dense retriever for large-scale retrieval. In *Proceedings of the ACM Web Conference 2023*, pages 3203–3213.

Shuying Zhang, Tianyu Zhao, and Tatsuya Kawahara. 2020. Topic-relevant response generation using optimal transport for an open-domain dialog system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4067–4077.

Chen Zhao, Chenyan Xiong, Jordan Boyd-Graber, and Hal Daumé Iii. 2021. Distantly-supervised dense

retrieval enables open-domain question answering without evidence annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9622.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7873–7887, Toronto, Canada. Association for Computational Linguistics.

A FOL Examples

Table 5 provides an example of the NL-query, FOL-query, NL-document, and FOL-document. The NL-query and NL-document stem from actual datasets. FOL-query and FOL-document are FOL formats of NL-queries and NL-documents, respectively, which are generated by GPT-4o in this paper.

NL-query	What is considered a business expense on a business trip?
FOL-query	$\forall$x (BusinessTrip(x) $\rightarrow$ BusinessExpense(x))
NL-documnet	I'm not saying I don't like the idea of on-the-job training too, but you can't expect the company to do that. Training workers is not their job - they're building software. Perhaps educational systems in the U.S. (or their students) should worry a little about getting marketable skills in exchange for their massive investment in education, rather than getting out with thousands in student debt and then complaining that they aren't qualified to do anything.
FOL-document	$\neg$Like(i, onTheJobTraining) $\wedge\neg$Expect(company, Train(workers)) $\neg$Job(company, Train(workers)) $\wedge$Job(company, Build(software)) $\forall$x (EducationalSystems(x) $\rightarrow$ Worry(x, MarketableSkills)) Invest(educationalSystems, education) $\wedge\exists$x (Student(x) $\wedge$StudentDebt(x)) $\rightarrow$ $\neg$Qualified(x, anything)

Table 5: An example of the NL-query, FOL-query, NL-document and FOL-document.

B Prompts

Prompt of NL-FOL Translation for Queries

Given some question. The task is to parse these questions into first-order logic formulars. The grammar of the first-order logic formular is defined as follows:
1) logical conjunction of expr1 and expr2: expr1 $\wedge$ expr2
2) logical disjunction of expr1 and expr2: expr1 $\vee$ expr2
3) logical exclusive disjunction of expr1 and expr2: expr1 $\oplus$ expr2
4) logical negation of expr1: $\neg$expr1
5) expr1 implies expr2: expr1 $\rightarrow$ expr2
6) expr1 if and only if expr2: expr1 $\leftrightarrow$ expr2
7) logical universal quantification: $\exists$x
8) logical existential quantification: $\forall$x

———
Here is an example:
Query:
Rina is either a person who jokes about being addicted to caffeine or is unaware that caffeine is a drug.
If Rina is either a person who jokes about being addicted to caffeine and a person who is unaware that caffeine is a drug, or neither a person who jokes about being addicted to caffeine nor a person who is unaware that caffeine is a drug, then Rina jokes about being addicted to caffeine and regularly drinks coffee.
###
Predicates:
Drinks(x) ::: x regularly drinks coffee.
Jokes(x) ::: x jokes about being addicted to caffeine.
Unaware(x) ::: x is unaware that caffeine is a drug.
Conclusion:
Jokes(rina) $\oplus$ Unaware(rina) ::: Rina is either a person who jokes about being addicted to caffeine or is unaware that caffeine is a drug.
((Jokes(rina) $\wedge$ Unaware(rina)) $\oplus$ $\neg$(Jokes(rina) $\vee$ Unaware(rina))) $\rightarrow$ (Jokes(rina) $\wedge$ Drinks(rina)) ::: If Rina is either a person who jokes about being addicted to caffeine and a person who is unaware that caffeine is a drug, or neither a person who jokes about being addicted to caffeine nor a person who is unaware that caffeine is a drug, then Rina jokes about being addicted to caffeine and regularly drinks coffee.

———
Here is an example:
Query:
Miroslav Venhoda loved music.
A Czech person wrote a book in 1946.
No choral conductor specialized in the performance of Renaissance.
###
Predicates:
Czech(x) ::: x is a Czech person.
ChoralConductor(x) ::: x is a choral conductor.
Author(x, y) ::: x is the author of y.
Book(x) ::: x is a book.
Specialize(x, y) ::: x specializes in y.
Conclusion:
Love(miroslav, music) ::: Miroslav Venhoda loved music.
$\exists$y $\exists$x (Czech(x) $\wedge$ Author(x, y) $\wedge$ Book(y) $\wedge$ Publish(y, year1946)) ::: A Czech person wrote a book in 1946.
$\neg$ $\exists$x (ChoralConductor(x) $\wedge$ Specialize(x, renaissance)) ::: No choral conductor specialized in the performance of Renaissance.

———
Below is the one you need to translate:
Query:
%QUERY%
———-

Prompt of NL-FOL Translation for Documents

Given a document. The task is to parse the document into first-order logic formulars. The grammar of the first-order logic formular is defined as follows:
1) logical conjunction of expr1 and expr2: expr1 $\wedge$ expr2
2) logical disjunction of expr1 and expr2: expr1 $\vee$ expr2
3) logical exclusive disjunction of expr1 and expr2: expr1 $\oplus$ expr2
4) logical negation of expr1: $\neg$expr1
5) expr1 implies expr2: expr1 $\rightarrow$ expr2
6) expr1 if and only if expr2: expr1 $\leftrightarrow$ expr2
7) logical universal quantification: $\exists$x
8) logical existential quantification: $\forall$x

Here is an example:
Document:
All people who regularly drink coffee are dependent on caffeine. People either regularly drink coffee or joke about being addicted to caffeine. No one who jokes about being addicted to caffeine is unaware that caffeine is a drug. Rina is either a student and unaware that caffeine is a drug, or neither a student nor unaware that caffeine is a drug. If Rina is not a person dependent on caffeine and a student, then Rina is either a person dependent on caffeine and a student, or neither a person dependent on caffeine nor a student.
###
Predicates:
Dependent(x) ::: x is a person dependent on caffeine.
Drinks(x) ::: x regularly drinks coffee.
Jokes(x) ::: x jokes about being addicted to caffeine.
Unaware(x) ::: x is unaware that caffeine is a drug.
Student(x) ::: x is a student.
Conclusion:
$\forall$x (Drinks(x) $\rightarrow$ Dependent(x)) ::: All people who regularly drink coffee are dependent on caffeine.
$\forall$x (Drinks(x) $\oplus$ Jokes(x)) ::: People either regularly drink coffee or joke about being addicted to caffeine.
$\forall$x (Jokes(x) $\rightarrow$ $\neg$Unaware(x)) ::: No one who jokes about being addicted to caffeine is unaware that caffeine is a drug.
(Student(rina) $\wedge$ Unaware(rina)) $\oplus$ $\neg$(Student(rina) $\vee$ Unaware(rina)) ::: Rina is either a student and unaware that caffeine is a drug, or neither a student nor unaware that caffeine is a drug.
$\neg$(Dependent(rina) $\wedge$ Student(rina)) $\rightarrow$ (Dependent(rina) $\wedge$ Student(rina)) $\oplus$ $\neg$(Dependent(rina) $\vee$ Student(rina)) ::: If Rina is not a person dependent on caffeine and a student, then Rina is either a person dependent on caffeine and a student, or neither a person dependent on caffeine nor a student.

Here is an example:
Document:
Miroslav Venhoda was a Czech choral conductor who specialized in the performance of Renaissance and Baroque music. Any choral conductor is a musician. Some musicians love music. Miroslav Venhoda published a book in 1946 called Method of Studying Gregorian Chant.
###
Predicates:
Czech(x) ::: x is a Czech person.
ChoralConductor(x) ::: x is a choral conductor.
Musician(x) ::: x is a musician.
Love(x, y) ::: x loves y.
Author(x, y) ::: x is the author of y.
Book(x) ::: x is a book.
Publish(x, y) ::: x is published in year y.
Specialize(x, y) ::: x specializes in y.
Conclusion:
Czech(miroslav) $\wedge$ ChoralConductor(miroslav) $\wedge$ Specialize(miroslav, renaissance) $\wedge$ Specialize(miroslav, baroque) ::: Miroslav Venhoda was a Czech choral conductor who specialized in the performance of Renaissance and Baroque music.
$\exists$x (ChoralConductor(x) $\rightarrow$ Musician(x)) ::: Any choral conductor is a musician.
$\forall$x (Musician(x) $\wedge$ Love(x, music)) ::: Some musicians love music.
Book(methodOfStudyingGregorianChant) $\wedge$ Author(miroslav, methodOfStudyingGregorianChant) $\wedge$ Publish(methodOfStudyingGregorianChant, year1946) ::: Miroslav Venhoda published a book in 1946 called Method of Studying Gregorian Chant.

Below is the one you need to translate:
Document:
%DOCUMENT%

Prompt of Query Generation for A - a

Given an example in information retrieval tasks. We refer to the query as a negative-constraint query. The query matches formulation **A - a**. **A** denotes Ginsberg's works and **a** denotes 'Howl'. The positive document mentions Ginsberg's works but does not mention 'Howl'. The negative document mentions Ginsberg's works and 'Howl'. Please provide a query based on the positive and negative documents provided.

EXAMPLE
Positive document:
Ginsberg took part in decades of non-violent political protest against everything from the Vietnam War to the War on Drugs. His poem ""September on Jessore Road"", calling attention to the plight of Bangladeshi refugees, exemplifies what the literary critic Helen Vendler described as Ginsberg's tireless persistence in protesting against ""imperial politics, and persecution of the powerless."" His collection ""The Fall of America"" shared the annual U.S. National Book Award for Poetry in 1974. In 1979 he received the National Arts Club gold medal and was inducted into the American Academy and Institute of Arts and Letters. Ginsberg was a Pulitzer. by gambling, and the cost of secondary education for Poe. He attended the University of Virginia but left after a year due to lack of money. Poe quarreled with Allan over the funds for his education and enlisted in the Army in 1827 under an assumed name. It was at this time that his publishing career began, albeit humbly, with the anonymous collection ""Tamerlane and Other Poems"" (1827), credited only to ""a Bostonian"". With the death of Frances Allan in 1829, Poe and Allan reached a temporary rapprochement. However, Poe later failed as an officer cadet at West Point, declaring...
Negative document:
Kerouac and William S. Burroughs. Ginsberg is best known for his poem ""Howl"", in which he denounced what he saw as the destructive forces of capitalism and conformity in the United States. In 1956, ""Howl"" was seized by San Francisco police and US Customs. In 1957, it attracted widespread publicity when it became the subject of an obscenity trial, as it described heterosexual and homosexual sex at a time when sodomy laws made homosexual acts a crime in every U.S. state. ""Howl"" reflected Ginsberg's own homosexuality and his relationships with a number of men, including Peter Orlovsky, his lifelong partner.
Query:
Introduce Allen Ginsberg's works, but do not mention 'Howl'.

Positive document: %POSITIVE DOCUMENT%
Negative document: %NEGATIVE DOCUMENT%

Below query is the one you need to generate, which make significant changes to the query style.
Query:
%QUERY%

Prompt of Query Generation (A - a) ∪ B

Given an example in information retrieval tasks. We refer to the query as a negative-constraint query. The query matches formulation **(A - a)** ∪ **B**. **A** denotes Allen Ginsberg's works, **a** denotes 'Howl', and **B** denotes Edgar Allan Poe's works. The positive document mentions Allen Ginsberg's and Edgar Allan Poe's works but does not mention 'Howl'. The negative document mentions Allen Ginsberg's works, Edgar Allan Poe's works and 'Howl'. Please provide a query based on the positive and negative documents provided.

EXAMPLE
Positive document:
Ginsberg took part in decades of non-violent political protest against everything from the Vietnam War to the War on Drugs. His poem ""September on Jessore Road"", calling attention to the plight of Bangladeshi refugees, exemplifies what the literary critic Helen Vendler described as Ginsberg's tireless persistence in protesting against ""imperial politics, and persecution of the powerless."" His collection ""The Fall of America"" shared the annual U.S. National Book Award for Poetry in 1974. In 1979 he received the National Arts Club gold medal and was inducted into the American Academy and Institute of Arts and Letters. Ginsberg was a Pulitzer. by gambling, and the cost of secondary education for Poe. He attended the University of Virginia but left after a year due to lack of money. Poe quarreled with Allan over the funds for his education and enlisted in the Army in 1827 under an assumed name. It was at this time that his publishing career began, albeit humbly, with the anonymous collection ""Tamerlane and Other Poems"" (1827), credited only to ""a Bostonian"". With the death of Frances Allan in 1829, Poe and Allan reached a temporary rapprochement. However, Poe later failed as an officer cadet at West Point, declaring...
Negative document:
Kerouac and William S. Burroughs. Ginsberg is best known for his poem ""Howl"", in which he denounced what he saw as the destructive forces of capitalism and conformity in the United States. In 1956, ""Howl"" was seized by San Francisco police and US Customs. In 1957, it attracted widespread publicity when it became the subject of an obscenity trial, as it described heterosexual and homosexual sex at a time when sodomy laws made homosexual acts a crime in every U.S. state. ""Howl"" reflected Ginsberg's own homosexuality and his relationships with a number of men, including Peter Orlovsky, his lifelong partner. by gambling, and the cost of secondary education for Poe. He attended the University of Virginia but left after a year due to lack of money. Poe quarreled with Allan over the funds for his education and enlisted in the Army in 1827 under an assumed name. It was at this time that his publishing career began, albeit humbly, with the anonymous collection ""Tamerlane and Other Poems"" (1827), credited only to ""a Bostonian"". With the death of Frances Allan in 1829, Poe and Allan reached a temporary rapprochement. However, Poe later failed as an officer cadet at West Point, declaring...
Query:
What themes are expressed in Allen Ginsberg's works (excluding "Howl")? Is there any similarity between Edgar Allan Poe's works and theirs?

Positive document: %POSITIVE DOCUMENT%
Negative document: %NEGATIVE DOCUMENT%

Below query is the one you need to generate, which make significant changes to the query style.
Query:
%QUERY%

Prompt of Query Generation (A - a) ∪ (B - b)

Given an example in information retrieval tasks. We refer to the query as a negative-constraint query. The query matches formulation **(A - a)** ∪ **(B - b)**. **A** denotes Ginsberg's works, **a** denotes 'Howl', **B** denotes Poe's works and **b** denotes 'The Raven'. The positive document mentions Ginsberg's and Poe's works but does not mention 'Howl' and 'The Raven'. The negative document 1 mentions Ginsberg's works, Poe's works' and 'Howl'. The negative document 2 mentions Ginsberg's works, Poe's works' and 'The Raven'. The negative document 3 mentions Ginsberg's works, Poe's works', 'Howl' and 'The Raven'. Please provide a query based on the positive and negative documents provided.

EXAMPLE
Positive document:
Ginsberg took part in decades of non-violent political protest against everything from the Vietnam War to the War on Drugs. His poem ""September on Jessore Road"", calling attention to the plight of Bangladeshi refugees, exemplifies what the literary critic Helen Vendler described as Ginsberg's tireless persistence in protesting against ""imperial politics, and persecution of the powerless."" His collection ""The Fall of America"" shared the annual U.S. National Book Award for Poetry in 1974. In 1979 he received the National Arts Club gold medal and was inducted into the American Academy and Institute of Arts and Letters. Ginsberg was a Pulitzer. produce his own journal ""The Penn"" (later renamed ""The Stylus""), though he died before it could be produced. by gambling, and the cost of secondary education for Poe. He attended the University of Virginia but left after a year due to lack of money. Poe quarreled with Allan over the funds for his education and enlisted in the Army in 1827 under an assumed name. It was at this time that his publishing career began, albeit humbly, with the anonymous collection ""Tamerlane and Other Poems"" (1827), credited only to ""a Bostonian"". With the death of Frances Allan in 1829, Poe and Allan reached a temporary rapprochement. However, Poe later failed as an officer cadet at West Point, declaring...
Negative document 1:
Kerouac and William S. Burroughs. Ginsberg is best known for his poem ""Howl"", in which he denounced what he saw as the destructive forces of capitalism and conformity in the United States. In 1956, ""Howl"" was seized by San Francisco police and US Customs. In 1957, it attracted widespread publicity when it became the subject of an obscenity trial, as it described heterosexual and homosexual sex at a time when sodomy laws made homosexual acts a crime in every U.S. state. ""Howl"" reflected Ginsberg's own homosexuality and his relationships with a number of men, including Peter Orlovsky...
Negative document 2:
Ginsberg took part in decades of non-violent political protest against everything from the Vietnam War to the War on Drugs. His poem ""September on Jessore Road"", calling attention to the plight of Bangladeshi refugees, exemplifies what the literary critic Helen Vendler described as Ginsberg's tireless persistence in protesting against ""imperial politics, and persecution of the powerless."" His collection ""The Fall of America"" shared the annual U.S. National Book Award for Poetry in 1974. In 1979 he received the National Arts Club gold medal and was inducted into the American Academy and Institute of Arts and Letters. Ginsberg was a Pulitzer. produce his own journal ""The Penn"" (later renamed ""The Stylus""), though he died before it could be produced. a firm wish to be a poet and writer...
Negative document 3:
Kerouac and William S. Burroughs. Ginsberg is best known for his poem ""Howl"", in which he denounced what he saw as the destructive forces of capitalism and conformity in the United States. In 1956, ""Howl"" was seized by San Francisco police and US Customs. In 1957, it attracted widespread publicity when it became the subject of an obscenity trial, as it described heterosexual and homosexual sex at a time when sodomy laws made homosexual acts a crime in every U.S. state. ""Howl"" reflected Ginsberg's own homosexuality and his relationships with a number of men, including Peter Orlovsky...
Query:
What are the similarities between Ginsberg's works (excluding 'Howl') and Poe's works (excluding 'The Raven')?

Positive document: %POSITIVE DOCUMENT%
Negative document 1: %NEGATIVE DOCUMENT 1%
Negative document 2: %NEGATIVE DOCUMENT 2%
Negative document 3: %NEGATIVE DOCUMENT 3%

Below query is the one you need to generate, which make significant changes to the query style.
Query:
%QUERY%

C Data Collection and Snippet

We use the introductory passages from Wikipedia dump as the corpus, as they are usually high-quality and contain most of the key information. We request three carefully selected experienced annotators to filter passages from Wikipedia. For queries of formulation **A - a** and **(A - a)** $\cup$ **B**, each positive and negative document corresponding to a query is composed of one passage from the Wikipedia dump, respectively. However, due to the complexity of formulation **(A - a)** $\cup$ **(B - b)**, we merge two passages that belong to one topic as a positive document, and the two merged passages are highly relevant. Negative documents for formulation **(A - a)** $\cup$ **(B - b)** are also obtained in this way. Then we prompt GPT-4o to generate queries based on the positive and negative documents. We ensure that the queries are as style-diverse as possible. That is, we do not just perform entity replacement, but pay more attention to the diversity of query mode. For example, there are three queries expressing negative constraints for formulation **A - a**:

1. *Investigate the role of nature in Walden, excluding Thoreau's critique of society.*
2. *Introduce the works of Emily Dickinson, but do not mention 'Because I could not stop for Death'.*
3. *Without referencing Victor Frankenstein's use of scientific knowledge, examine the role of technology in Frankenstein.*

Finally, annotators also select several irrelevant passages with queries to fill into the corpus.

Table 6 introduce snippets of NegConstraint dataset. Entities marked in red and green denote entities in negative-constraint conditions. For formulation **A - a**, the negative document mentions "Howl". For formulation **(A - a)** $\cup$ **B**, the negative document mentions "Howl". For formulation **(A - a)** $\cup$ **(B - b)**, the negative document 1 mentions "Howl", the negative document 2 mentions "The Raven", and the negative document 3 mentions "Howl" and "The Raven".

A - a	Query	Introduce Allen Ginsberg's works, but do not mention ' Howl'.
	Postive document	. . . His poem 'September on Jessore Road', calling attention to the plight of Bangladeshi refugees, exemplifies what the literary critic Helen Vendler described as Ginsberg's tireless persistence in protesting against ""imperial politics, and persecution of the powerless. . .
	Negative document	In 1956, 'Howl' was seized by San Francisco police and US Customs...'Howl' reflected Ginsberg's own homosexuality and his relationships with a number of men, including Peter Orlovsky, his lifelong partner. . .
(A - a) ∪ B	Query	What themes are expressed in Allen Ginsberg's works other than 'Howl' and Edgar Allan Poe's works?
	Postive document	Ginsberg took part in decades of non-violent political protest against everything from the Vietnam War to the War on Drugs. His poem ""September on Jessore Road"", calling attention to the plight of Bangladeshi refugees, exemplifies what the literary critic Helen Vendler described as Ginsberg's tireless persistence in protesting against ""imperial politics, and persecution of the powerless. . .
	Negative document	Kerouac and William S. Burroughs. Ginsberg is best known for his poem 'Howl', in which he denounced what he saw as the destructive forces of capitalism and conformity in the United States. In 1956, 'Howl' was seized by San Francisco police and US Customs. . . .
(A - a) ∪ (B - b)	Query	What themes do Allen Ginsberg's works other than 'Howl' and Edgar Allan Poe's works other than 'The Raven' express?
	Postive document	Ginsberg took part in decades of non-violent political protest against everything from the Vietnam War to the War on Drugs..., His collection 'The Fall of America' shared the annual U.S. National Book Award for Poetry in 1974. In 1979 he received the National Arts Club gold medal and was inducted into the American Academy and Institute of Arts and Letters...,credited only to 'a Bostonian'. With the death of Frances Allan in 1829, Poe and Allan reached a temporary rapprochement. However, Poe later failed as an officer cadet at West Point, declaring...
	Negative document 1	Kerouac and William S. Burroughs. Ginsberg is best known for his poem 'Howl', in which he denounced what he saw as the destructive forces of capitalism and conformity in the United States...'Howl' reflected Ginsberg's own homosexuality and his relationships with a number of men, including Peter Orlovsky, his lifelong partner. by gambling, and the cost of secondary education for Poe...
	Negative document 2	. . . In January 1845, Poe published his poem 'The Raven' to instant success. His wife died of tuberculosis two years after its publication. For years, he had been planning to...
	Negative document 3	In 1956, 'Howl' was seized by San Francisco police and US Customs... 'Howl' reflected Ginsberg's own homosexuality and his relationships with a number of men, including Peter Orlovsky, his lifelong partner. a firm wish to be a poet and writer, and he ultimately parted ways with John Allan...In January 1845, Poe published his poem 'The Raven' to instant success. His wife died of tuberculosis two years after its publication. For years, he had been planning to

Table 6: Snippets of NegConstraint dataset. Entities marked in red and green denote negative-constraint conditions.