The Search for Conflicts of Interest: Open Information Extraction in Scientific Publications

Garima Gaur
Inria, Institut Polytechnique de Paris
Palaiseau, France
garima.gaur@inria.fr

Oana Balalau
Inria, Institut Polytechnique de Paris
Palaiseau, France
oana.balalau@inria.fr

Ioana Manolescu
Inria, Institut Polytechnique de Paris
Palaiseau, France
ioana.manolescu@inria.fr

Prajna Devi Upadhyay
BITS Pilani Hyderabad
Hyderabad, India
prajna.u@hyderabad.bits-pilani.ac.in

Abstract

A conflict of interest (COI) appears when a person or a company has two or more interests that may directly conflict. This happens, for instance, when a scientist whose research is funded by a company audits the same company. For transparency and to avoid undue influence, public repositories of relations of interest are increasingly recommended or mandated in various domains, and can be used to avoid COIs. In this work, we propose an LLM-based open information extraction (OpenIE) framework for extracting financial or other types of interesting relations from scientific text. We target scientific publications in which authors declare funding sources or collaborations in the acknowledgment section, in the metadata, or in the publication, following editors' requirements. We introduce an extraction methodology and present a knowledge base (KB) with a comprehensive taxonomy of COI centric relations. Finally, we perform a comparative study of disclosures of two journals in the field of toxicology and pharmacology.

1 Introduction

Many modern institutions require inputs from experts prior to decisions that affect people's health, finances, rights, etc. For instance, expert input is needed to approve new drugs or medical procedures, to award contracts or public grants, etc. We call *expert* a person whose input may be called upon for such a decision, or who may actually take the decision. For public trust in the process, experts should not have *Conflicts of Interest (COIs)*, that is: they should not stand to gain from a decision that they can impact. To avoid COIs, *transparency databases* are built, cataloging known relations of interest between individuals and organizations, and consulted to avoid soliciting experts with stakes in the decision. For instance, the US Sunshine Act requires medical doctors to state their relationships, e.g., with drug and medical device manufacturers; the EU database EurosForDocs harvests such information from different EU countries, etc.

In this work, we investigate if reliable and valuable Information Extraction (IE) can be performed from *disclosure texts* using LLMs. In the context of scientific articles, the disclosure text can appear in a variety of forms – from an explicit declaration of COI as manuscript metadata to an *Acknowledgment*, or a *Funding* section at the end of the manuscript. We observed that while LLM can produce good quality triples as compared to traditional methods, the extracted triples need further refinement for improving the overall quality and, hence, usability of the collection of triples. To this end, we propose an end-to-end pipeline that relies on in-context learning for raw triple extraction and, further, leverages rule-based and clustering-based methods to transform the triples into a meaningful, well-structured *knowledge base* (KB). Our methodology can be replicated to build transparency KBs in a variety of contexts.

As a step forward, we applied our methodology to construct a KB capturing relations of interest from journal articles in the field of toxicology and pharmacology. We selected the biomedical domain due to large impact on the public's health, as witnessed for instance by the "Forever Pollution" project undertaken by journalists from 16 countries, and motivated by prior research highlighting possible real-world harm resulting from COIs in this domain (see Section 2). Inspired by prior journalistic work which identified experts involved in COIs with the food industry, we selected: a journal in which these experts published: "Regulatory Toxicology and Pharmacology" (RTP, in short), and, as a basis for comparison, a second one from the same scientific area, namely "Toxicology and Ap-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13922–13936

plied Pharmacology" (TAP, in short). Our analysis highlights very strong involvement of the industry in one scientific journal, in striking contrast with the other.

To summarize, our contributions are as follows:

- We present a **complete OpenIE pipeline** that combines the strength of LLMs and classic cleaning and normalizing techniques, like filtering and clustering, to produce high quality triples. It leverages the superior capabilities of LLMs through in-context learning for smaller decomposed tasks.
- We present a **rich, nuanced, yet manageable-size vocabulary** (300+) describing relations which may appear in declared or extracted COI statements. Further, we **align these relationships into a taxonomy, by connecting them to an existing set** of 11 COI relations identified in recent work (Hardy et al., 2023).
- We build a KB by collectively extracting $30K$**+ triples** from RTP and TAP articles published over two decades. Further, we perform a comparative analysis of industry involvement across these journals.

Our code and KB are publicly available[1].

2 Related Work

Open Information Extraction Open Information Extraction (OIE) seeks to identify the subject, relation, and object of a sentence in a domain-independent manner, unlike closed-domain information extraction, where the extracted information should belong to known domains. A comprehensive review of open information extraction techniques is (Pai et al., 2024). OIE has evolved significantly over the years – from rule-based or traditional machine learning approaches (Yates et al., 2007), to neural approaches (Kolluru et al., 2020b,a; Stanovsky and Dagan, 2016; Upadhyay et al., 2023), to recent approaches based on large language models (LLMs) (Qi et al., 2023). Stanovsky and Dagan (2016); Kolluru et al. (2020a) formulated OIE as a sequence-labelling problem, while Cui et al. (2018); Kolluru et al. (2020b) cast it as a sequence to sequence generation task. This was further extended to handle named entities by implementing soft constraints during training and inference (Upadhyay et al., 2023).

The emergence of powerful LLMs (GPT-4 (OpenAI et al., 2024), Llama-3 (Grattafiori et al., 2024)) highlight the promise of LLMs for OIE, either through fine-tuning (Lu et al., 2023) or through few-shot prompting (Ling et al., 2023). Our work extends this line by showing that few-shot prompting alone can be competitive not only for OIE but also for relation normalization. At the same time, we demonstrate that purely generation-based methods are sometimes brittle, and that lightweight linguistically motivated preprocessing rules remain useful to enforce prompt adherence.

KG Construction The LLM-based methods (Arsenyan et al., 2024; Zhu et al., 2024; Lairgi et al., 2024; Zhang and Soh, 2024) for KG construction are quite effective in end-to-end extraction of well-structured triples from the textual input. However, these methods, such as (Arsenyan et al., 2024; Zhang and Soh, 2024), either expect an ontology/schema as input to guide the relation extraction and semantic deduplication task, or methods like (Zhu et al., 2024; Lairgi et al., 2024) use LLMs for computing or/and validating semantic equivalence of extracted relations. The former category of solutions makes a strong assumption that does not hold, as relations are often not known beforehand. The latter relies on LLMs in an end-to-end fashion. Zhu et al proposed one such method, AutoKG (Zhu et al., 2024), and listed generalization to specialized scientific text as the limitation of AutoKG.

Mining Conflicts of Interest Anadiotis et al. (2021) extract COI phrases from *explicit disclosure statements (part of the article metadata)* and from PDF articles, and apply Named Entity Recognition (NER) on them. However, COI relationships are not extracted. Graham et al. (2022) identify three broad classes of COIs: *Type 1* (personal fees, travel, board memberships, and non-financial support), *Type 2* (grants and research support), and *Type 3* (stock ownership and industry employment). They use custom NER and a COI term dictionary to extract and classify COI relations from 200K explicit PubMed disclosure statements in these three classes. Hardy et al. (2023) propose a more comprehensive set of 11 COI relationships. Graham et al. (2024) propose a dataset of over 38K COI statements from PubMed, mostly from explicit metadata, but also (when this is not present) from the papers' full text. The authors targeted papers rele-

[1] https://gitlab.inria.fr/cedar/coi-openie

vant for frequently-prescribed drugs; no information extraction is attempted on the COIs.

Compared with the above, our work: (i) *harvests COI phrases from PDFs also*, since explicit disclosure statements are rarely present, as noted in (Anadiotis et al., 2021; Graham et al., 2024); such phrases increase the diversity of relationships and entities involved; (ii) proposes a *fine-granularity relationship vocabulary* covering a large variety of situations; (iii) contributes a *more generic dataset of COI triples*, where *the subject and/or the object can be a Person, or an Organization NE, or an article*. Further, (iv) we provide a *taxonomy of COI relationships* by aligning our relationships to the 11 ones introduced in (Hardy et al., 2023). This will help streamline the automatic detection and classification of COI relationships, a pressing need noted, e.g., in (Graham et al., 2024; Xun et al., 2024; McCartney, 2024).

3 Extracting Relations of Interest from Disclosure Text

Extracting relations of interest from the disclosure statements involves acquiring and cleaning the data (Section 3.1), followed by the extraction of the relations of interest (Section 3.2), and finally the creation of a knowledge base (Section 3.3). We present an overview of our approach in Figure 1.

3.1 Data Collection and Cleaning

Data Collection We retrieve all publications from RTP and TAP over a period of **20 years, from 1993 to 2022**. For each publication, we extract relations of interests from two sources: the publication **metadata**, available as **XML**, and the **PDF** of the publication itself. In the XML metadata, authors are given the option to fill an element labeled *CoiStatement*) with free text. In addition to the relations of interest, we also extract from the bibliographic metadata (XML) **the author names and affiliations**. Affiliations are interesting because being employed by an entity (company or other organization) is clearly also a relation of interest by itself. In the PDF article itself, we discard all but the part of each article (typically towards the end) that contains certain *phrases of interest*, which we built manually discussing with a domain expert (such as "Acknowledgment", "Statement of Interest", etc.; see Appendix A for the full list).

Data Cleaning We focus on *explicitly declared* conflicts of interest; in particular, we discard the articles that do not use any of these phrases of interest. As we will see (Section 4), *very few articles have a CoIStatement in the metadata*, whereas *a majority of articles state some relations of interest*. Note that *latent (implicit, not explicit)* conflicts of interest may possibly be detected from an article, however, we do not consider them in this work.

We perform *co-reference resolution* and *authors' name normalization* on the COI phrases from both the metadata and the PDF content (Appendix B).

3.2 Open Information Extraction

In this section, we present the methodology for extracting relations of interest from the sentences we have retrieved from the journals.

Triple Extraction Given the strong performance of state-of-the-art LLMs (Team et al., 2024), we perform the extraction using few shot learning. We prompt an LLM with the task description and multiple examples of the task. We detail the prompt in the Appendix E.1. For a given input, the LLM may extract triples that, while being correct, are not quite relevant to our agenda of extracting potential conflict of interest relations. For instance, for an input string *"This research was funded by Pfizer Ltd, New York, US"*, the LLM can output a list of triples: *[<this research, was funded by, Pfizer New York>, <New york, is located in, US>]*. In this scenario, we focus on only those triples that capture a relevant relation. This leads to our next task of identifying and filtering relevant triples. We refer to this task as *triple refinement*.

Triple Refinement We perform Named Entity Recognition (NER) by prompting the LLM with the standard name of the task, and the expected output format of the annotations; the prompt is in Appendix E.2. We filter the triples as follows: *the subject and object of an extracted triple must each be a named entity of the type Person, Organization or Grant*. A Grant entity is often mentioned next to the Organization financing it. For instance, in the triple *(this study, was funded by, EU commission (Grant no 11232))*, the object returned by the LLM contains the *EU commission* Organization and the *Grant No 11232* Grant. We group such complex-object (or complex-subject) triples with those having only one Organization NE in the respective position, by fusing the Grant and the Organization in a single Organization entity. Based on the NER annotations, we keep only the triples between the following combinations of

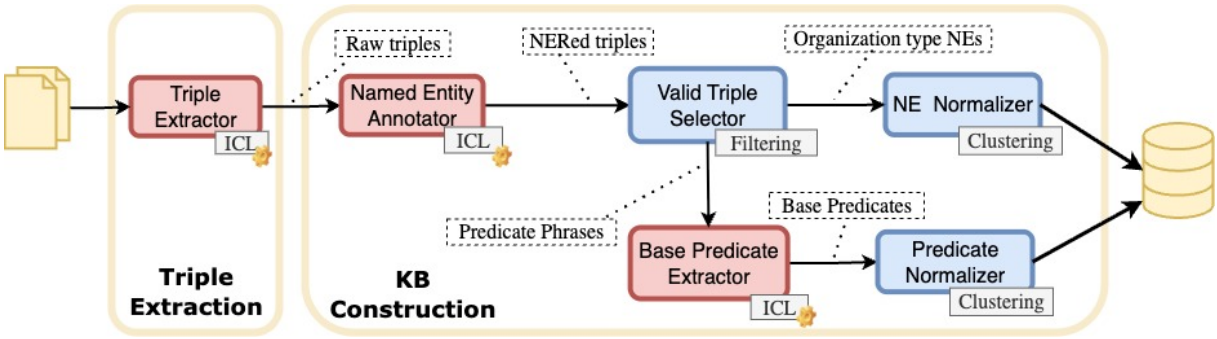

Figure 1: Overview of our pipeline that takes as input a collection of articles and produces a KB. The steps in red boxes rely on LLM via in-context learning (ICL), and the steps in blue are based on classical deterministic methods.

(subject, object): (Person, Person), (Organization, Organization), (Person, Organization), (Organization, Person). We add additional extractions that do not follow these rules, but still represent valid interest relations:

- Phrases with implicit object: For example, the triple *(Authors, thank Sigrid Roesener for, proofreading)* does not have any NE in the object phrase but contains an entity of type Person in the predicate. Therefore, we reformulate this into the valid triple *(Authors, thanked for proofreading, Sigrid Roesener)*.
- Phrases with implicit subject: In some LLM extracted triples the subject is under-specified, for e.g., triple *(Financial support, was provided by, NHS)* conveys that the authors were provided financial support by NHS to conduct the corresponding research work. Therefore, the triple should be converted to *(Authors, were provided financial support by, NHS)*. If neither the subject nor the predicate phrase has any NE and the object has an NE, then often the implicit subject is the authors or the study. Therefore, such triples must be included in the valid triple collection.

Triples thus obtained exhibit a large variety of lexical forms in their subjects, objects, and predicates. Thus, we added to our pipeline steps focused specifically on normalizing them, as follows.

3.3 A Knowledge Base of Relations of Interest

Predicate Normalization This task can be compared to the task of relation/predicate classification, however, in our case, a fixed set of classes, i.e., a set of normalized predicates is not available. Our aim is to group semantically similar predicate phrases, while also preserving sufficient variety to capture many different relations. In essence, each predicate phrase captures a core relation. For example, the predicate phrases *received pilot project grant from* and *received basic science research program grant from* can both be standardized into *has received grant*, which captures their essential, common meaning. We refer to the phrase capturing the core relation as the **base predicate**. Based on the base predicates, we group the predicate phrases and represent them using a normalized form. In our example, both predicate phrases will be represented by *received grant*. We relied on few-shot learning for identifying base predicate for a given predicate phrase . This prompt is shown in the Appendix E.3.

The above steps still resulted in thousands of base predicates. To facilitate analysis, we *clustered* them to obtain a nuanced yet manageable-size vocabulary of different relations, as follows. We relied on the classic hierarchical agglomerative clustering and performed two rounds of clustering. In the first round, we *cluster the base predicates*. This first round follows a two-step procedure: first, we obtain highly coherent clusters obtained by Agglomerative clustering with a threshold parameter t_1. Note that the threshold distance parameter provides an upper bound on the distance between clusters that can be merged, i.e., the smaller the threshold distance, the more coherent the clusters. We use the Silhouette score to measure the cluster coherence. We keep the clusters with a higher Silhouette score ($\geq s1$) and re-cluster the remaining cluster predicates with a higher threshold distance t_2. The different thresholds are determined empirically using a ground truth set. The details of parameter tuning are presented in Section 4.2.

In the second round, we select from each cluster its *representative* (i.e. the predicate closest to the centroid). We then *cluster the representatives*. If a

cluster's Silhouette score is greater than the threshold $s1$, we merge the corresponding base predicate clusters (from the first round), else, we keep the clusters as is. This second round allows us to further generalize predicates.

Entity Normalization The subject and the object of the extracted triples, together, make up the collection of entities in the dataset. We call *entity phrase* the *subject* or *object* in an extracted triple. An entity phrase can be of three types: (i) a proper noun mentioning a person or an organization; (ii) a phrase referring to authors or the article itself, e.g., *authors of the paper*, *this study*, *the manuscript*, etc., and (iii) a grant number. We anonymized the author names by replacing them with a numeric identifier. To disambiguate, in our entity set, we replace phrases referring to the authors, or to the article, with standardized entries of the form `A_<ID>` or `P_<ID>`, representing the authors of the article with identifier `<ID>`, respectively, the article itself. Entity phrases that just contain grant numbers do not need normalization. The quality of our entity phrases can still be improved, as different entity phrases may refer to the same entity in real-life. For instance, entity phrase *FDA*, *Food and drug administration*, or *Food and drug administration (FDA)* all refer to the US FDA. We identify **abbreviations** in the set of Organizations, by selecting those are in uppercase only, and consist of a single word. We look up the abbreviations in the set of Organizations, and group together all the Organizations that *contain* the same abbreviation. This unifies a significant percentage of Organizations, e.g., *FDA* and *Food and drug administration (FDA)*, however, *Food and drug administration* does not (yet) join them. Therefore, in our next step: (i) in each group, we drop the acronym from all the names, e.g., *Food and drug administration (FDA)* becomes *Food and drug administration*; (ii) we cluster using Agglomerative clustering all the Organizations, considering a group obtained in the previous step as one organization. In our example, this ensures that the original name *Food and drug administration* is associated to the same bucket as *FDA*.

4 Experiments

We now describe experiments validating the methodology presented in Section 3.

Dataset From the RTP, respectively, TAP journals we retrieved the entire publication set from 1993 to 2022. Table 1 shows, for each journal: the number of published articles (Papers); number of articles with explicit COI metadata (COI); among these, articles with actual information in that field as opposed to stating "we have nothing to disclose" (∃COI); number of papers containing phrases of interest (Phrases) ; number of triples we extracted from paper (#T_Paper), respectively, from explicit COI statements (#T_COI). Only about **1%** of papers contain useful COI metadata, whereas roughly **99%** include relevant phrases in the PDF. Although this may be slightly overestimated ($\approx 10\%$ error; see Limitations), such phrases remain far more frequent, underscoring the value of harvesting them in addition to COI metadata, unlike prior studies (Perlis et al., 2005; Hardy et al., 2023).

	Papers	COI	∃COI	Phrases	#T_Paper	#T_COI
RTP	2721	85	30	2618	11044	126
TAP	5753	240	10	5524	21561	70

Table 1: Dataset statistics

Models We test the Gemma 2 models (Team et al., 2024) of 2B, 9B, and 27B and GPT, the `gpt-3.5-turbo-0125` version. We run the models with temperature of 0. Gemma 2 27B is run in 8 bit precision and the other models in full precision. We ran the Gemma models on a V100 32G GPU and the GPT model using the OpenAI API.

4.1 Evaluation of Triple Extraction

We evaluate the quality of the triples extracted using our tailored prompt (see Appendix E.1) for *relations of interests* extraction. We used the gold set proposed in (Upadhyay et al., 2023) for the assessment. The gold set consists of 1113 pairs of *(sentence, triple)* obtained by annotating 282 conflict-of-interest statements from Pubmed. We used two well-established approaches for performance evaluation of the *open information extraction* methods – Carb (Bhardwaj et al., 2019) and Wire57 (Lechelle et al., 2019), see Appendix C for a detailed description of the metrics.

We compare the performance of different LLMs against a BERT-based method (Upadhyay et al., 2023) and report the results in Table 2 and 3. GPT-3.5 outperforms all the other methods. The notable drop in precision of Gemma 2 (27B) is due to the higher number of extractions per test sample in comparison to the other models. We note that the differences in the performance of the baseline OpenIE method as compared to the reported performance in the original work (Upadhyay et al., 2023) is attributed to minor issues that we fixed in the

Model	Precision	Recall	F1 Score
GPT3.5	**0.761**	**0.852**	**0.804**
Gemma 2 (2B)	0.760	0.803	0.781
Gemma 2 (9B)	0.731	0.842	0.783
Gemma 2 (27B)	0.675	0.805	0.734
OpenIE (Upadhyay et al., 2023)	0.722	0.699	0.711

Table 2: Carb metric on the triple extraction task

Model	Precision	Recall	F1 Score
GPT3.5	**0.755**	**0.782**	**0.768**
Gemma 2 (2B)	0.750	0.669	0.709
Gemma 2 (9B)	0.737	0.764	0.750
Gemma 2 (27B)	0.679	0.750	0.713
OpenIE (Upadhyay et al., 2023)	0.706	0.585	0.640

Table 3: WiRe57 metric on the triple extraction task

gold set and in the implementation of Wire57.

4.2 Qualitative Evaluation of KB Construction

We select a set of 200 triples by randomly choosing an equal number of triples from our corpus and existing IE task benchmark (Upadhyay et al., 2023). We used this **evaluation set** for both qualitative and quantitative analyses of our proposed method.

Predicate Normalization Recall that predicate normalization includes *base predicate extraction* from predicate phrases, and then *normalized predicate set generation* by clustering those base predicates. We evaluate them separately, and then provide an end-to-end evaluation (in Section 4.3).

a) Extracting base predicates from predicate phrases The triples extracted from RTP, respectively, TAP journal articles contain 2862, respectively, 3563 distinct predicate phrases, out of which 458 are common. As explained in Section 3, we need to extract the base predicate and features from each predicate phrase.

To evaluate the quality of our prompting approach, we assign three annotators to each triple of our evaluation set. The annotators were presented with the task description, i.e. the prompt of the model. The annotators all have C1 level of English and higher education, however they are not native speakers. For measuring *inter-annotator agreement* (IAA), we used the pairwise F1 metric which is widely used for token span-specific tasks like NER, (Brandsen et al., 2020). We evaluated the performance of each model's extractions by comparing them with each annotator's extraction. We report the average metric score for each model in Table 4. We observed a high agreement among the annotators (referred to as Human in the table).

Model	Precision	Recall	F1 Score
GPT3.5	0.777	0.7756	0.776
Gemma 2 (2B)	0.746	0.685	0.714
Gemma 2 (9B)	0.756	0.761	0.758
Gemma 2 (27B)	**0.854**	**0.786**	**0.818**
Human	0.806	0.829	0.817

Table 4: Evaluation on the task of base and feature extraction from a predicate

While Gemma 2 (27B) has the best performance among the the models, we retained GPT3.5 for this task to have a pipeline relying on a single model, and because its advantage on the triple extraction task (Tables 2, 3) was larger. We obtain 2329, respectively, 2450 base predicates from RTP and TAP. Note that multiple predicate phrases can have the same base predicate.

b) Constructing the normalized predicate set Recall from Section 3.3 that we organize base predicates in a hierarchy of clusters, where the clusters at the top of the hierarchy correspond to our smallest set of normalized predicates.

We first encoded the base predicates using the recent Pearl model (Chen et al., 2024), specialized for embedding noun and verb phrases. We used Scikit Learn's agglomeration clustering algorithm, with the *linkage* parameter set to value `ward`. We used Silhouette scores to measure cluster quality (a high score is desirable), while also aiming at a reasonably small number of clusters.

To evaluate the quality of our clustering methodology, we create a **ground truth set of base predicates and their correct labels** as follows. We first generate a potential label for each extracted base predicate of the evaluation set using the Agglomerative clustering. If a base predicate belongs to a cluster with a high Silhouette score, then we assign the cluster representative as its potential label, otherwise, the base predicate itself is its potential ground label. We ask 2 annotators to evaluate whether the assigned potential label was best suited among other labels. We observed substantial inter-annotator agreement with Cohen's kappa score of 0.64. In case of annotator disagreement, a third annotator is asked to select the most suitable label. Recall that our clustering method assigns a *representative* predicate to each base predicate depending on which cluster it belongs to. We compute cosine-similarity of the assigned representative with the ground truth labels and assigning the most similar label as the predicted label. We com-

	t_1	R1 clust.	t_2	R2 clust.	95% clust.
RTP	0.889	432	1.20	406	246
TAP	0.879	480	1.20	422	161

Table 5: Thresholds and numbers of predicate clusters

	Org1	Abbrev.	Org2	Clust2	Total Org
RTP	3400	341	409	121	2924
TAP	7672	723	1185	235	5523

Table 6: Organization clusters

pute the F1 score as our evaluation metric. We tuned the distance threshold by testing multiple values on the evaluation set and selecting the one with the highest F1 score. The best result was obtained at a threshold of 0.812, yielding an F1 score of 0.836. To obtain a more compact and normalized predicate set, we slightly increased the threshold, merging nearby clusters to reduce their number. With this adjustment, the F1 score on the evaluation set was 0.804.

Table 5 shows the thresholds t_1 and the resulting number of round 1 (R1) clusters for both datasets. From this round, we keep clusters with Silhouette score ≥ 0.2, and we re-cluster the predicates from the remaining clusters. In the second round, we used the threshold $t_2 = 1.2$; Table 5 shows the number of clusters after round $R2$. The triple coverage distribution of the resultant clusters is long-tailed; we only keep the largest clusters that, together, cover 95% of triples. The numbers of clusters appear at right in Table 5. In the final predicate lists, 39 are common (**368 distinct predicates** in all). Sample clusters are in Appendix G.

Organization Normalization As discussed in Section 3.3, we normalized organizations in triples by abbreviation containment and agglomerative clustering (see Appendix D). Table 6 shows the numbers of distinct Organizations mentioned in both journals (Org1) and the number of abbreviations under the column Abbrev. The column Org2 shows how many Organizations were clustered based on their containing a common abbreviations, and Clust2 the number of clusters with more than one element.

Taxonomy of COI relationships As stated in Section 2, in the recent (Hardy et al., 2023), human experts proposed a set of 11 COI relations: *analyze, collected data, coordinated, designed, funded, participated in, reviewed, supplied, supplied data, supported*, and *wrote*; we denote this set $\boxed{R_{11}}$. To advance modeling and understanding of COIs, we seek to *organize our normalized predicates in a taxonomy* by relating them to those in R_{11}.

Evaluation Task	Precision
T1 (Extraction Correctness)	0.926
T2 (NER Correctness)	0.949
T3 (Predicate Normalization Correctness)	0.926
T4 (Entity Normalization Correctness)	0.938
End-to-end Correctness	0.802

Table 7: End-to-end evaluation of our methodology

To this end, for each of our 368 normalized predicate, we compute the most similar relation $r \in R_{11}$, as the one with the closest Pearl (Chen et al., 2024) embedding (according to the cosine similarity). We analyzed for each R_{11} relation, their closes match in our predicated. We observed that the match was correct in the following proportions (i.e. the precision): analyze (28/31), collected data (4/9), coordinated (3/14), designed (3/3), funded (49/64), participated in (17/45), reviewed (12/14), supplied (28/75), supplied data (7/14), supported (24/36), wrote (15/20). We note that we considered a match as correct even if the direction of the predicate was wrong, i.e. funded and received funding should be considered as two distinct relations of interest. In the wrongly matched predicates, we observed the following: the embedding similarity works very well, and the actual errors are due to missing relations of interest in the R_{11} relations. Notable missing elements are: is employee of (stronger than funded), is stockholder, reports conflicts of interest (unclear of what kind), has affiliation, acknowledges/is grateful. We also note that the *supported* predicate in R_{11} is very general, and it is sometimes matched with predicates closer to funded. The R_{11} relations lack the richness needed to capture the nuances of the relations stated by the users, in the above metadata and/or in the papers.

4.3 End-to-end Evaluation

We assessed our method on the following evaluation tasks:

T1 Given a pair of triple and associated text, *can this triple be extracted from the text?*

T2 *Are the named entities correctly identified* and labeled in the extracted subject and object?

T3 Does the *normalized form of the predicate correctly capture the semantics of the extracted predicate phrase*?

T4 Is the *normalized form of entity correct*?

We manually evaluated our method on the evaluation set, assigning binary scores for each of the four tasks. For T2 and T4, a triple was scored 1 only if both subject and object annotations were correct. Task-wise precision is reported in Table 7. A triple was considered correct only if it scored 1 on all four tasks. We then conducted a comprehensive error analysis as follows.

For task $T1$, apart from the correctness of the extracted subject, predicate, and object, we also assessed if the extracted predicate is a *relation of interest* that can potentially indicate a conflict of interest. For instance, triple ⟨*FDA, has headquarters at, Maryland* ⟩ is not relevant for our problem domain. We categorized primary errors in triple extractions task T1 into the following categories, **(C1)** erroneous input text, **(C2)** incorrect extracted predicate phrase, and **(C3)** incorrect extracted subject phrase. These 3 types of errors contributed to about 80% of errors in T1 with 26% of type C1, 36% of C2, and 18% of C3. The C1 category error are mainly due to poor text extraction from the PDF. The errors of type C2 and C3 are more indicative of the IE capabilities of LLMs. We observed mainly 3 patterns for type C2 errors – extraction on non-COI predicates (43%), extraction semantically opposite predicates (14%) (e.g., predicate phrase *supported* is extracted instead of *was supported by*), and extraction of incomplete predicate (43%) (e.g., from the text *"authors are grateful to John for technical advice"*, the extracted triple is ⟨ *authors, are grateful to, John* ⟩). The incorrect subject phrases are due to the failure of LLMs to correctly identify subjects in long text inputs.

We observed two types of error patterns for task T2 – first, **(C4)** confusing acronymized forms of author name with Organization entity; second, **(C5)** absence of NE in the incorrect subject or object phrase. We observed 33% errors are on type C4, and the rest are of type C5. The type C5 errors are often when the incorrectly extracted entity (subject/object) phrase does not contain any NE.

In task T3, we evaluated how well the normalized form of a predicate captures the extracted predicate phrase. This task holistically evaluates the quality of both the substeps of predicate normalization – base predicate extraction, and our clustering-based predicate normalization method. We categorized frequent errors for T3 in two types – **(C6)** assignment of semantically close but not correct normalized predicate, and **(C7)** assignment of inverse normalized predicate to the extracted predicate phrase. For instance, in type C6 errors, the predicate phrase *co-funder* is normalized to *co-owner* predicate. While both the predicates are semantically close, they do not necessarily *always* represent the same relation in real-world. The type C7 errors are of the form where predicate like, *was funded by* is assigned a normal form *funded*. Here, *was funded by* is the inverse relation of *funded*. Among all the errors in task T3, 39% are of type C6, and 46% are of type C7. Due to the deterministic nature of type C7 errors, they are easier to handle. For instance, an incorrectly normalized triple ⟨ this work, funded, FDA⟩ can be fixed by inspecting the NE type of subject and object. Here, the subject NE type of predicate *funded* can only be Organization. This wrongly normalized triple can be corrected by interchanging the subject and object, i.e. ⟨ FDA, funded, this work⟩.

	Company	Non-Company	Coverage
RTP	32.63%	36.84%	68.98%
TAP	9.28%	55.73%	65.02%

Table 8: Organization labels in COI statements.

The errors in task T4 can mostly be attributed to the errors in task T1 and T2. We categorized errors into, first, **(C8)** incorrect normalized entity due to error in entity (subject/object) extraction by LLM, and, second, **(C9)** error in correctly identifying the NE in the entity phrase. These errors are expected as the correctness of normalization of the entity is dependent on the correctness of the entity phrase extraction and, subsequently, named entity recognition from the extracted phrase. We observed 54% and 36% of errors in T4 are of type C8 and C9.

Based on the analysis, the key takeaways are: **a)** to mitigate the garbage-in-garbage-out issue, a PDF text extractor with better layout identification of journal articles is desired, **b)** prompts with explicit intent for the OpenIE task and the input data (check prompt in Appendix E.1) yields quality results, **c)** postprocessing is essential as LLMs still confuse syntactically similar but semantically opposite phrases.

5 Comparative Analysis of the Journals

In this section, we investigate if our extracted triples of relations of interest can serve as useful input for computational social science investigations of publishing practices.

Company vs. Non-Company Organizations Our

	Company	Non-Company	Coverage
RTP	47.41%	44.57%	91.97%
TAP	5.72%	84.93%	90.65%

Table 9: Distribution of affiliations in papers (all same-affiliation authors of one paper count as 1).

corpus contains organizations (i) as employers (affiliations) of the authors, and (ii) acknowledged in COI statements, whether in the metadata or from the PDF articles. Their form varies; organization name are sometimes paired with different cities and countries. We normalize them as follows. First, we extract the *city* and *country* of each entry using GPT 3.5 and we create a corresponding entry of the form *(orgName, city, country)*. One organization can have multiple divisions at different geographical locations, for example *Pfizer Worldwide Research and Development, Sandwich, UK* and *Pfizer Inc, Cambridge, USA* are different entities under the same umbrella company. To avoid false fusions of organizations, we group the (orgName, city, country) by city and country, and then, for each city and country, we cluster the affiliations.

After normalization, we classify each organization as **company** or **non-company**, with a cascade of three methods, as follows. First, we look for keywords such as *Company*, *Corporation*, in the organization name, and if one is present, we classify the organization as Company; similarly, if *University* or similar terms are present, we consider it is not a company. The lists of keywords used for assigning both labels appear in Appendix F. Second, for each remaining organization, we query the Wikidata online KB with the organization name, and fetch its `instance of` (wdt:P31) property values (human-readable type names); in these types, we search with the same keywords, and classify accordingly. Third, for those organizations absent from Wikidata, we prompt GPT-3.5 for a classification into **company**, **non-company**, and **unsure**.

Company Involvement in the Journals Table 8 shows the distribution of company and non-company organizations in the COI statements of the two journals, together with the coverage (percentage of organizations labeled either company or non-company). We see that **companies are three times more frequent in RTP COI statements than in TAP's**. Complementing this, Table 9 studies the distribution of companies among the author affiliations, counting an author affiliation once for each paper. The table shows that **company employers are nine times more present** in RTP than in TAP.

	No comp. affil.	≥1 comp. affil.	Comp-only affil.	Comp. COI
RTP	37.63%	50.03%	38.96%	39.11%
TAP	79.20%	3.45%	1.85%	13.49%

Table 10: Company involvement in individual articles.

The coverage for affiliation organizations, around 90%, is much better than for COI-mentioned ones, around 65%. This is because organization names are declared with more care in the author affiliation metadata than in paper phrases; also, phrases in the paper often acknowledge grants or funding programs, whose classification as company or non-company is harder. Table 10 shows a global per-article perspective: how many articles have *no company affiliation*, *at least one*, *only company affiliations*, and *a COI with a company*. The difference is striking, e.g., **the frequency of company-only articles in RTP is 20 times higher than in TAP**.

Finally, we have computed the frequencies of **(affiliation, COI-acknowledged organization)** pairs, where a pair (a, o) results as soon as in one article with one author affiliated to a, organization o is acknowledged. Among the **50 most frequent pairs**, in TAP, only **5** involve companies: Merck (acknowledged by a Merck branch in Germany, and an Italian research institute), GlaxoSmithKline (acknowledged by a GSK branch, and the University of Surrey in the UK), and Genentech (acknowledged by a GSK branch). The others involve public agencies funding science, ministries of research, etc. In contrast, in the 50 most frequent pairs in RTP, **48** involve companies, notably: Amgen, BIAL, GSK, American Tobacco, and LRSS, a consortium of cosmetics companies.

6 Conclusion

We have presented an OIE approach, based on LLM prompting, clustering, and external knowledge bases, for extracting COI triples from disclosure statements. We propose a new KG, and a set of 300+ fine-granularity relationship predicates. Finally, we show a vastly stronger industry presence in a journal than in another one on the same topics. Our resources should enable more social science studies on this important topic.

Acknowledgments. We thank Gary Fooks, Professor of Criminology, and Stéphane Horel, journalist, for their valuable insights on the scientific articles in the domains of toxicology and pharmacology. The financial support for this work is provided by DATAIA, Hi!Paris center, and BITS Pilani New Faculty Seed Grant.

7 Limitations

Our **list of keywords (Appendix A)** used to select phrases containing acknowledgments (Section 3.1) gave good results on the data we considered, but it may need to be changed for new settings to ensure no relevant phrase is missed.

The **keywords used for our company/non-company classification** (Appendix F, Section 4.2) may also be imperfect; for instance, while *institute* indeed corresponds to non-company actors in our experience, some counter-examples may exist.

Further, we found that *in a few rare cases, one organization featured both a company and a non-company keyword*. In our dataset, this occurred in exactly two cases:

- "United States Environmental Protection Agency, Washington D.C, USA", which, in Wikidata, has the types *government agency* and *publishing company*. The second is clearly wrong, but errors (and in particular ontology and typing errors) are known to exist in large knowledge bases such as Wikidata. Similarly,
- "Mayo Clinic, Rochester, USA" has the Wikidata types *educational institution* and *business*. Here, both types are correct; this is a business to its patients, and a training ground for medical doctors ("institution" matched our "Institute" keyword).

We curated these by hand, considering them as *non-company*.

The Mayo Clinic example also highlights that the education-related (University, College, Department) keywords may prevent institutions that are educational and for-profit (thus, companies) to be considered as companies. Different choices can be made here, possibly based on a wider use of external resources, e.g., harvesting information from the Web, etc. to clarify an organization's status.

The **extraction of phrases from PDF articles** lead to some errors in the RTP papers. Specifically, because Acknowledgments are often towards the end, where the bibliography starts, some paragraphs we kept based on our filtering (Section 3.1) comprise some references.

As a consequence, some Organizations mentioned in the RTP COI statements are in fact journals. They do not impact our findings in Section 5, since they are classified *unsure*, and we manually discarded them when building the top-50 most frequent (affiliation, acknowledged organization) pairs. We could extend our LLM-based classification to look also for journal as a possible class, in order to automate the weeding out of such erroneous data.

References

Angelos Christos Anadiotis, Oana Balalau, Francesco Chimienti, Mhd Yamen Haddad, Stéphane Horel, Youssr Youssef, Théo Bouganim, Ioana Manolescu, and Helena Galhardas. 2021. Discovering Conflicts of Interest across Heterogeneous Data Sources with ConnectionLens. In *ACM International Conference on Information and Knowledge Management (CIKM 2021)*, Online, Australia.

Vahan Arsenyan, Spartak Bughdaryan, Fadi Shaya, Kent Wilson Small, and Davit Shahnazaryan. 2024. Large language models for biomedical knowledge graph construction: Information extraction from emr notes. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, page 295–317. Association for Computational Linguistics.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam. 2019. CaRB: A crowdsourced benchmark for open IE. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6262–6267, Hong Kong, China. Association for Computational Linguistics.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Lihu Chen, Gaël Varoquaux, and Fabian M. Suchanek. 2024. Learning high-quality and general-purpose phrase representations. In *EACL Findings*, pages 983–994. Association for Computational Linguistics.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction.

S. Scott Graham, Zoltan P. Majdik, Joshua B. Barbour, and Justin F. Rousseau. 2022. Associations between aggregate NLP-extracted conflicts of interest and adverse events by drug product. *Studies in Health Technology and Informatics*, 290:405–409.

S. Scott Graham, Jade Shiva, Nandini Sharma, Joshua B. Barbour, Zoltan P. Majdik, and Justin F. Rousseau. 2024. Conflicts of interest publication disclosures: Descriptive study. *JMIR Data*, 5:e57779.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The Llama 3 herd of models.

Hardy Hardy, Derek Ruths, and Nicholas B. King. 2023. Who controlled the evidence? question answering for disclosure information extraction. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 340–349. PMLR.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020a. Openie6: Iterative grid labeling and coordination analysis for open information extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3748–3761. Association for Computational Linguistics.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020b. IMoJIE: Iterative memory-based joint open information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5871–5886, Online. Association for Computational Linguistics.

Yassir Lairgi, Ludovic Moncla, Rémy Cazabet, Khalid Benabdeslem, and Pierre Cléau. 2024. itext2kg: Incremental knowledge graphs construction using large language models. Springer-Verlag.

William Lechelle, Fabrizio Gotti, and Phillippe Langlais. 2019. WiRe57 : A fine-grained benchmark for open information extraction. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 6–15, Florence, Italy. Association for Computational Linguistics.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Yanchi Liu, Wei Cheng, Haoyu Wang, Zhengzhang Chen, Takao Osaki, Katsushi Matsuda, Haifeng Chen, and Liang Zhao. 2023. Improving open information extraction with large language models: A study on demonstration uncertainty.

Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. 2023. PIVOINE: Instruction tuning for open-world entity profiling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15108–15127, Singapore. Association for Computational Linguistics.

Margaret McCartney. 2024. *Conflicts of interest in healthcare-where are we now?* Ph.D. thesis, The University of St Andrews.

OpenAI, Josh Achiam, Steven Adler, et al. 2024. GPT-4 technical report.

Liu Pai, Wenyang Gao, Wenjie Dong, Lin Ai, Ziwei Gong, Songfang Huang, Li Zongsheng, Ehsan Hoque, Julia Hirschberg, and Yue Zhang. 2024. A survey on open information extraction from rule-based model to large language model. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9586–9608.

R.H. Perlis, C.C. Perlis, Y. Wu, C. Hwang, M. Joseph, and A.A. Nieremberg. 2005. Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. *American Journal of Psychiatry*, 162:1957–1960.

Ji Qi, Kaixuan Ji, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Lei Hou, Juanzi Li, and Bin Xu. 2023. Mastering the task of open information extraction with large language models and consistent reasoning environment.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Prajna Upadhyay, Oana Balalau, and Ioana Manolescu. 2023. Open information extraction with entity focused constraints. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1285–1296, Dubrovnik, Croatia. Association for Computational Linguistics.

Yangqin Xun, Janne Estill, Joanne Khabsa, Ivan D Florez, Gordon H Guyatt, Susan L Norris, Myeong Soo Lee, Akihiko Ozaki, Amir Qaseem, Holger J Schünemann, et al. 2024. Reporting conflicts of interest and funding in health care guidelines: The RIGHT-COI&F checklist. *Annals of internal medicine*, 177(6):782–790.

Alexander Yates, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. 2007. TextRunner: Open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 25–26, Rochester, New York, USA. Association for Computational Linguistics.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An LLM-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: recent capabilities and future opportunities. *World Wide Web*, 27(5).

A Paragraphs of Interest

From the text obtained from the PDF of the papers, we have retained paragraphs that start with one of the following: “Acknowledgement", “Conflict of Interest", “Statement of Interest", “Competing Interest", “Declaration of Conflict of Interest", “Declaration of Interest", “Disclosure", “Funding", “Compliance" and “Competing Financial Interest".

B Data Cleaning

We perform *co-reference resolution* on the CoI phrases from both the XML metadata and the PDF-derived JSON content. This step ensures that all the pronouns present in the texts we use are associated with the nouns they refer to. For this purpose, we used the Coreferee tool[2] from the open-source NLP library spaCy. Further, in some of the phrases of interest, article author names are replaced by their initials, e.g., *Rogely Waite Boyce* is written as *RWB* or *Bruno L. Abbadi* is referred to as *L.A.B* in the articles. We resolve short forms by matching different permutations of the first character of the names of all the authors of an article, as in other works on PubMed data, e.g., (Perlis et al., 2005).

C Extraction Metrics

WiRe57 is a one-to-one matching metric where each system extraction is paired with a single gold extraction if they share at least one token in the subject, relation, and object. Precision measures the proportion of system tokens in the gold extraction, while recall captures the percentage of gold tokens in the system output. CaRB, in contrast, adopts a many-to-one matching approach, allowing multiple gold extractions to map to a single system extraction for recall calculation, preventing penalties when a system extraction corresponds to multiple gold extractions. As a result, CaRB typically yields higher recall than WiRe57. Reporting both ensures a balanced evaluation.

D Organization Normalization Clustering

The agglomerative clustering of these clusters' representatives, together with all the other Organization names (which did not contain abbreviations) used a distance threshold of 0.4. We used the Silhouette score threshold value of 0.5 and 0.57 to cluster the organization names in RTP and TAP triples respectively. The threshold values are decided based on the manual inspection of the clusters. Our clustering is *conservative*, i.e., we primarily avoided false positives, that would cluster together names of different Organizations. Overall, this reduced number of Organizations by 14% and 28% in the RTP and TAP sets, respectively.

[2] https://spacy.io/universe/project/coreferee

E Open Information Extraction Prompts

E.1 Relation Extraction Prompt

We prompt the LLM model to generate the natural language description of the task. In a chat session, we ask LLM to perform an Open IE task, and based on the output, we instruct the LLM to perform better on the same input. After multiple rounds of improvement of the extractions, we ask the model to differentiate between its most successful attempt at the task and another low-quality output task. From the response of the model of this question, we extracted the description of the tasks and used that as part of the main prompt for the IE task. The prompt we developed for *relation extraction* has four main blocks:

- **Task description:** The description is the following: *“Can you extract atomic triples in the format of <subject,predicate,object> in the following text? By choosing more specific predicates, aim to create triples that better reflect the nuances of the relationships described in the text. Please use the information only available in the text."*
- **Data context description**: We inform the model of the source of the text and the intent for extracting the triples. We believe this can help the model focus on specific relations that are relevant to our problem. The description is the following:
  “These sentences are from the journal articles. The intent of the triple extraction is to capture the conflict of interest relations among the authors and organizations."
- **Output format:** We ensured the ease of parsing the LLM response, by providing detailed instructions for the model to output triples in tabular form. The description of the output is the following:
  “The output should be a table with 3 columns

where first column stores the subject of the extracted triple, second column has the predicate and the third column stores the corresponding object."

- **In-context example:** We add a single example with the input and the expected output in the tabular format. The example is as following:

  Input text: Menno V. Huisman reports unrestricted grants from and personal fees from Boehringer Ingelheim, Pfizer BMS, Bayer Health Care, Aspen and Daiichi Sankyo, outside the submitted work.

  Extracted triples:
  <Menno V. Huisman,reports unrestricted grants from,Boehringer Ingelheim>,
  <Menno V. Huisman,reports personal fees from,Boehringer Ingelheim>,
  <Menno V. Huisman,reports unrestricted grants from,Bayer Health Care>,
  <Menno V. Huisman,reports personal fees from,Bayer Health Care>,
  <Menno V. Huisman,reports unrestricted grants from,Pfizer BMS>,
  <Menno V. Huisman,reports personal fees from,Pfizer BMS>,
  <Menno V. Huisman,reports unrestricted grants from,Bayer Health Care>,
  <Menno V. Huisman,reports unrestricted grants from,Aspen>,
  <Menno V. Huisman,reports personal fees from,Aspen>,
  <Menno V. Huisman,reports unrestricted grants from,Daiichi Sankyo>,
  <Menno V. Huisman,reports personal fees from,Daiichi Sankyo>

  We note the triples are given in a tabular format in the original prompt.

E.2 Prompt for Named Entity Recognition

The prompt we have used for named entity recognition is the following:

"In the following sentence please perform named entity recognition. Return the output in a tabular format where the table contains 2 columns, the first column with the named entity, the second column with the named entity type."

E.3 Prompt for predicate normalization

The prompt we have developed for predicate normalization is the following:

- **Task description:** *"Can you split the predicate phrases into the main predicate and its attributes such that the main predicate contains the conflict of interest relation concisely with context?"*

- **Context:** *"Please maintain the context and clarity of the original phrase. In the following examples, note that for some of the cases it is important to maintain a detailed context and for others it is not required. While extracting the main predicate, use your judgement to maintain the context or not. Remember that we are trying to get these extractions for capturing conflict of interest relations."*

- **Output format:** We ensured the ease of parsing the LLM response, by providing detailed instructions for the model to output triples in tabular form. The description of the output is the following:

  "can you format the output in tabular form with 4 columns, first column for the main predicate, second column for the attribute, third column for the attribute type and the fourth column indicates if you have included the detailed context in the main predicate or not."

- **In-context example:** *Following are some examples of this task:*

  `Example 1`
  Predicate phrase: received NIEHS Grant from
  main predicate: received Grant
  attribute: NIEHS Grant
  attribute type: noun
  isContextMaintained: No

  `Example 2`
  Predicate phrase: provided supplemental funds through
  main predicate: provided fund
  attribute: supplemental
  attribute type: adjective
  isContextMaintained: No

  `Example 3`
  Predicate phrase: provided animal care support for
  Main Predicate: provided animal care support
  Attribute: none
  Attribute Type: none
  isContextMaintained: Yes

Example 4
Predicate phrase: partially supported by funding provided by
Main Predicate: supported by funding
Attribute: partially
Attribute Type: adverb
isContextMaintained: No

Example 5
Predicate phrase: acknowledge the facilities, scientific and technical assistance of
Main Predicate: acknowledge the facilities, scientific and technical assistance
Attribute: None
Attribute Type: None
isContextMaintained: Yes

Example 6
Predicate phrase: acknowledges helpful english writing and secretarial assistance from
Main Predicate: acknowledges English writing and secretarial assistance
Attribute: helpful
Attribute Type: adjective
isContextMaintained: Yes

Example 7
Predicate phrase: received economic support from
Main Predicate: received economic support
Attribute: None
Attribute Type: None
isContextMaintained: Yes

Example 8
Predicate phrase: was provided financial support for study by
Main Predicate: was provided financial support
Attribute: None
Attribute Type: None
isContextMaintained: No

Example 9
Predicate phrase: was provided financial support by
Main Predicate: was provided financial support
Attribute: None
Attribute Type: None
isContextMaintained: Yes

F Affiliation Classification Keywords

The keywords used to classify affiliations as company or non-company are:

- Company: *Limited, Company, Companies, LLC, LLP, Corporation, Ltd, Consultancy, Consulting, Consultant*
- Non-Company: *Institute, College, Department, School, University, Ministry, National, Federal, Government, Facility*

G Example of predicate clusters

We present a few examples of the output of our multiround clustering in Table 11. The representative predicate (Column 1) represents the normalized form of all the predicates that belongs to the cluster. The second column (clusters merged count) contains the number of cluster that are merged in round 2. The last three clusters in the table are obtained after round 1 of clustering, therefore the cluster rep. count is 0. The base predicate count specifies the number of base predicates that are normalized to the corresponding representative predicate. Depending on whether the cluster formed after round 1 or after round 2, the value on the rightmost column (Clustered predicate) will contain the list of base predicates or the list of clustered representatives respectively.

Representative predicate	Clusters merged count	Base predicate count	Clustered (base/representative) predicates
supported by grant	22	356	['was supported by grants', 'supported by grants', 'received grant', 'was supported by grant', 'received grants', 'supported by grant-in-aid', 'grant number', 'acknowledge grant support', 'receive grant', 'acknowledge grant', 'supported by grant', 'received research grant', 'recipient of grant', 'supported by grant numbers', 'grant', 'has grant', 'supported by a grant', 'acknowledges research grant', 'received grant-in-aid', 'receives grant', 'supported by research grant', 'funded by grant']
reports conflicts of interest	3	11	['declared competing interest', 'reports conflicts of interest', 'have conflict of interest']
had involvement in analysis	5	48	['had involvement in data interpretation', 'performed analysis', 'had involvement in analysis', 'sponsored analysis of data', 'had involvement in collection of data']
thanked for technical assistance	4	32	['thank for technical peer review', 'provided technical assistance', 'acknowledge technical assistance', 'thanked for technical assistance']
is a consultant	5	36	['served as a consultant', 'is a consultant', 'is an independent consultant', 'is paid consultant', 'is a consultant retained']
acknowledge assistance	0	15	['gratefully acknowledge assistance', 'acknowledge assistance', 'acknowledge support and assistance', 'acknowledges help', 'gratefully acknowledge help', 'acknowledged help', 'acknowledge help', 'acknowledge help and support', 'acknowledge the help', 'acknowledge assistance and advice']
received scholarship	0	15	['was awarded ernst mach scholarship', 'received research scholarship', 'received scholarships', 'receives scholarship', 'recipient of scholarship', 'received bursary', 'received scholar award', 'received scholarship', 'was awarded scholarship', 'received graduate student bursary', 'recipient of a scholarship', 'were recipients of scholarship', 'received graduate scholarship', 'received scholarship award']
acknowledge contributions	0	7	['acknowledge contribution to development', 'acknowledge contribution', 'recognise contributions', 'acknowledge vital contribution', 'acknowledge contributions and support', 'acknowledge significant contributions', 'acknowledge contributions', 'acknowledged contribution']

Table 11: Examples of clusters build by our multi-round clustering method