Article:
Preiss, J. orcid.org/0000-0002-2158-5832 (2025) A hybrid approach to literature-based discovery: combining traditional methods with LLMs. Applied Sciences, 15 (16). 8785. ISSN: 2076-3417

https://doi.org/10.3390/app15168785

applied sciences

Article

A Hybrid Approach to Literature-Based Discovery: Combining Traditional Methods with LLMs

Judita Preiss

Information School, University of Sheffield, Sheffield S10 2AH, UK; judita.preiss@sheffield.ac.uk

Abstract

We present a novel hybrid approach to literature-based discovery (LBD) which exploits large language models (LLMs) to enhance traditional LBD methodologies. We explore the use of LLMs to address significant LBD challenges: (1) the extraction of factual subject–predicate–object relations from publication abstracts using few-shot learning and (2) the filtering of unpromising candidate hidden knowledge pairs (CHKPs) using a variant of the LLM-as-a-judge paradigm with and without the addition of domain-specific information using retrieval augmented generation. The approach produces relations with greater coverage and results in a lower number of CHKPs compared to LBD based on relations extracted with, e.g., SemRep, improving the prediction and efficiency of knowledge discovery. We demonstrate the utility of the method using a drug-repurposing case study and suggest that emerging AI technologies can be used to assist in knowledge discovery from the ever-growing volume of the scientific literature.

Keywords: literature-based discovery; large language models; relation extraction; retrieval augmented generation

1. Introduction

The increasing volume of the scientific literature presents a challenge for knowledge discovery. The manual synthesis of information cannot keep up with the rapid growth of publications across diverse fields, leading to missed connections. Literature-based discovery (LBD) is a data mining approach capable of connecting disparate literature studies [1] and uncovering new connections within specialized fields, addressing situations where the volume of publications cannot be processed by a human (e.g., [2,3]). The original LBD technique, the ABC model [4]—due to the connection between A and C it proposes based on known relations between A and B and B and C—hinges on the accuracy and quality of the relations extracted from texts, as well as the ability to filter the resulting candidate hidden knowledge pairs (CHKPs) to remove uninformative candidates [5]. This work introduces a hybrid data analysis approach, which explores the use of large language models (LLMs) to address both of these points.

Early LBD works utilised simple text mining techniques, such as word co-occurrence in titles, to define relations [6]. This method was effective for investigating suspected (closed) connections where the literature could be reduced to titles containing A or C and only the linking term B was sought. However, even when the relations were refined, the approach had a tendency to generate an overwhelming number of CHKPs when used in open mode, where all connections from every term are followed for two steps [7].

The need to reduce the quantity of CHKPs generated by open LBD has fuelled research into enhancements to existing methods for effective data mining. Prior approaches have

Academic Editor: Ming Liu

Received: 20 July 2025
Revised: 3 August 2025
Accepted: 5 August 2025
Published: 8 August 2025

Citation: Preiss, J. A Hybrid Approach to Literature-Based Discovery: Combining Traditional Methods with LLMs. *Appl. Sci.* **2025**, *15*, 8785. https://doi.org/10.3390/app15168785

included the following: (a) the removal of terms based on frequency (e.g., [8,9]), (b) the mapping of terms to specialized lexicons such as the Unified Medical Language System (UMLS) [10] to prevent connections via ambiguous terms [11], (c) the removal of terms based on semantic type (e.g., [12,13]), (d) the restriction of the type of discovery (e.g., cancer biology [14] or protein interaction [15]), (e) the restriction of relations to be connected [16], and (f) the re-ranking of generated CHKPs (e.g., [17–21]). However, the exponential growth of academic publications [22] requires a constant refinement of these approaches.

Recent advances in deep learning have improved the performance of many natural language processing (NLP) tasks in the biomedical field [23], including LBD (for an overview, see [24]). While pre-trained language models have been used for LBD [25], many successful applications of data mining, particularly in areas such as drug repurposing, have predominantly relied on knowledge graph-based deep learning, with the knowledge provided by existing tools, such as SemMedDB [26] (e.g., [3]), the Global Network of Biomedical Relationships [27] (e.g., [28]), or named entity extraction (e.g., [29]). The use of large, generative, language models (LLMs) has, however, been relatively low for the purpose of LBD since they have a propensity for adhering to information they were trained on rather than proposing new connections [30].

Through the vast quantity of training data, LLMs have access to an enormous amount of knowledge. While this appears to be a limitation when proposing new connections, it suggests their suitability for use within CHKP filtering. Since a large quantity of proposed CHKPs represent background, well-known knowledge (which, due to being widely familiar, does not appear explicitly in publications [25]), we propose that LLMs could be used to filter these CHKPs. By instructing them to act as judges—a modification of the LLM-as-a-judge paradigm—they can assess each CHKP against their training data to determine whether it represents generally known, background knowledge. This theoretically distinct approach goes beyond conventional filtering methods that rely on frequency or semantic constraints.

In this work, we therefore present a novel data mining approach that extensively explores the integration of machine learning and AI through LLMs for LBD, involving them in every crucial step except the initial CHKP generation:

1. Using few-shot learning to extract factual subject–predicate–object relations from publication abstracts, achieving greater coverage than established tools such as SemRep.
2. Investigating the impact of different training examples within few-shot learning in Step 1, comparing manually annotated instances with examples derived from cited facts.
3. Using zero-shot learning for filtering background knowledge CHKPs, in an LLM-as-a-judge setup, with hallucinations ruled out using retrieval augmented generation.

The replication of existing discoveries is used to demonstrate the usefulness of the LLM-generated relations. A small-scale timeslicing evaluation indicates their superior suitability for LBD, producing less background knowledge and achieving higher precision against the gold standard than LBD based on SemRep relations.

The remainder of this paper is structured as follows: Section 2 describes related work, while Section 3 outlines the experiments carried out in this work. Discussions are presented in Section 4 with conclusions and future work appearing in Section 5.

2. Related Work

The growing quantity of the scientific literature indicates that effective knowledge discovery techniques, such as literature-based discovery (LBD), are required. The original LBD approach, the ABC model [6], uses simple inference to propose a connection between previously unconnected terms A and C if there are known, published, connections from A to B and from B to C, which appear in separate publications. Its functionality was initially demonstrated on a suspected connection between *Raynaud's syndrome*and *fish oil*—both *Raynaud's syndrome* and *fish oil* were found to have already published links to *blood viscosity*, thereby allowing a connection to be confirmed through this (*B*) term.

The number of practical applications of LBD is growing: within the biomedical field, this includes, for example, adverse drug event prediction [31], drug development [32] and drug repurposing [33]. Particularly in the latter case, the economic benefits are substantial: a drug that has already been safety tested can bypass 6–7 years of preclinical and early-stage research when being investigated for a new use [34]. While the novel data analysis approach presented in this paper is applicable to all uses of LBD, our evaluations specifically focus on its utility in drug repurposing.

2.1. Resources

Several resources are frequently employed in large-scale text and data mining for LBD. These will be introduced first, as they are required for relation extraction (Section 2.2) and subsequently LBD (Section 2.3).

2.1.1. MEDLINE and PubMed

PubMed, and its biomedical subset MEDLINE, are the National Library of Medicine's (NLM's) databases of publications, on which many LBD investigations are based. MEDLINE mainly contains paper titles and abstracts from the biomedical domain, while PubMed offers broader domain coverage and more recently also contains articles' full texts. Although some works utilise full texts (e.g., [35]), the majority of LBD systems exploit titles and/or abstracts only.

2.1.2. UMLS

The Unified Medical Language System (UMLS) [10] is another widely used resource. It comprises the metathesaurus, the semantic network and the SPECIALIST lexicon and tools.

2.1.3. UMLS Metathesaurus

The UMLS metathesaurus unifies a number of source vocabularies into a single database of biomedical and health-related concepts. The concepts, identified using a concept unique identifier (CUI), link together different ways to refer to the same concept in a natural text. A number of manually identified relationships between CUIs (related concepts) are also included within the metathesaurus. The main use of the metathesaurus in LBD is for disambiguation: if terms are accurately mapped to their CUIs, connections made through terms with identical spelling but different meaning can be avoided [36]. In addition, performing this mapping can link abbreviations to their long form.

2.1.4. The Semantic Network

The Semantic Network groups CUIs together into broad categories known as semantic types with each UMLS concept assigned at least one semantic type. This semantic typing is frequently used to constrain the hidden connections proposed by LBD, for example, to drug (*Chemicals & drugs*)–treatment (*Disease or Syndrome*) pairs (e.g., [37]).

2.1.5. MetaMap

MetaMap [38], a tool within the UMLS suite, is widely used for mapping biomedical text to UMLS concepts. Its availability for local execution and regular releases of preprocessed versions of MEDLINE by the NLM significantly aid large-scale biomedical text mining. In this work, we utilize version 24 of MetaMapped MEDLINE, and MetaMap 2020 with the 2020AA USAbase strict data model is used for local processing.

2.2. Relation Extraction

The extraction of reliable semantic relations is an important component of the ABC model (e.g., [39] or [40]). High-quality relations enable further refinements, such as focusing on specific types of interactions (for example, relations between chemicals, genes and diseases as explored by [27]). Also, the availability of relations allows the construction of large-scale knowledge graphs, which can be exploited by LBD (e.g., [41]).

2.2.1. SemRep and Refinements

SemRep is a widely used rule-based approach to semantic relation extraction tuned to the biomedical domain. Built upon MetaMap, it incorporates a step for mapping concepts to UMLS CUIs. For example, for the sentence in list 1, SemRep extracts the semantic relations in list 2 (CUIs have been mapped to their term representation).

1. *Raynaud's phenomenon (RP) is commonly observed in fingers and toes of patients with connective tissue diseases (CTDs).*
2. *Connective Tissue Diseases* PROCESS_OF *Patients.*
   Toes PART_OF *Patients.*
   Fingers LOCATION_OF *Raynaud Phenomenon.*
   Toes LOCATION_OF *Raynaud Phenomenon.*

The tool extracts approximately 70 different predications, with roughly half being negative (such as NEG_TREATS). A database of SemRep processed abstracts from MEDLINE, SemMedDB [26], is publicly available and widely used for biomedical data analysis, including LBD (e.g., [33] or [42]). However, a recent evaluation of SemRep (version 1.8) on the SemRep test collection [43] revealed limitations in its performance, with 0.55 precision, 0.34 recall, and 0.41 F_1 [44]. Even after accounting for test collection and evaluation setup issues, the recall remained low at 0.42 (with an F_1 of 0.52), suggesting that the technique requires enhancement for effective data mining.

To address the recall limitations of SemRep, machine learning approaches, such as classification, have been explored [45]. This involved fine-tuning language models such as PubMedBERT (now BiomedBERT) [46] on entities extracted by SemRep following a pre-training step using contrastive learning. This approach was found to be complementary to SemRep's annotations, with 0.81 recall, 0.62 precision and 0.70 F_1.

Beyond rule-based systems such as Public Knowledge Discovery Engine for Java (PKDE4J) [47] and its transformer-based refinement BioPREP [48], other applications of transformer techniques to relation extraction, often focusing on specific subsets of relations, have also emerged (e.g., [49,50]). However, fine-tuning pre-trained transformer models typically requires a substantial annotated datasets (e.g., [3]), and the resulting relations are not always optimally suited for LBD.

Recently, there has been a shift towards the use of large language models (LLMs) for relation extraction. Given the enormous quantity of data used to pre-train LLMs, relations can be extracted in a zero-shot setting—with no examples provided in the prompt [51]—or in a few-shot setting, with a small number of examples included in the prompt [52]. However, the specific focus of each work (e.g., clinical trials only) and varied different evaluation datasets make direct performance comparisons difficult across different LLM-based approaches.

2.2.2. Factuality Dataset

The publicly available FACTUALITY dataset allows the confidence level of SemRep-generated relations to be assessed. The distinction between genuine facts, conjectures, or doubtful statements is valuable for high-quality input to LBD (e.g., [53]). Kilicoglu et al. [54] manually annotated SemRep relations extracted from 500 PubMed abstracts with one of seven factuality values (FACT, PROBABLE, POSSIBLE, DOUBTFUL, COUNTERFACT, UNCOMMITTED, and CONDITIONAL), providing a valuable resource for refining relation extraction (available from https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR.html, accessed 1 August 2024).

2.3. Literature-Based Discovery

The original ABC approach to LBD was integrated into a number of larger-scale systems, such as BITOLA [55], Arrowsmith [56] and FACTA+ [57]. In their open form, these systems often generated an excessively large number of candidate hidden knowledge pairs (CHKPs), making the manual identification of promising avenues an extremely laborious task. To reduce the quantity of CHKPs, co-occurrence was replaced with semantic relations, and the semantic types of the terms involved in connections were restricted [58]. Further refinements involved the use of graph-based approaches for inference, based on knowledge graphs build from semantic relations extracted from publications (e.g., [19]). Despite these refinements, even targeted LBD systems can still yield a large quantity of links; for example, Syafiandini et al. [59] found 2,740,314 paths between 35 Xanthium compounds and three types of diabetes.

Recent advances in deep learning have significantly improved the state of the art across many biomedical domain tasks (as noted in Section 2.2.1). Many applications of deep learning to LBD exploit the graph structure of extracted relations, whether by using language models to enhance similarity in graph construction [60] or for filtering relations within the resulting graph [3]. Other neural network approaches, such as CNNs [61] and autoencoders [62], have also been investigated for their utility in LBD.

There is also a clear utility of LLMs within the biomedical domain, for example, by exploiting their ability to answer scientific questions [63] or aiding clinical decision-making [64]. LLMs boast a number of advantages over traditional approaches: most importantly, being based on a massive dataset, they do not require structured inputs and are less susceptible to the scalability issues experienced by simple inference. However, so far, works exploiting LLMs for direct hidden knowledge suggestion in LBD have indicated that the proposed hidden knowledge is of low technical depth and novelty [65], frequently reverting to well-established information [66]. This observation suggests that it is important to use LLMs where they excel: we therefore propose a hybrid approach, using LLMs to extract knowledge from texts and to identify background knowledge among the proposed CHKPs, while employing the established ABC model for CHKP generation.

2.3.1. Filtering

The effective filtering or re-reranking of the CHKPs resulting from LBD determines its usefulness. We can broadly categorize valid relation-based CHKPs into three groups: (1) connections depicting background knowledge—information that is so widely known that it is unlikely to appear in knowledge bases; (2) minor variations of existing knowledge, e.g., involving synonyms or already closely linked studies such as *fish oil—Raynaud phenomenon* deduced from a known link between *fish oil* and *Raynaud disease*; and (3) genuinely valuable CHKPs, such as connections between disjointed studies [67].

Many filtering approaches do not target CHKPs by type. For example, the authors of [3] reduce the set of predicate types within their LBD system to 15 based on their expected utility. Similarly, restricting the semantic types of subjects (source) and objects (target) of semantic relations (e.g., [12,13,37]) focuses on the application domain rather than eliminating specific CHKP groups. While these methods reduce the overall quantity of CHKPs, type 1 and 2 CHKPs are still prevalent in the remaining lists.

2.3.2. Filtering Using Generative AI

The large quantity of data that LLMs are pre-trained with enables them to be used for content evaluation and filtering, including tasks such as removing noisy data from datasets (e.g., [68]). However, the application of LLMs for CHKP filtering in LBD has not yet been explored. Within the LLM-as-a-judge paradigm, an LLM is used to evaluate the output of other LLMs, checking their agreement with human annotation, yielding agreement levels of human experts [69]. We propose modifying this paradigm to utilise an LLM to evaluate the proposed, automatically created CHKPs—which were designed to represent novel hypotheses—to determine whether they represent valuable hidden knowledge or already known background information.

2.3.3. Evaluation of LBD

The evaluation of CHKPs is difficult as there is, by definition, no gold standard available ([70,71]). Generally, studies therefore perform one of three evaluations [5]:

1. Expert and clinical evaluation, where an expert examines the CHKPs proposed by the LBD system. Such connections are usually proposed by running LBD in a heavily restricted mode to account for experts' specializations.
2. The replication of existing discoveries, where a new system is judged on its ability to replicate discoveries made by past LBD systems.
3. Timeslicing, in which a chronological cut-off date is selected with publications prior to the cut-off point used by the LBD model to generate CHKPs, which are then evaluated against relations extracted from publications after the cut-off point [72].

It should be noted that the number of past discoveries is limited, restricting this evaluation type [73] and potentially introducing biases [74]. However, the timeslicing approach, which—unlike expert evaluation—allows for large-scale evaluation, also suffers from a number of limitations: (1) not all valid CHKPs will have been discovered (yet), and (2) the evaluation is heavily reliant on the accuracy of relations extracted from the literature post cut-off.

3. Methodology

This section describes the novel approaches explored in this work: (1) the use of FSL-based relation extraction is compared to SemRep triples using an application (ABC LBD model), and (2) LLMs are explored for the identification of already known CHKPs. The individual experiments are described below.

3.1. Pre-Inference Filtering

To standardize representations and provide access to semantic-type information, concepts are mapped to the UMLS. Given our focus on drug repurposing, source term semantic types are restricted to *chemicals & drugs* and *genes & molecular sequence* and target terms to *disease or syndrome*. This filtering step is applied before inference to increase speed, meaning only relations with source (subject) semantic type concepts or target (object) semantic types proceed to inference.

3.2. Semantic Relations

Since accurate semantic relation extraction is crucial for LBD, we explore three separate approaches for this text mining task.

3.2.1. SemMedDB Semantic Relations

SemMedDB version `semmedVER43_2024_R`, used in this work, contains 130,480,195 distinct predications extracted from 37,233,341 articles using a rule-based approach. Removing predications based on frequency and their likelihood of being helpful for LBD leaves the following 5 most represented relations: TREATS, AFFECTS, AUGMENTS, PRODUCES and STIMULATES.

3.2.2. Factuality Dataset-Based Relation Extraction

To encourage the extraction of factual relations, FACT relations from the manually annotated FACTUALITY dataset (Section 2.2.2) are used to provide examples to an LLM performing relation extraction. Examples were sampled from relations remaining after removing (1) relations containing a null subject or object, (2) duplicate relations from the same abstract, (3) relations which did not include the 5 most represented relations, and (4) relations which did not contain the desired semantic types (Section 3.2.1). This dataset is referred to as `fact_triple`.

It is important to note that a single abstract can contain more than one factual triple. Due to this, the number of shots may be lower than the number of factual triples included in a prompt since the number of shots refers to the number of abstracts contained in the prompt. Additionally, the potential of more than one factual triple per abstract means that balancing over predicates would not be straightforward, and a different technique would need to be employed for the selection of examples from the S2ORC data (Section 3.2.3). The LLM's ability to generalize, along with the prevalence of TREATS relations in the FACTUALITY dataset (in order, there are 174 TREATS relations compared to 39 instances of AFFECTS), indicated the suitability of random selection of examples for few-shot learning. A fixed seed was used to ensure reproducibility and incremental building of examples (i.e., that examples for 10-shot learning add 5 abstracts to those used for 5-shot learning), with the same examples used for all models explored.

3.2.3. Cited Information-Based Relation Extraction

An alternative approach to finding examples for LLM FSL-based relation extraction is based on the hypothesis that novel contributions of publications will be cited by other researchers. Such descriptions are expected to be factual and represent significant contributions of publications. They are not restricted to a specific domain and can avoid biases that may be present when examples are selected for manual annotation.

The S2ORC corpus [75], which contains many full texts of articles, was used to access the *introduction* and *related work* sections (identified using GROBID [76]). These sections frequently contain short summaries of other academic publications alongside citations of the source—these summaries should overlap with the abstract and/or conclusion of the original work, yielding examples for FSL relation extraction. Training examples were drawn from (1) English articles (identified using the `langdetect` library), (2) articles with less than 50 references (to avoid survey papers, which may dilute specific contributions), and (3) articles with an abstract length under 500 words with a minimum of 5 words per paragraph (both being indicative of GROBID failure).

Since some of the summaries could be lengthy, with only a fragment describing the contribution of a cited work, these were parsed using the `constituent_treelib` library [77] (the `ConstituentTree.SpacyModelSize.Large` was used for greater accuracy) to extract the clause containing the reference, imposing a minimum length of 6 words on extracted clauses.

The resulting examples are formed of a sentence fragment alongside a corresponding abstract or conclusion (dataset `cited_sent`). Since these sentences may not contain a valid relation (either by virtue of not containing a valid UMLS term or not containing one of the selected predicates), a second dataset, `cited_sent_wrel`, is created comprising sentences that contain a valid SemRep relation of the desired type. A third dataset, `cited_triple`, retains only the SemRep triples alongside an abstract or conclusion. Restricting articles to the Microsoft Academic field of study (present in S2ORC) of *Medicine*, yields a second set of these datasets (`med_cited_sent`, `med_cited_sent_wrel`, `med_cited_triple`).

3.3. LLMs for Relation Extraction Using FSL

Relation extraction based on the examples identified in Sections 3.2.2 and 3.2.3 is performed using FSL: the LLM is prompted to extract novel contributions from a publication's abstract in either sentence or semantic relation form and is provided with a number of examples from the appropriate list (exact prompts used are shown in Appendices A.1 and A.2). This approach has a number of advantages:

1. Semantic relations can be extracted from publications in any domain.
2. For FSL, only a (relatively) small number of examples is required to guide the generative process.
3. The LLM can be directed to identify only relations representing novel contributions of a publication.

Furthermore, the LLM is constrained to producing TREATS, AFFECTS, AUGMENTS, PRODUCES or STIMULATES predicates (the most common and LBD-relevant predicates), which encourages the focus to be on relations likely to be involved in knowledge discovery.

Following relation extraction, the output of the LLM is automatically adjusted to fit within the expected parameters by exploring frequently occurring patterns and manually deciding whether a rewrite is appropriate for each pattern. For example, the output of a relation in the form X `PRODUCES` *a rise in* Y is automatically adjusted to X `AUGMENTS` Y. Further validation is performed by passing the output generated by the LLM through MetaMap, which converts it to a standardized form, ensuring that a valid triple is present in the output of the LLM. Table 1 summarizes the information passed to the FSL model for each approach—note that the relation/text pairs need to indicate to the LLM which piece of information is important in a paragraph of text, and therefore, it is not important whether the source of the text is an abstract or a conclusion.

Three recently publicly available models are employed: two Llama 3.1 models [78] (8B and 70B) and Cohere's CommandR+ model. These models differ in the training data and tasks used to create them, the context lengths they allow and the number of parameters they contain (see Table 2 (the `llama2-13B` model is included in the table as it is used for a later experiment)). Note that both the `llama3.1-70B` and the `commandR+` models were loaded in their quantized (4 bit) versions (the largest Llama 3.1 model, Llama 3.1 405B, exceeded available computational resources for this work). The decision to employ the quantized CommandR+ and Llama3.1-70B models was based on a preliminary small-scale experiment on a subset of the Drug–Drug Interaction benchmark data, DDIExtraction 2013, where drug–drug interactions needed to be classified into one of four different types [79]. This task was selected for its similarity to relation extraction. A 5-fold cross validation was performed with 10 shots randomly selected from the training portion and 50 randomly selected test examples used for evaluation. The results and execution time of the locally run models can be seen in Table 3. While the non-quantized Llama3.1-70B achieved the highest accuracy (61.2%), its average query execution time of 134.4 s was deemed unfeasible for a scalable deployment given our resources. The tasks used for training the models are likely to be more important for their performance and generalizability to new tasks than their knowledge cut-offs (which can be found in the final column of Table 2); however, information about neither the tasks nor details of the training data is available for any of the models, and therefore, empirical evaluations of all three models were performed on the relation extraction task described in this work.

Table 1. A summary of each extracted fact type's associated information: `med` indicates a restriction to the medical domain, while `wrel` represents sentences that contain at least one extractable SemRep semantic relation.

Input ID	Type	Source	Title	Abstract	Conclusion
fact_triple	triple	Factuality	✓	✓	×
cited_sent	sentence	S2ORC	✓	✓	✓
med_cited_sent	sentence	S2ORC	✓	✓	✓
cited_triple	triple	S2ORC+SemRep	✓	✓	✓
med_cited_triple	triple	S2ORC+SemRep	✓	✓	✓
cited_sent_wrel	sentence	S2ORC(+SemRep)	✓	✓	✓
med_cited_sent_wrel	sentence	S2ORC(+SemRep)	✓	✓	✓

Table 2. Generative AI models used for FSL.

Identifier	Huggingface Model ID	Context	Params	Cutoff
llama2-13B	meta-llama/Llama-2-13b-chat-hf	4K	13B	Sep 2022
llama3.1-8B	meta-llama/Meta-Llama-3.1-8B-Instruct	128K	8B	Dec 2023
llama3.1-70B	meta-llama/Meta-Llama-3.1-70B-Instruct	128K	70B	Dec 2023
commandR+	CohereForAI/c4ai-command-r-plus-4bit	128K	104B	Oct 2023

Table 3. Medical relation extraction on subset of DDIExtraction 2013 benchmark.

Model	Accuracy		Execution Time per Query	
	Mean	Stdev	Mean	Stdev
llama3.1-8B	24.8%	7.2%	0.15 s	0.01 s
llama3.1-70B	48.8%	14.1%	1.56 s	0.09 s
llama3.1-70B non-quantized	61.2%	7.0%	134.40 s	11.66 s
commandR+	57.6%	10.8%	2.09 s	0.10 s

All three models were evaluated on 25 randomly selected withheld examples of the FACTUALITY dataset, using a fixed seed of 234 and temperature of 0.1 (allowing some flexibility in generation). The number of shots (examples) was varied between 5 and 20 in increments of 5. To assess the semantic closeness of the LLM responses to the expected output, sentence transformer similarity [80] (`all-MiniLM-L6-v2`) was employed, a method better suited than traditional summarization metrics like the ROUGE score [81] for detecting semantically equivalent expressions [82].

Figure 1 contains the average similarity achieved by the top-performing models across the test examples, with the x-axis indicating the model name, number of shots, and training data source (as listed in Table 1). While the highest performance is based on FSL examples from the FACTUALITY dataset, the CommandR+ results using cited facts (from the medical domain with examples containing at least one SemRep relation) also demonstrate high average similarity, showing the utility of the method especially for domains where an equivalent of the FACTUALITY dataset is not available. Overall, the CommandR+ model significantly outperforms the Llama 3.1 models on this task. When exploring the reason for the difference, it was observed that the Llama 3.1 models were found to generally produce a single fact for each input, while the CommandR+ models frequently generated more than one fact. Based on these results, the 10-shot CommandR+ model using FACTUALITY-based triples was selected to extract relations for LBD in this work.

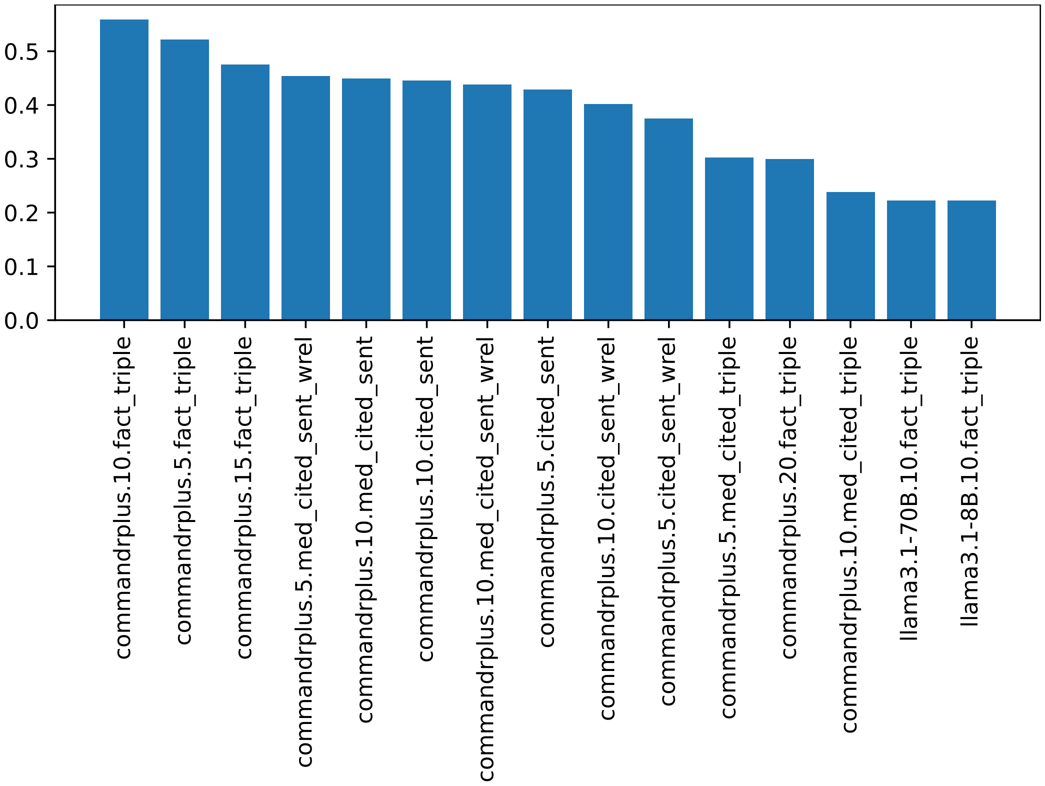

Figure 1. Models arranged by highest similarity: the x-axis value contains the model name, the number of shots, and the source of examples as listed in Table 1.

3.4. CHKP Generation

Two aspects are explored in this section: (1) the suitability of the relation extraction approach introduced in Section 3.3 for simple inference-based LBD and (2) the applicability of LLMs for filtering CHKPs.

3.4.1. Evaluation Using Existing Discoveries

Should LBD based on LLM-generated relations be unable to replicate existing discoveries, they would indicate that they may not be suitable for LBD. The following established discoveries are explored in this work:

- Raynaud–fish oil [6]. Literature time interval: 1960–1985. Terms sought: *raynaud*, *fish oil*.
- Migraine–magnesium [83]. Literature time interval: 1980–1984. Terms sought: *migraine*, *magnesium*.
- Alzheimer–Indomethacin [84]. Literature time interval: 1966–1996. Terms sought: *alzheimer*, *indomethacin*.
- Thalidomide–Myasthenia gravis [85]. Literature term interval: 1950–2000. Terms sought: *thalidomide*, *myasthenia gravis*.

While the number of replicated discoveries is small, this is primarily due to a lack of publicly available, well-documented discoveries that are suitable for LBD benchmarking and align with our specific use case of identifying treatments for diseases (note that this is a common evaluation issue: LBD works frequently only use the first two discoveries [71]). Many documented LBD discoveries are either focused on highly specialized domains (e.g., specific genes and cancer types [14]) or do not include the necessary literature range for replication.

Table 4 highlights a significant difference between relations extracted by SemRep and our FSL approach. Both approaches do not manage to extract wanted relations (TREATS, AFFECTS, AUGMENTS, PRODUCES or STIMULATES) from all articles (the *Articles* column) that contain the desired terms (the *Unique Articles* column). This is expected since, for some publications, only very short content is present in PubMed, e.g., PMID 6163587 is only represented by its title "Raynaud's phenomenon and scleroderma". However, relations are extracted from more articles using FSL consistently, e.g., for the *Raynaud–fish oil* discovery, about a third of abstracts (519 out of 1,472) produce a wanted relation, while only 430 abstracts give rise to relations using SemRep, indicating a broader coverage of the method.

Table 4. Number of relations extracted using each semantic relation extraction method for each replication example.

Discovery	Unique	SemRep		Fact_TRIPLE	
	Articles	Relations	Articles	Relations	Articles
Raynaud	1472	899	430	882	519
Migraine	4278	3513	1577	4434	2472
Alzheimer	36,096	53,871	19,204	43,062	25,316
Myasthenia gravis	8891	7315	3167	4669	4107

LBD applied simple inference to the resulting relations, which were filtered as follows: (1) relations including 1071 highly frequent terms (in SemRep and/or FSL) (such as *patient*) were removed along with terms based on frequent patterns (such as ... *workers*), (2) semantic types of source terms were restricted to a subset of *Chemicals & Drugs* (with semantic types such as *Indicator, Reagent, or Diagnostic Aid* removed) and *Gene & Molecular Sequences*, and (3) semantic types of target terms were restricted to *Disease or Syndrome*, *Mental or Behavioural Dysfunction* and *Neoplastic Process*. To ensure only **hidden** knowledge pairs are proposed, all SemRep TREATS relations (as listed in SemMedDB) appearing in works published before the end of the discovery's literature year interval (i.e., 1985 for the *Raynaud—fish oil* discovery) are removed from the result of the simple inference. SemRep TREATS relations are used in this step, rather than relations specific to each extraction approach, as the execution of the FSL approach over the entire PubMed collection would not be feasible given our resources. Table 5 shows that all four discoveries were replicated using the LLM FSL relations and that this approach consistently produced a significantly lower number of CHKPs.

Table 5. Replication of existing discoveries

Discovery	SemRep CHKPs		Fact_TRIPLE CHKPs	
	Reproduced	|CHKPs|	Reproduced	|CHKPs|
Raynaud	✓	99	✓	33
Migraine	✓	3693	✓	1400
Alzheimer	✓	151,241	✓	85,412
Myasthenia gravis	✓	3829	✓	3590

3.4.2. LLMs for Background Knowledge Identification

The columns |CHKPs| in Table 5 indicate that the number of CHKPs increases greatly with the number of studies, despite heavily restricted studies. Since removing background knowledge could reduce this number, we explore the use of LLMs, with their massive training data, via zero-shot learning for its automatic identification (see Appendix B.1 for the prompt used).

To validate the capability of LLMs for this task, a dataset of relationships listed in the UMLS is constructed with 50 examples of treatment relations (MAY_BE_TREATED_BY, MAY_TREAT, TREATED_BY and TREATS) randomly selected from examples added in each version (see Table 6 for UMLS versions used and their release dates). An additional dataset was created from a list of clinical trials around the world, available from https://clinicaltrials.gov/ (the listing downloaded on 4 August 2024 was used in this work). For each study, the data contains a start and end date, the study type and a list of conditions and interventions. Fifty examples of interventional trials with a start date of 1 May 2024 or later containing a DRUG intervention were extracted. The LLMs listed in Table 2 were tasked with indicating whether each example represented background knowledge. To ensure that the Llama models followed instructions closely, these were invoked using a transformer pipeline [86]. However, in this mode, the `llama3.1-70B` (non-quantized) model became unusable, taking 2 h 30 min to answer 6 queries, and the accuracy of the quantized version was lower than the non-quantized `llama3.1-8B`. Its results are therefore not included in Table 6, which shows the proportion of background knowledge in the three datasets found by each model.

Table 6. Percentage of background knowledge identified by each model (average over 5 runs).

Source	Source Release Date	llama2-13B	llama3.1-8B	commandR+
2006AA	13 February 2006	72%	83%	74%
2023AA	6 May 2024	64%	88%	82%
2023AB	6 November 2023	54%	99%	88%
2024AA	6 May 2024	42%	85%	80%
Clinical trials	–	40%	60%	70%

Aligning the performance in Table 6 with the models' knowledge cut-offs shown in Table 2 illustrates the reduced utility of an ageing LLM for identifying contemporary background knowledge. More recent models achieve similar performance on the UMLS data, as a larger proportion of the publications are expected to be included in the models' training data. Despite the `commandR+` model having a greater number of parameters than the `llama3.1-8B` model, the data used to train the `llama3.1-8B` model is more up to date and has been stated to be more varied, which may justify its slightly higher performance.

While all three models report a lower percentage of background knowledge within the clinical trial test set than in the older UMLS data, this is higher than expected. The reasoning provided by the `commandR+` model was manually explored for 10 randomly selected condition–intervention pairs where background knowledge was detected. In 9 of the 10 cases, the trials compared a possible new treatment to existing known treatments between which there was a published link found using a Google internet search, suggesting that a suspected link is often closely linked to existing works and therefore that the current prompt was likely to also identify CHKPs with a low novelty value.

3.4.3. Timeslicing Evaluation

A small-scale timeslicing experiment was performed to evaluate the raw percentage of CHKPs that match a SemRep TREATS relation appearing in SemMedDB from publications between 1 March 2023 and 31 December 2024. The analysis focused on CHKPs generated from publications between 1 January 2023 and 28 February 2023. This narrow time window was selected based on the following constraints:

1. The computational cost of running LLM-based fact extraction and CHKP validation for a larger time window becomes high, as each publication abstract within the window requires fact extraction and each CHKP (whether SemRep or FSL generated) needs to be background knowledge-checked.

2. The evaluation step checks for the presence of CHKPs in publications (i.e., SemMedDB) after the timeslice. However, `semmedVER43_2024_R` only contains publications through 2024 (with the majority of included publications appearing before 8 May 2024).

Table 7 demonstrates that even on a larger scale, the number of CHKPs generated from FSL relations is lower than from SemRep-based relations and that a higher percentage of these CHKPs were valid (i.e., discovered by December 2024). The overall low percentage of discovered CHKPs is likely attributable to the relatively short evaluation time period. One thousand CHKPs were randomly selected for the exploration of background detection, with the results finding 28% of the SemRep-based CHKPs and 19% of the FSL-based CHKPs to be already known. The lower quantity of background knowledge found with FSL suggests that this technique's relations are better suited to the task.

Table 7. Timeslice evaluation of CHKPs from Jan and Feb 2023 against SemRep TREATS relations.

Source	| Triples |	| CHKPs |	Discovered by Dec 2024
SemRep	836,950	207,123	0.32%
FSL	683,915	124,086	0.45%

4. Discussion

4.1. Semantic Triple Extraction

Both the replication results (Section 3.4.1) and the timeslicing experiment (Section 3.4.3) confirm the suitability of FSL-based relation extraction for LBD. A careful selection of examples allows the approach to overcome changes in word distributions seen in different domains [87], which leads to a reduction in accuracy when general approaches are employed on domain-specific data. The FSL approach has also been shown to extract relations from more publications than SemRep, and this section explores a manual analysis of the differences between the relations extracted by the two approaches.

The relations extracted from the literature subsets for the replication experiments show under 1% exact overlap of relations between the two approaches. Even when ignoring the predicate and focusing solely on exact matches for subjects and objects, the overlap percentage only doubled. This increase appears to be due to the FSL approach frequently proposing AUGMENTS instead of STIMULATES and TREATS instead of AFFECTS. The low overlap cannot be solely attributed to the publications that SemRep was unsuccessful with while FSL extracted relations, as this represents around 15% of publications that FSL extracted relations from. Exploring non-aligning triples indicates that the FSL-based approach frequently produces more specific concepts, such as the following:

- FSL: *nonsteroidal anti-inflammatory–analgesic agents* vs. Semrep: *Agent*.
- FSL: *allergic histamine release* vs. Semrep: *Histamine Release*.
- FSL: *coronary artery contractions* vs. Semrep: *Contraction*.

Additional discrepancies in alignment are due to variations in MetaMap mapping, where concepts, despite originating from the same abstract, were mapped differently. For example, FSL contains *salivary secretion*, while the corresponding SemRep concept is *secretion of saliva*. (It is important to note that this is unlikely to be due to any inconsistency in MetaMap, but rather the result of slight changes made to the concepts by the LLM.)

Overall, the differences between the FSL-based approach and SemRep are substantial enough to indicate that designing relation extraction specifically for a task, such as LBD, has the potential to result in significant changes.

4.2. Identification of Background Knowledge

Sections 3.4.2 and 3.4.3 indicate that LLMs can be used to determine background knowledge—they identify a high proportion of already known facts (Section 3.4.2) without labeling all CHKPs as known (Section 3.4.3). This suggests that the LLM's decisions are unlikely to be based on hallucinated knowledge. To further confirm this, a further experiment—based on retrieval augmented generation (RAG)—is performed.

RAG [88] allows an LLM's in-built knowledge (derived from its training data) to be augmented with additional information, typically from a knowledge graph. This provides the LLM access to both its training data and the additional information contained in the knowledge base. When using an LLM to detect background knowledge, a significant reduction in the quantity identified in the experiment from Section 3.4.3 when additional information is supplied to the LLM, would suggest that the original—un-augmented—LLM was hallucinating its decisions.

4.2.1. Knowledge Base Creation

To enable RAG, a knowledge base (KB) was set up on `neo4j aura` (https://www.neo4j.com, accessed on 4 August 2024), the free version of `neo4j`, which is limited to 400,000 relations. The KB was populated using three inputs:

- SemRep: To fit within the KB's capacity, 15 years of TREATS relations (until 2023-03-01, the end of the timeslice) with source term semantic types restricted as in Section 3.1 were included. Preference was given to longer string relations in the case of partial matches—for example, if two relations with the same source and predicate have the target terms *Influenza A (H3N2)* and *influenza A*, then the relation containing *influenza A* is removed. While this limitation may suggest limitations due to KB incompleteness, it is expected that publications prior to this date have been included in the vast training data used for all models. Publications before this date are also expected to have reached their citation saturation point [89], with publications advancing their research also either contained in the training data or in (at least) the KB.
- UMLS: Non-null MAY BE TREATED BY, MAY TREAT, TREATED BY and TREATS relations with different source and target values (with filtering as above) were extracted from the UMLS. The English preferred term was used in place of CUIs.
- Clinical Trial: Intervention–condition pairs (also known as disease–drug treatment pairs in [90]) were extracted from trials completed before 1 March 2023 that reported positive results from http://www.clinicaltrials.gov (accessed on 4 August 2024). If a trial investigated multiple interventions, these were separated to form multiple TREATS relations. Filtering was performed to remove terms referring to placebo treatments.

After performing a case-insensitive duplicate drop, the resulting KB contains 399,688 relationships and 107,200 nodes.

4.2.2. Timeslicing with RAG

To determine the impact of RAG on LLM-based background knowledge detection, Langchain's (https://www.langchain.com/, accessed 1 November 2024) retrieval chains were used to augment the prompt. The KB was queried for matches and close matches of the two terms within a CHKP and the additional information was passed to the LLM as context. The model was instructed to use this information in addition to its own knowledge to answer the question (the prompt template is shown in Appendix B.2 with the query remaining identical to direct background detection).

Table 8 presents the results of augmented queries on the 1000 CHKP sample from Section 3.4.3, providing a breakdown of the LLM's classifications. The quantity of background knowledge remains lower for FSL-based LBD than when this is SemRep relation-based. RAG augmentation resulted in a slight increase (1%) in background knowledge detection with the FSL approach, but a small decrease (5%) in SemRep. The purpose of RAG is to provide up-to-date or highly specific information that the base LLM might not have. If our KB contained a high density of facts that were not present in the LLM's training data, we would expect to see a more significant change in the classification results. The fact that the changes are marginal suggests that most of the information in our 400k-edge KB is already "known" to the LLM (which was expected as it was created from information published before the LLM's knowledge cut-off). The effect of RAG was therefore to confirm, rather than to augment, the model's existing knowledge. However, since both approaches showed an increase in the number of 'Unclear' labels, these were manually explored.

Aligning the output with [91]'s division of error types, specifically **factuality** (whether the response is factually correct), **refusal** (whether the response is a refusal to answer) and **contradiction** (whether the response is inconsistent with itself), the majority of the new 'Unclear' labels are due to contradictions within the response while the majority of 'Unclear' labels arising from the unaugmented LLM response correspond to refusals. A frequent source of contradictions was the confirmation of the context at the beginning of the response (whether or not the context contained the exact relation) with the answer and the explanation to the question following (with the connection not subsequently found).

Exploring the CHKPs whose classification changed from 'known' when RAG was added shows that the LLM's response sometimes claimed the existence of a direct relationship when only an indirect one could be supported with evidence. This did not appear to be a hallucination, rather a potential shortcoming of the prompt design (note that various versions of the prompt were tried to emphasize the need for a direct connection, but it did not appear to be possible to eliminate this category of error completely while retaining the option to use synonyms). For example, when querying the existence of a direct connection between *anticonvulsant medications* and *diabetic retinopathy*, the unaugmented LLM responds with this being a known connection but also adds the following:

> Given the potential metabolic impacts of anticonvulsants and considering the broader context of managing chronic diseases like diabetes, there appears to be a rationale supporting further investigation into any possible connections between anticonvulsants and diabetic retinopathy.

This suggests that while the LLM accurately identifies potential relevance, its interpretation of a 'direct connection' can sometimes be broader than intended for strict LBD filtering.

Table 8. Results of filtering with and without enhancement.

	No RAG			With RAG		
Source	**Yes**	**No**	**Unclear**	**Yes**	**No**	**Unclear**
Semrep	278	707	15	227	705	72
Factuality	187	788	25	199	736	65

4.3. Application to Nonmedical Domains

While the evaluation in this work is restricted to the biomedical domain, specifically its drug repurposing subfield of drug, the approach is not limited to this subfield or domain. Existing works (see, e.g., [92] for an overview) have applied LBD to nonmedical domains, relying on linguistic and statistical approaches to extract concepts from data (since the domains often do not possess controlled vocabularies or ontologies). While biomedical tools (such as UMLS semantic types and MetaMap) are used to filter information in the use case presented in this work, neither of the LLM approaches requires these. To adapt to a non-biomedical domain, or a non-drug repurposing domain, a small number of abstracts (the current work suggests 10 examples would be sufficient) should be annotated with relationships between facts in the domain (in the style of the FACTUALITY dataset) to provide information for FSL-based fact extraction. The simple inference step (which runs on the output of the FSL-extracted facts) proposes CHKPs relevant to the domain and this is followed by background knowledge detection, which runs in a zero-shot mode, requiring no domain examples, and therefore is applicable to any domain.

5. Conclusions

This work explores the use of LLMs for LBD in line with their strengths: (1) to extract factual semantic relations from biomedical text and (2) to identify knowledge pairs that represent background knowledge so they can be removed from CHKPs generated by simple inference LBD.

Using a small number of semantic relations representing facts targeted to the LBD task (by restricting the predicate) for the few-shot learning of a large language model is shown to yield a different set of semantic relations to the highly utilised tool, SemRep. This approach exhibits higher coverage than SemRep relations, with more specific terms mentioned. When used for LBD, a higher proportion of the resulting CHKPs are shown to be valid.

This work avoids the shortfall of LLMs mainly suggesting already known or closely related facts [60] by continuing to employ simple inference for LBD, followed by using LLMs to identify already known CHKPs from the resulting set. The importance of the LLM's knowledge cut-off for this task is demonstrated and it is shown to successfully filter existing knowledge with hallucination not found to pose a problem.

The next step would involve an expert evaluation of the resulting discoveries, including clinical trials as appropriate. To this end, the discoveries yielded by the timeslicing evaluation presented in the paper are being made publicly available. In addition, expert involvement could evaluate the inverse application of LLMs—when an LLM is prompted to determine whether a possible candidate hidden knowledge pair through its linking term 'makes sense'. Further re-ranking could also be offered by combining the proposed approach with other methods, such as, e.g., distance within a knowledge graph [27].

Funding: This research received no external funding.

Data Availability Statement: The CHKPs generated in this work are available from https://github.com/juditapreiss/llm_background_knowledge, accessed on 4 August 2024.

Acknowledgments: For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CHKP	Candidate hidden knowledge pair
CUI	Concept unique identifier
FSL	Few-shot learning
KB	Knowledge base
LBD	Literature-based discovery
LLM	Large language models
NLM	National Library of Medicine
RAG	Retrieval augmented generation
UMLS	Unified Medical Language System

Appendix A. Prompts for Few-Shot Learning

Appendix A.1. Extraction of Novel Contribution Sentence

List novel facts introduced in the title / abstract / conclusion of the academic publication shown after <<<< below - these should represent the novel contributions of the paper. Only list facts which appear in the paper, do not generate any new titles, abstracts, conclusion or summaries. Place each fact on a new line, and start the line with a dash. Output as many facts as can be extracted from the title / abstract / conclusion segment, while listing each fact as a complete sentence. Some examples are:

Appendix A.2. Extraction of Semantic Relations

List novel facts introduced in the title / abstract / conclusion of the academic publication shown after <<<< below - these should represent the novel contributions of the paper. Only list facts which appear in the paper, do not generate any new titles, abstracts, conclusion or summaries. The novel contributions should be listed as semantic triples (i.e., subject relation object triples), where the only possible relations are: TREATS, AFFECTS, AUGMENTS, PRODUCES, STIMULATES - do not suggest other relations. Place each triple on a new line, separating the three elements with tabs. Output as many facts as appear in the title / abstract / conclusion segment, while listing each fact as a complete sentence. Some examples are:

Appendix B. Prompts for Zero-Shot Background Detection

Appendix B.1. Direct Background Knowledge Detection

Based on your biomedical knowledge, answer the following research question. Is there an already known and established {{PREDICATION}} connection between {{SOURCE_TERM}} and {{TARGET_TERM}}? You should answer this question as a researcher, starting your answer with 'Yes' if a relation already exists or 'No' if it does not, even if it seems plausible, with reasoning following.

In this work, PREDICATION is always the TREATS relation, and the source and target terms are inserted for each relation pair.

Appendix B.2. Background Detection with RAG

You are an assistant for question-answering tasks. Use your own knowledge in addition to that included in the context. The context represents relations extracted from a knowledge base, found by searching for the two terms provided. Ignore irrelevant relations but feel free to explore synonyms, hypernyms and hyponyms of treatments and diseases to make your decision. If you don't know the answer, just say that you don't know. Please start your answer with 'Yes' or 'No', with reasoning following. Question: {{ question }} Context: {{ context }} Answer:

The query is identical to that in direct background detection (shown in Appendix B.1).

References

1. Hahn-Powell, G. Machine Reading for Scientific Discovery. Ph.D. Thesis, University of Arizona, Tucson, AZ, USA , 2018.
2. Kostoff, R.N.; Patel, U. Literature-related discovery and innovation: Chronic kidney disease. *Technol. Forecast. Soc. Change* **2015**, *91*, 341–351. [CrossRef]
3. Zhang, R.; Hristovski, D.; Schutte, D.; Kastrin, A.; Fiszman, M.; Kilicoglu, H. Drug repurposing for COVID-19 via knowledge graph completion. *J. Biomed. Inform.* **2021**, *115*, 103696. [CrossRef]
4. Swanson, D.R.; Smalheiser, N.R. An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artif. Intell.* **1997**, *91*, 183–203. [CrossRef]

5. Henry, S.; McInnes, B.T. Literature Based Discovery: Models, methods, and trends. *J. Biomed. Inform.* **2017**, *74*, 20–32. [CrossRef] [PubMed]
6. Swanson, D.R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **1986**, *30*, 7–18. [CrossRef]
7. Preiss, J.; Stevenson, M.; Gaizauskas, R. Exploring Relation Types for Literature-based Discovery. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 987–992. [CrossRef]
8. Cohen, T.; Widdows, D.; Schvaneveldt, R.W.; Davies, P.; Rindflesch, T.C. Discovering discovery patterns with predication-based Semantic Indexing. *J. Biomed. Inform.* **2012**, *45*, 1049–1065. [CrossRef]
9. Pratt, W.; Yetisgen-Yildiz, M. LitLinker: Capturing connections across the biomedical literature. In Proceedings of the K-CAP03: International Conference on Knowledge Capture 2003, Sanibel Island, FL, USA, 23–25 October 2003; pp. 105–112.
10. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. [CrossRef]
11. Srinivasan, P. Text mining: Generating hypotheses from MEDLINE. *J. Am. Soc. Inf. Sci. Technol.* **2004**, *55*, 396–413. [CrossRef]
12. Weeber, M.; Vos, R.; Klein, H.; de Jong-van den Berg, L.T.W. Using concepts in literature-based discovery: Simulating Swanson's Reynaud—fish oil and Migraine—magnesium discoveries. *J. Am. Soc. Inf. Sci. Technol.* **2001**, *52*, 548–557. [CrossRef]
13. Wilkowski, B.; Fiszman, M.; Miller, C.M.; Hristovski, D.; Arabandi, S.; Rosemblat, G.; Rindflesch, T.C. Graph-Based Methods for Discovery Browsing with Semantic Predications. *AMIA Annu. Symp. Proc.* **2011**, *2011*, 1514–1523. [PubMed]
14. Pyysalo, S.; Baker, S.; Ali, I.; Haselwimmer, S.; Shah, T.; Young, A.; Guo, Y.; Högberg, J.; Stenius, U.; Narita, M.; et al. LION LBD: A literature-based discovery system for cancer biology. *Bioinformatics* **2018**, *35*, 1553–1561. [CrossRef]
15. Škrlj, B.; Kokalj, E.; Lavrač, N. PubMed-Scale Chemical Concept Embeddings Reconstruct Physical Protein Interaction Networks. *Front. Res. Metrics Anal.* **2021**, *6*, 644614. [CrossRef] [PubMed]
16. Hristovski, D.; Friedman, C.; Rindflesch, T.C.; Peterlin, B. Exploiting semantic relations for literature-based discovery. *AMIA Annu. Symp. Proc.* **2006**, *2006*, 349–353. [PubMed]
17. Swanson, D.R.; Smalheiser, N.R.; Torvik, V.I. Ranking Indirect Connnections in Literature-Based Discovery: The Role of Medical Subject Headings. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 1427–1439. [CrossRef]
18. Wren, J.D. Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinform* **2004**, *5*, 145. [CrossRef] [PubMed]
19. Cameron, D.; Kavuluru, R.; Rindflesch, T.C.; Sheth, A.P.; Thirunarayan, K.; Bodenreider, O. Context-Driven Automatic Subgraph Creation for Literature-Based Discovery. *J. Biomed. Inform.* **2015**, *54*, 141–157. [CrossRef]
20. Bruza, P.; Song, D.; McArthur, R. Abduction in semantic space: Towards a logic of discovery. *Log. J. IGPL* **2004**, *12*, 97–109. [CrossRef]
21. Hristovski, D.; Peterlin, B.; Mitchell, J.A.; Humphrey, S.M. Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.* **2005**, *74*, 289–298. [CrossRef]
22. Hunter, L.; Cohen, K.B. Biomedical Language Processing: Perspective What's Beyond PubMed? *Mol. Cell* **2006**, *21*, 589–594. [CrossRef]
23. Doneva, S.E.; Qin, S.; Sick, B.; Ellendorff, T.; Goldman, J.P.; Schneider, G.; Ineichen, B.V. Large language models to process, analyze, and synthesize biomedical texts: A scoping review. *Discov. Artif. Intell.* **2024**, *4*, 107. [CrossRef]
24. Cesario, E.; Comito, C.; Zumpano, E. A survey of the recent trends in deep learning for literature based discovery in the biomedical domain. *Neurocomputing* **2024**, *568*, 127079. [CrossRef]
25. Preiss, J. Avoiding Background Knowledge: Literature Based Discovery From Important Information. *BMC Bioinform.* **2022**, *23* (Suppl. 9), 570. [CrossRef]
26. Kilicoglu, H.; Shin, D.; Fiszman, M.; Rosemblat, G.; Rindflesch, T.C. SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Bioinformatics* **2012**, *28*, 3158–3160. [CrossRef]
27. Percha, B.; Altman, R.B. A global network of biomedical relationships derived from text. *Bioinformatics* **2018**, *34*, 2614–2624. [CrossRef]
28. Sosa, D.N.; Derry, A.; Guo, M.; Wei, E.; Brinton, C.; Altman, R.B. A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. *Pac. Symp. Biocomput.* **2020**, *25*, 463–474.
29. Wang, Q.; Li, M.; Wang, X.; Parulian, N.; Han, G.; Ma, J.; Tu, J.; Lin, Y.; Zhang, R.H.; Liu, W.; et al. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*; Sil, A., Lin, X.V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 66–77. [CrossRef]
30. Taleb, I.; Navaz, A.N.; Serhani, M.A. Leveraging Large Language Models for Enhancing Literature-Based Discovery. *Big Data Cogn. Comput.* **2024**, *8*, 146. [CrossRef]
31. Shang, N.; Xua, H.; Rindflesch, T.C.; Cohen, T. Identifying plausible adverse drug reactions using knowledge extracted from the literature. *J. Biomed. Inform.* **2014**, *52*, 293–310. [CrossRef]

32. Zhang, R.; Cairelli, M.J.; Fiszman, M.; Kilicoglu, H.; Rindflesch, T.C.; Pakhomov, S.V.; Melton, G.B. Exploiting literature-derived knowledge and semantics to identify potential prostate cancer drugs. *Cancer Inform.* **2014**, *13*, 103–111. [CrossRef] [PubMed]
33. Sosa, D.N.; Altman, R.B. Contexts and contradictions: A roadmap for computational drug repurposing with knowledge inference. *Briefings Bioinform.* **2022**, *23*, bbac268. [CrossRef] [PubMed]
34. Mullard, A. Drug repurposing programmes get lift off. *Nat. Rev. Drug Discov.* **2012**, *11*, 505. [CrossRef]
35. Dai, S.; You, R.; Lu, Z.; Huang, X.; Mamitsuka, H.; Zhu, S. FullMeSH: Improving large-scale MeSH indexing with full text. *Bioinformatics* **2019**, *36*, 1533–1541. [CrossRef]
36. McInnes, B.T.; Pedersen, T.; Carlis, J. Using UMLS Concept Unique Identifiers (CUIs) for word sense disambiguation in the biomedical domain. *AMIA Annu. Symp. Proc.* **2007**, *2007*, 533–537.
37. Yetisgen-Yildiz, M.; Pratt, W. Using statistical and knowledge-based approaches for literature-based discovery. *J. Biomed. Inform.* **2006**, *39*, 600–611. [CrossRef] [PubMed]
38. Aronson, A.R.; Lang, F.M. An overview of MetaMap: Historical perspective and recent advances. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 229–236. [CrossRef] [PubMed]
39. Yang, H.T.; Ju, J.H.; Wong, Y.T.; Shmulevich, I.; Chiang, J.H. Literature-based discovery of new candidates for drug repurposing. *Briefings Bioinform.* **2016**, *18*, 488–497. [CrossRef] [PubMed]
40. Gopalakrishnan, V.; Jha, K.; Jin, W.; Zhang, A. A survey on literature based discovery approaches in biomedical domain. *J. Biomed. Inform.* **2019**, *93*, 103141. [CrossRef]
41. Zhao, D.; Wang, J.; Sang, S.; Lin, H.; Wen, J.; Yang, C. Relation path feature embedding based convolutional neural network method for drug discovery. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 59. [CrossRef] [PubMed]
42. Hristovski, D.; Stare, J.; Peterlin, B.; Dzeroski, S. Supporting discovery in medicine by association rule mining in Medline and UMLS. *Stud. Health Technol. Inform.* **2001**, *84*, 1344–1348.
43. Kilicoglu, H.; Rosemblat, G.; Fiszman, M.; Rindflesch, T.C. Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinform.* **2011**, *12*, 486. [CrossRef]
44. Kilicoglu, H.; Rosemblat, G.; Fiszman, M.; Shin, D. Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinform.* **2020**, *21*, 188. [CrossRef]
45. Ming, S.; Zhang, R.; Kilicoglu, H. Enhancing the coverage of SemRep using a relation classification approach. *J. Biomed. Inform.* **2024**, *155*, 104658. [CrossRef] [PubMed]
46. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc.* **2021**, *3*, 3458754. [CrossRef]
47. Song, M.; Kim, W.C.; Lee, D.; Heo, G.E.; Kang, K.Y. PKDE4J: Entity and relation extraction for public knowledge discovery. *J. Biomed. Inform.* **2015**, *57*, 320–332. [CrossRef]
48. Hong, G.; Kim, Y.; Choi, Y.; Song, M. BioPREP: Deep learning-based predicate classification with SemMedDB. *J. Biomed. Inform.* **2021**, *122*, 103888. [CrossRef]
49. Lai, P.T.; Wei, C.H.; Luo, L.; Chen, Q.; Lu, Z. BioREx: Improving biomedical relation extraction by leveraging heterogeneous datasets. *J. Biomed. Inform.* **2023**, *146*, 104487. [CrossRef]
50. Mitra, A.; Rawat, B.P.S.; McManus, D.D.; Yu, H. Relation Classification for Bleeding Events From Electronic Health Records Using Deep Learning Systems: An Empirical Study. *JMIR Med. Inform.* **2021**, *9*, e27527. [CrossRef]
51. Huguet Cabot, P.L.; Navigli, R. REBEL: Relation Extraction By End-to-end Language generation. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021*; Moens, M.F., Huang, X., Specia, L., Yih, S.W.T., Eds.; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 2370–2381. [CrossRef]
52. Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; Sontag, D. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 1998–2022. [CrossRef]
53. Bahaj, A.; Ghogho, M. A step towards quantifying, modelling and exploring uncertainty in biomedical knowledge graphs. *Comput. Biol. Med.* **2025**, *184*, 109355. [CrossRef] [PubMed]
54. Kilicoglu, H.; Rosemblat, G.; Rindflesch, T.C. Assigning factuality values to semantic relations extracted from biomedical research literature. *PLoS ONE* **2017**, *7*, e0179926. [CrossRef]
55. Hristovski, D.; Rindflesch, T.; Peterlin, B. Using Literature-based Discovery to Identify Novel Therapeutic Approaches. *Cardiovasc. Hematol. Agents Med. Chem.* **2013**, *11*, 14–24. [CrossRef] [PubMed]
56. Smalheiser, N.R. The Arrowsmith Project: 2005 Status Report. In *Discovery Science*; Hoffmann A., Motoda, S.T.H., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3735.
57. Tsuruoka, Y.; Miwa, M.; Hamamoto, K.; Tsujii, J.; Ananiadou, S. Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics* **2011**, *27*, i111–i119. [CrossRef]

58. Hu, X.; Li, G.; Yoo, I.; Zhang, X.; Xu, X. A semantic-based approach for mining undiscovered public knowledge from biomedical literature. In Proceedings of the 2005 IEEE International Conference on Granular Computing, Beijing, China, 25–27 July 2005; Volume 1, pp. 22–27.
59. Syafiandini, A.F.; Song, G.; Ahn, Y.; Kim, H.; Song, M. An automatic hypothesis generation for plausible linkage between xanthium and diabetes. *Sci. Rep.* **2022**, *12*, 17547. [CrossRef]
60. Sybrandt, J.; Tyagin, I.; Shtutman, M.; Safro, I. AGATHA: Automatic Graph Mining And Transformer based Hypothesis Generation Approach. In Proceedings of the CIKM '20: 29th ACM International Conference on Information & Knowledge Management, New York, NY, USA, 19–23 October 2020; pp. 2757–2764. [CrossRef]
61. Crichton, G.; Baker, S.; Guo, Y.; Korhonen, A. Neural networks for open and closed Literature-based Discovery. *PLoS ONE* **2020**, *15*, e0232891. [CrossRef]
62. Zeng, X.; Zhu, S.; Lu, W.; Liu, Z.; Huang, J.; Zhou, Y.; Fang, J.; Huang, Y.; Guo, H.; Li, L.; et al. Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* **2020**, *11*, 1775–1797. [CrossRef]
63. Birhane, A.; Kasirzadeh, A.; Leslie, D.; Wachter, S. Science in the age of large language models. *Nat. Rev. Phys.* **2023**, *5*, 277–280. [CrossRef]
64. Iannantuono, G.M.; Bracken-Clarke, D.; Floudas, C.S.; Roselli, M.; Gulley, J.L.; Karzai, F. Applications of large language models in cancer care: Current evidence and future perspectives. *Front Oncol.* **2023**, *13*, 1268915. [CrossRef]
65. Wang, Q.; Downey, D.; Ji, H.; Hope, T. SciMON: Scientific Inspiration Machines Optimized for Novelty. *arXiv* **2024**, arXiv:2305.14259. [CrossRef]
66. Nedbaylo, A.; Hristovski, D. Implementing Literature-based Discovery (LBD) with ChatGPT. In Proceedings of the 2024 47th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 20–24 May 2024; pp. 120–125. [CrossRef]
67. Peng, Y.; Bonifield, G.; Smalheiser, N.R. Gaps within the biomedical literature: Initial characterization and assessment of strategies for discovery. *Front. Res. Metrics Anal.* **2017**, *2*, 3. [CrossRef] [PubMed]
68. Bolding, Q.; Liao, B.; Denis, B.J.; Luo, J.; Monz, C. Ask Language Model to Clean Your Noisy Translation Data. In Proceedings of the Findings of EMNLP, Singapore, 6–10 December 2023.
69. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In Proceedings of the NIPS'23: 37th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 10–16 December 2023.
70. Yetisgen-Yildiz, M.; Pratt, W. Evaluation of literature-based discovery systems. Literature-Based Discovery. In *Information Science and Knowledge Management*; Bruza, P., Weeber, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; Volume 15, pp. 101–113.
71. Thilakaratne, M.; Falkner, K.; Atapattu, T. A systematic review on literature-based discovery workflow. *PeerJ Comput. Sci.* **2019**, *5*, e235. [CrossRef]
72. Yetisgen-Yildiz, M.; Pratt, W. A new evaluation methodology for literature-based discovery. *J. Biomed. Inform.* **2009**, *42*, 633–643. [CrossRef]
73. Moreau, E. Literature-based discovery: Addressing the issue of the subpar evaluation methodology. *Bioinformatics* **2023**, *39*, btad090. [CrossRef]
74. Ganiz, M.C.; Pottenger, W.M.; Janneck, C.D. *Recent Advances in Literature Based Discovery*; Technical Report LU-CSE-05-027; Lehigh University: Bethlehem, PA, USA, 2005.
75. Lo, K.; Wang, L.L.; Neumann, M.; Kinney, R.; Weld, D. S2ORC: The Semantic Scholar Open Research Corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4969–4983.
76. GROBID. 2008–2021. Available online: https://github.com/kermitt2/grobid (accessed on 4 August 2024).
77. Halvani, O. Constituent Treelib—A Lightweight Python Library for Constructing, Processing, and Visualizing Constituent Trees v0.0.8. 2023. Available online: https://pypi.org/project/constituent-treelib/ (accessed on 4 August 2024).
78. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971. [CrossRef]
79. Segura-Bedmar, I.; Martínez, P.; Herrero-Zazo, M. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, GA, USA, 14–15 June 2013; pp. 341–350.
80. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Kerrville, TX, USA, 2019.
81. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
82. Supriyono; Wibawa, A.P.; Suyono; Kurniawan, F. A survey of text summarization: Techniques, evaluation and challenges. *Nat. Lang. Process. J.* **2024**, *7*, 100070. [CrossRef]

83. Swanson, D.R. Migraine and magnesium—11 neglected connections. *Perpectives Biol. Med.* **1988**, *31*, 526–557. [CrossRef]
84. Smalheiser, N.R.; Swanson, D.R. Indomethacin and Alzheimer's disease. *Neurology* **1996**, *46*, 583. [CrossRef]
85. Weeber, M.; Vos, R.; Klein, H.; de Jong-van den Berg, L.T.; Aronson, A.R.; Molema, G. Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide. *J. Am. Med. Inform. Assoc.* **2003**, *10*, 252–259. [CrossRef] [PubMed]
86. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2020**, arXiv:1910.03771.
87. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef] [PubMed]
88. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
89. Ponomarev, I.V.; Williams, D.E.; Hackett, C.J.; Schnell, J.D.; Haak, L.L. Predicting highly cited papers: A Method for Early Detection of Candidate Breakthroughs. *Technol. Forecast. Soc. Change* **2014**, *81*, 49–55. [CrossRef]
90. Xu, R.; Li, L.; Wang, Q. dRiskKB: A large-scale disease-disease risk relationship knowledge base constructed from biomedical text. *BMC Bioinform.* **2014**, *15*, 105. [CrossRef] [PubMed]
91. Hosking, T.; Blunsom, P.; Bartolo, M. Human Feedback is not Gold Standard. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2024.
92. Hui, W.; Lau, W.K. Application of Literature-Based Discovery in Nonmedical Disciplines: A Survey. In Proceedings of the 2nd International Conference on Computing and Big Data, New York, NY, USA, Taichung, Taiwan, 18–20 October 2019; pp. 7–11. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.