GENRES: Rethinking Evaluation for Generative Relation Extraction in the Era of Large Language Models

Anonymous ACL submission

Abstract

The field of relation extraction (RE) is experiencing a notable shift towards generative relation extraction (GRE), leveraging the capabilities of large language models (LLMs). However, we discovered that traditional relation extraction (RE) metrics like precision and recall fall short in evaluating GRE methods. This shortfall arises because these metrics rely on exact matching with human-annotated reference relations, while GRE methods often produce diverse and semantically accurate relations that differ from the references. To fill this gap, we introduce GENRES for a multi-dimensional assessment in terms of the topic similarity, uniqueness, granularity, factualness, and completeness of the GRE results. With GENRES, we empirically identified that (1) precision/recall fails to justify the performance of GRE methods; (2) human-annotated referential relations can be incomplete; (3) prompting LLMs with a fixed set of relations or entities can cause hallucinations. Next, we conducted a human evaluation of GRE methods that shows GENRES is consistent with human preferences for RE quality. Last, we made a comprehensive evaluation of fourteen leading LLMs using GENRES across document, bag, and sentence level RE datasets, respectively, to set the benchmark for future research in GRE.

1 Introduction

Relation Extraction (RE) is one of the most critical tasks in natural language processing (Han et al., 2020). In essence, RE transforms unstructured text into structured, actionable knowledge (e.g., knowledge graphs). However, the traditional RE methods only mine the predefined patterns referring to the predefined sets of relations and entities, thus often struggling to capture the complexity of natural language. Recently, Large Language Models (LLMs) like GPT (OpenAI, 2023), promise a transition to Generative Relation Extraction (GRE). LLM-based GRE methods are capable of comprehending the input texts and then identifying complex relationships without the constraints of predefined patterns in a zero-shot manner. This is particularly advantageous when there is a scarcity of training data, and the input texts are varied.

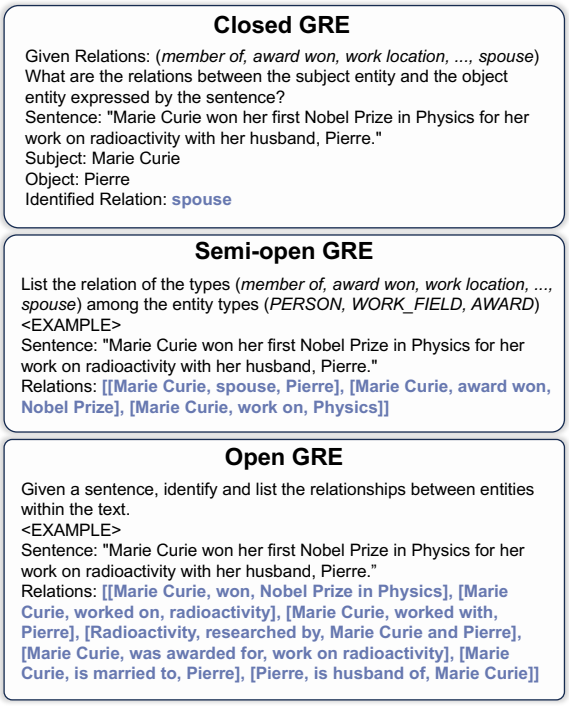

Figure 1: **Generative Relation Extraction (GRE):** Contrasting Closed and Semi-open GRE's type constraints with Open GRE's reliance on source text alone.

Existing applications of LLMs in GRE are either performing binary classification tasks (Li et al., 2023a) given entity pairs and a set of predefined relation types, or given restricted entity types (Wadhwa et al., 2023a; Zhu et al., 2023), which overlook extensive novel relations and entities beneath the text. Notably, to unlock the full power of LLMs in GRE, we advocate a transformation from "defining a set of relation types" → "finding matches between entities" to "exploring as many relations and entities as possible without limitation" → "re-

finement" (Paulheim, 2017; Liu et al., 2018). This strategy elicits LLMs' implicit knowledge to discover a wider array of relationships with minimal predefined constraints (Hao et al., 2023), which we define as "Open GRE" that can be applied to knowledge graph construction for various downstream tasks (Baralis et al., 2013; Wang et al., 2018; Mohamed et al., 2020; Zeng et al., 2022; Jiang et al., 2023b). We illustrate the difference of GRE strategies in Figure 1.

The versatility of GRE, however, poses significant challenges in evaluation (Wadhwa et al., 2023a). Specifically, we identified that traditional relation extraction (RE) metrics like precision and recall only capture the exact matching with human-annotated reference relations, while GRE methods often produce diverse and semantically accurate relations that differ from the references. As such, we argue that precision in GRE should be verified against the source text, and recall should be based on soft matching to accommodate the output flexibility of generative models. Furthermore, a proficient model should not only cover crucial information in the text but also avoid redundant results, ensuring the extracted knowledge is both comprehensive and atomistic. To navigate these new dimensions, we introduce **GENRES** (**GEN**erative **R**elation **E**xtraction **S**coring), a multi-dimensional framework tailored for evaluating GRE. Our key contributions are as follows.

- We demonstrate the effectiveness of GENRES for evaluating GRE tasks, emphasizing its superiority over traditional metrics.
- We benchmark the open GRE performance of fourteen leading LLMs through GENRES, and paving the way for future research and development of better LLM-based GRE methods.

2 Preliminaries

Definition 1 (Source Document) *A source document $\mathcal{D}$ is a piece of free-text, which can be a sentence, a passage, or a document.*

Definition 2 (Extracted Triples) *A triple $\tau = \langle s|r|o\rangle$ is a structure formatting a piece of free text into a subject s, a relation r, and an object o. Example: For a sentence "Alice lives in Champaign.", "Alice" is the subject, "live in" is the relation, and "Champaign" is the object. Together, they form a triple $\langle Alice|live\_in|Champaign\rangle$. We define $\mathcal{T}_\mathcal{D} = [\tau_1, \tau_2, ...]$ as a list of triples extracted from the source document $\mathcal{D}$.*

2.1 Generative Relation Extraction

GRE uses a generative large language model (LLM) to extract relational triples from a source document $\mathcal{D}$. The model functions on an autoregressive basis at the token level, expressed as $P(x_t|x_1, x_2, \ldots, x_{t-1}, \mathcal{D})$, where x_t represents the t^{th} token in the output sequence. The process generates a sequence of tokens that are structured into triples $\mathcal{T}_\mathcal{D} = [\tau_1, \tau_2, \ldots]$. We categorize existing GRE methods as follows:

- **Closed GRE** (Li et al., 2023a): Given (1) source context, (2) entity pairs in the context, and (3) a set of predefined relation types, prompt the LLM to classify the relation type between the entity pairs to compose each triple τ_i.
- **Semi-open GRE** (Wadhwa et al., 2023a): Given (1) source context, (2) a predefined set of relation types, and (3) a predefined set of entity types, prompt the LLM to extract triples τ_i.
- **Open GRE**: Given source context, prompt the LLM to extract triples as many as possible.

3 GENRES

Evidenced by previous work conducting semi-open GRE (Wadhwa et al., 2023a), traditional metrics for RE like hard matching precision/recall/F1 are inadequate to evaluate GRE tasks as the LLM generations are flexible. To fill in this gap, we introduce GENRES, an automated multi-aspect evaluation framework for GRE. GENRES are composed of a series of sub-scores defined as follows.

3.1 Topical Similarity Score

We compute the topical similarity score (TS) to measure the information abundance of the extracted triples $\mathcal{T}_\mathcal{D}$ compared to the source text $\mathcal{D}$. Here, we employ a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003), an algorithm that represents each document as a blend of a certain number of latent topics, for topic modeling. We concatenate the elements in each triple so that $\mathcal{T}'_\mathcal{D} = [\tau'_1, \tau'_2, ...] = [s_1 \oplus r_1 \oplus o_1, s_2 \oplus r_2 \oplus o_2, ...]$. TS is computed as:

$$t(\mathcal{D}, \mathcal{T}'_\mathcal{D}) = e^{-\sum_{i=1}^{K} LDA(\mathcal{D})_i \cdot \log\left(\frac{LDA(\mathcal{D})_i}{LDA(\mathcal{T}'_\mathcal{D})_i}\right)} \quad (1)$$

which is based on the *KL-divergence* of two topical distributions. A higher TS indicates that the extracted triples closely align with the topical content of the source document, reflecting effective and

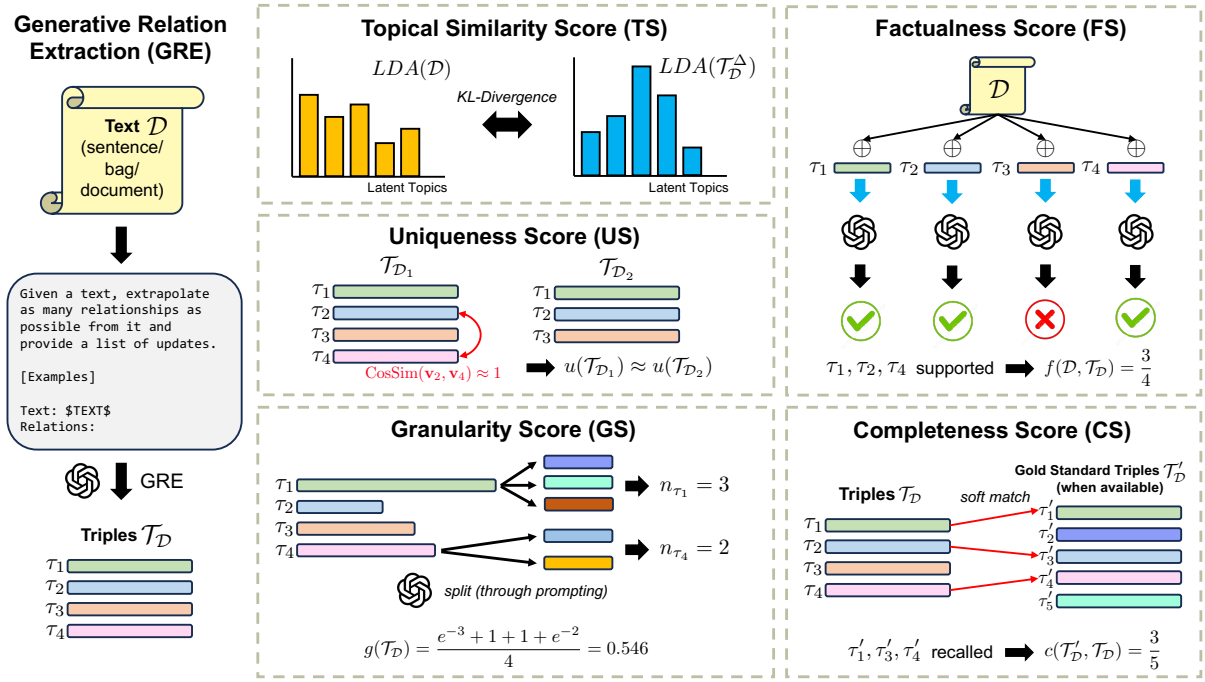

Figure 2: **GENRES framework for the evaluation of generative relation extraction (GRE).** ***Left***: An example showing the GRE process to extract triples $\mathcal{T}_\mathcal{D}$ from a source text $\mathcal{D}$ through prompting generative large language model. ***Right***: illustration of sub-scores contained in GREScore regarding: Topical Similarity (§3.1), Uniqueness (§3.2), Fatualness (§3.3), Granularity (§3.4), and Completeness (§3.5).

relevant information extraction, while a lower TS suggests that the extracted triples may be missing key topical elements from the source.

3.2 Uniqueness Score

Uniqueness Score (US) assesses the diversity of the extracted triples $\mathcal{T}_\mathcal{D}$ in the GRE, emphasizing the importance of extracting varied and distinct relationships. Given $\mathcal{T}_\mathcal{D} = [\tau_1, \tau_2, \ldots, \tau_n]$, with each triple τ_i encoded in a vector $\mathbf{v}_i$ using word embeddings, the US is computed as follows:

$$u(\mathcal{T}_\mathcal{D}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (\text{CosSim}(\mathbf{v}_i, \mathbf{v}_j) > \phi) \tag{2}$$

where $\text{CosSim}(\mathbf{v}_i, \mathbf{v}_j)$ is the cosine similarity between the vector representations of triples τ_i and τ_j. ϕ is a predefined similarity threshold. The normalization factor $n(n-1)$ accounts for all pairings where $i \neq j$. A higher US indicates greater diversity among the triples, while a lower US suggests more similarity and potential redundancy.

3.3 Factualness Score

Factualness Score (FS) quantifies the extent to which extracted triples, denoted as $\mathcal{T}_\mathcal{D}$, align with the information in the source text $\mathcal{D}$. This metric is crucial for gauging the hallucinations (Zhang et al., 2023), a phenomenon where LLMs fabricate the content not present in the source text. Building on the foundations laid by prior research (Min et al., 2023; Jiang et al., 2021), FS employs a detailed triple-wise verification process. Each triple τ within $\mathcal{T}_\mathcal{D}$ undergoes a thorough check to confirm whether it is supported by factual evidence in $\mathcal{D}$:

$$f(\mathcal{D}, \mathcal{T}_\mathcal{D}) = \frac{1}{|\mathcal{T}_\mathcal{D}|} \sum_{\tau \in \mathcal{T}_\mathcal{D}} [\![\tau \text{ is supported by } \mathcal{D}]\!] \tag{3}$$

where $[\![\tau \text{ is supported by } \mathcal{D}]\!]$ is an indicator function that returns 1 if the triple is factual and 0 if it is not. In this study, we adopt the approach from previous work (Min et al., 2023) and utilize an LLM as the fact-checking tool. Specifically, we employ GPT-3.5-Turbo-Instruct as the fact checker, with the methodology detailed in Appendix B.2. A high FS signifies that a substantial portion of the extracted triples are factually consistent with the source text. On the contrary, a low FS indicates a higher incidence of hallucinated or unsupported data. Employing this metric is vital to guarantee the reliability and trustworthiness of the information generated by the model.

3.4 Granularity Score

The Granularity Score (GS) evaluates the level of detail of the extracted triples $\mathcal{T}_\mathcal{D}$ from the source text $\mathcal{D}$. It is based on the premise that triples should

capture the optimal granularity of information, not too coarse. The GS aims to penalize triples that are overly broad and could be further split into more precise statements. The process involves an assessment of each triple's potential to be split into more granular sub-triples. This can be performed by prompting an LLM to evaluate if a given triple can be divided into additional, more specific triples. The number of possible splits is represented by n_τ for each triple τ.

The Granularity Score for the extracted triples $\mathcal{T}_\mathcal{D}$ is calculated using the formula:

$$g(\mathcal{T}_\mathcal{D}) = \frac{1}{|\mathcal{T}_\mathcal{D}|} \sum_{\tau \in \mathcal{T}_\mathcal{D}} e^{-n_\tau} \quad (4)$$

where e^{-n_τ} is the exponential decay function based on the number of splits n_τ, which assigns a lower score to triples that can be split into more sub-triples (indicating they are too broad or general). Therefore, a lower Granularity Score indicates that the triples could be broken down further, while a higher score suggests that the triples are at an appropriate level of specificity.

3.5 Completeness Score

The Completeness Score (CS) evaluates how comprehensively the extracted triples $\mathcal{T}_\mathcal{D}$ cover the information present in the source text $\mathcal{D}$. This score is analogous to the recall metric in information retrieval and is particularly important when gold standard triples $\mathcal{T}'_\mathcal{D}$ are available for comparison. CS is assessed by determining the proportion of gold standard triples that are successfully captured by the extracted triples. For each gold standard triple τ', we find the best matching triple τ from $\mathcal{T}_\mathcal{D}$, using cosine similarity of their embeddings as the *soft matching* criterion. If the cosine similarity exceeds a specified threshold ϕ, the triple τ is considered a match. CS is then computed as:

$$c(\mathcal{T}'_\mathcal{D}, \mathcal{T}_\mathcal{D}) = \frac{|\{\tau' \in \mathcal{T}'_\mathcal{D} | \exists \tau \in \mathcal{T}_\mathcal{D}, \text{sim}(\tau, \tau') \geq \phi\}|}{|\mathcal{T}'_\mathcal{D}|} \quad (5)$$

where $\text{sim}(\tau, \tau') = \text{CosSim}(emb(\tau), emb(\tau'))$ calculates the cosine similarity between the embeddings of the extracted triple and the gold standard triple. The threshold ϕ is pre-defined to determine the acceptable level of similarity for a match. A higher CS indicates that the extracted triples effectively capture the complete range of information as represented by the "gold standard". It is worth noting that CS is optional as precise human annotations are expensive and not always available.

4 Experiments

4.1 Datasets

In our evaluation, we examine several RE datasets with a focus on their performance in GRE using test sets enriched with detailed human annotations. These include: **CDR** (Li et al., 2016), a document-level dataset with 1,500 PubMed abstracts highlighting chemical-disease interactions; **DocRED** (Yao et al., 2019), also document-level, derived from Wikipedia and Wikidata, featuring extensive entity, coreference, and relational annotations across 5,053 documents; **NYT10m** and **Wiki20m** (Han et al., 2019), both bag-level[1] datasets from The New York Times and Wikipedia, respectively, with manually annotated test sets; and **TACRED** (Zhang et al., 2017) and **Wiki80** (Han et al., 2018), sentence-level datasets, the former comprising 106,264 examples across various text sources and the latter containing 56,000 instances with 80 relations from Wikipedia and Wikidata. These datasets collectively offer a comprehensive view of RE capabilities across various levels and sources.

We adopt a random sampling method to select the test sets from the above datasets. We randomly choose {200, 500, 800} samples for the document-, bag-, and sentence-level evaluations[2].

4.2 Implementation

For topical similarity score (TS), we train six LDA models with {50, 100, 150, 150, 150, 150} latent topic numbers and {1500, 5051, 11086, 14257, 38140, 22400} samples (document/bag/sentence) for CDR, DocRED, NYT10m, Wiki20m, TACRED, and Wiki80, respectively. For evaluations (US and CS) using word embedding, we retrieve the embedding for each entity and relation in the triple using `text-embedding-ada-002`, and perform element-wise addition to obtain the triple embedding.[3] Based on our tests, we set the similarity threshold ϕ at 0.95. All local LLMs are run

[1] A "bag" of information that share the same entity pair.

[2] For the Wiki20m dataset (bag-level), we deviated from this approach due to the predominance of low-quality random samples, often containing only a single ground-truth triple. We first refined the dataset to include samples with two triples, narrowing it down to 3,526 samples. From this filtered pool, 500 samples were randomly selected.

[3] Concatenation should be employed instead when the direction of the relation is concerned

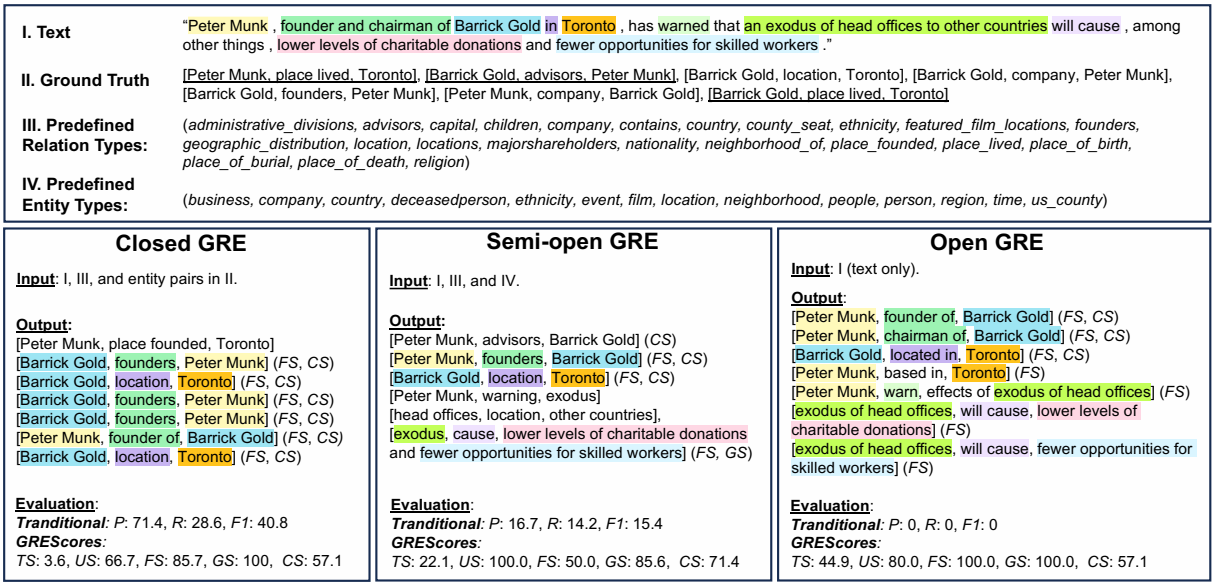

Figure 3: **Comparative Analysis of GRE Methods and Evaluation Metrics using the NYT10m Dataset.** The diagram showcases the outcomes of closed, semi-open, and open Generative Relation Extraction (GRE) strategies. The distinct entity and relation spans are color-coded, with factual triples specifically highlighted. The extracted triples that affect FS, CS (soft recall), and GS are listed with the corresponding labels. We underline the ground truth labels that are inaccurate or cannot be inferred from the source text.

	CDR				NYT10m			
	C	S	O	GT	C	S	O	GT
#tri	10.1	6.8	16.1	10.1	1.4	2.9	5.8	1.4
#tok	6.6	4.0	8.3	5.8	4.6	2.0	7.0	4.5
P	58.8	1.1	0.4	-	29.3	5.2	0.0	-
R	58.7	0.8	0.7	-	26.6	12.7	0.0	-
F1	58.8	0.7	0.5	-	27.5	6.5	0.0	-
TS	11.9	35.5	**77.6**	9.6	10.3	13.4	**54.2**	8.7
US	31.8	58.2	**89.6**	33.4	87.5	**91.5**	83.0	69.3
FS	64.4	62.0	**96.8**	93.5	72.3	33.7	**84.0**	84.1
GS	**92.0**	78.5	54.2	98.1	**87.4**	79.9	71.9	93.1
CS	**58.4***	56.7	47.8	100	**62.3***	20.3	53.4	100

*Closed GRE, due to its use of predefined entity pairs for relation classification, inherently exhibits high triple similarity. Hence, we further check relation embedding similarity for the best soft matching of triples.

Table 1: **Different GRE strategies measured by different metrics including traditional P/R/F1 and GENRES.** "C", "S", "O", and "GT" denote Closed, Semi-open, Open GRE, and ground truth, respectively. GPT-3.5-Turbo-Instruct was used as the LLM. We **highlight** the highest GREScores for each dataset.

on 8 NVIDIA A100 GPUs. All prompts used are detailed in Appendix B.

4.3 Performance of Different GRE Strategies

We conducted evaluations of closed, semi-open, and open GRE on the CDR and NYT10m datasets. The expansive relation sets and the absence of defined entity types in other datasets render them incompatible with closed and semi-open GRE, owing to the limitations of context window constraints. This limitation emphasizes the flexibility of open GRE, which operates unconstrained by predefined relation types or entity types, proving its adaptability to a wider array of datasets. The comparative results of these evaluations are presented in Table 1. Combined with our example shown in Figure 3, we summarize the key observations as follows.

Traditional metrics are not ideal for GRE evaluation, especially in semi-open and open GRE settings. Figure 3 illustrates that despite open GRE's high-quality extractions based on FS and CS, they score zero across these metrics. This occurs because Precision/Recall/F1 depend on exact matching of triples, which are nearly impossible without predefined relation/entity sets, as evidenced by the zero scores for these metrics on the NYT10m dataset in Table 1. This finding syncs with Wadhwa et al. (2023a)'s conclusion.

Human annotations sometimes are unreliable. In Figure 3, we underline several mistakes (e.g., "[Barrick Gold, advisors, Peter Munk], [Barrick Gold, place lived, Toronto]") in the the ground truth where "Barrick Gold" is a company but incorrectly recognized as a person. Such inaccurate labels are unlikely to be correctly predicted by LLMs. This suggests that traditional metrics that purely rely on ground truth triples, are even inadequate for closed GRE, and more so for semi-open and open GRE.

The imposition of predefined relation sets or entity types can misguide LLMs to generate in-

		CDR							DocRED						
		#tri	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***	***#tri***	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***
	Ground Truth	10.1	5.8	9.6	33.4	93.5	98.1	100	12.4	6.0	8.4	64.0	94.4	81.9	100
LLaMA	Vicuna-7B	6.8	8.4	57.8	86.9	84.7	44.6	30.7	7.4	9.9	23.1	81.9	93.4	46.8	28.3
	Vicuna-33B	6.4	10.5	73.0	89.2	**97.3**	38.4	32.0	10.8	9.8	34.7	82.8	97.2	49.6	36.9
	LLaMA-2-7B	5.6	6.7	48.6	92.0	62.0	44.9	25.7	2.7	3.2	12.8	93.3	34.0	60.6	12.1
	LLaMA-2-70B	10.8	8.1	**74.8**	87.6	96.6	**57.8**	**51.0**	13.8	8.7	**39.2**	82.6	**97.3**	**60.9**	**39.2**
	WizardLM-70B	10.2	7.8	65.4	**94.1**	76.4	46.2	32.6	5.8	3.6	24.3	**94.9**	37.9	56.7	12.8
GPT	text-davinci-003	12.7	8.3	76.7	87.2	96.8	**55.4**	44.3	15.3	8.5	40.1	84.2	97.6	59.8	46.2
	GPT-3.5-Turbo-Inst.	16.1	8.3	77.6	89.6	96.8	54.2	47.8	17.8	8.9	47.8	85.6	98.1	56.2	44.7
	GPT-3.5-Turbo	11.2	11.4	81.7	89.2	**98.2**	40.3	30.2	15.0	9.9	**50.4**	84.0	98.5	49.1	36.5
	GPT-4	14.3	9.3	81.7	91.0	97.9	49.1	46.3	17.8	8.7	48.6	82.8	**98.6**	59.6	47.3
	GPT-4-Turbo	18.6	8.5	**82.1**	**91.9**	96.8	53.1	**48.8**	21.5	8.7	50.0	**87.4**	97.6	**63.1**	**49.3**
others	Mistral-7B-Inst.	14.2	9.1	69.0	74.9	93.5	51.1	**40.0**	11.3	9.6	30.2	76.4	94.1	55.2	27.5
	Zephyr-7B-Beta	25.9	8.8	49.1	79.5	70.1	**57.7**	29.3	18.6	8.6	27.9	79.4	94.7	**64.7**	37.1
	Galactica-30B	0.2	0.3	4.1	1.1	0.9	44.4	0.0	0.0	0.0	8.6	0.0	0.0	0.0	0.0
	OpenChat-3.5	8.6	12.6	**78.7**	**91.9**	**97.4**	38.2	31.8	15.4	8.9	**39.7**	**82.1**	**98.1**	61.7	**43.4**

Table 2: **GENRES evaluation of Open GRE on *document-level* datasets.** Scores (%) are averaged across documents. *#tri* and *#tok* denote the number of triples per document and the number of tokens per triple, respectively. We **highlight** the highest within-group scores. Galactica's low scores are due to its limited size of context window.

		NYT10m							Wiki20m						
		#tri	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***	***#tri***	***#tok***	***TS***	***US***	***FS***	***GS***	***CS***
	Ground truth	1.4	4.5	8.7	69.3	84.1	93.1	100	2.0	6.3	4.4	21.2	88.7	85.1	100
LLaMA	Vicuna-7B	3.1	7.8	42.0	86.4	80.0	60.2	38.9	3.0	7.5	48.3	67.8	50.0	68.6	37.3
	Vicuna-33B	4.7	7.2	**47.8**	80.1	75.1	65.2	46.5	4.1	7.0	**49.8**	56.4	84.4	75.4	46.1
	LLaMA-2-7B	3.1	6.0	35.4	82.2	78.9	69.2	38.4	3.1	6.3	37.9	**73.8**	73.4	75.6	36.0
	LLaMA-2-70B	5.0	6.9	45.4	83.0	**81.7**	**71.8**	**52.4**	4.1	6.9	45.2	62.0	**87.1**	**78.4**	**50.2**
	WizardLM-70B	4.4	4.2	30.5	**88.9**	43.9	68.9	27.6	3.6	5.6	43.1	67.8	67.3	75.0	40.9
GPT	text-davinci-003	4.9	7.1	50.6	81.4	85.8	69.3	52.6	3.7	8.2	51.8	56.9	91.3	73.3	43.5
	GPT-3.5-Turbo-Inst.	5.8	7.0	54.2	83.0	84.0	**71.9**	53.4	4.8	7.7	54.0	60.3	90.1	78.9	43.8
	GPT-3.5-Turbo	4.1	6.2	43.3	82.3	68.2	62.8	29.8	3.6	7.7	48.2	61.8	80.2	72.7	32.5
	GPT-4	5.1	7.4	56.2	81.8	89.0	68.2	52.6	3.8	8.1	**59.0**	56.2	**93.2**	77.2	40.0
	GPT-4-Turbo	5.3	7.8	**58.1**	**84.2**	**89.6**	69.1	**53.7**	4.2	7.6	56.4	**62.0**	92.4	**81.2**	**52.7**
others	Mistral-7B-Inst.	5.7	7.4	40.6	77.6	75.4	62.9	36.5	4.0	6.9	43.3	57.0	83.6	69.9	40.1
	Zephyr-7B-Beta	7.8	7.2	36.5	80.8	64.9	**73.8**	47.0	5.2	6.8	40.3	**65.5**	75.5	**79.0**	45.9
	Galactica-30B	8.3	8.7	29.7	48.4	52.4	60.6	37.0	6.0	8.4	35.3	49.4	65.2	66.8	38.6
	OpenChat-3.5	5.2	7.2	**54.0**	**84.7**	**84.3**	69.7	**55.3**	4.3	7.0	**57.5**	61.8	**90.5**	76.0	**47.7**

Table 3: **GENRES evaluation of Open GRE on *bag-level* datasets.** Scores (%) are averaged across bags. *#tri* and *#tok* denote the number of triples per bag and the number of tokens per triple, respectively. We **highlight** the highest within-group scores.

accurate triples. For instance, as seen in Figure 3, closed GRE misclassifies the relation between "Peter Munk" and "Toronto" as "place founded" based on limited choices from the relation set, despite the text not supporting this inference. Similarly, semi-open GRE's entity recognition becomes problematic when it erroneously divides "exodus of head offices" into separate entities "exodus" and "head offices", leading to less coherent and less meaningful triples.

It is also obvious that the range of information captured by extracted triples widens from closed GRE to open GRE. Closed and semi-open GRE, which limit the types of relations or entities, often yield extractions with a narrower scope. This constriction hampers the completeness of the captured information, a fact corroborated by the TS metrics presented in Table 1. Furthermore, providing a more diverse relation set to semi-open GRE, such as the one in NYT10m (as opposed to the more limited CDR, which restricts entity types to chemicals and diseases), results in a significant drop in granularity (GS). In contrast, open GRE maintains stability, underscoring the benefit of eschewing predefined relation/entity types. Although closed GRE records the highest GS and CS, it is benefited from taking extra input entity pairs, which are not provided to simi-open and open GRE.

		TACRED							Wiki80						
		#tri	*#tok*	*TS*	*US*	*FS*	*GS*	*CS*	*#tri*	*#tok*	*TS*	*US*	*FS*	*GS*	*CS*
	Ground Truth	1.4	4.6	15.8	92.7	87.0	94.9	100	1.0	5.8	5.9	100	90.1	84.4	100
LLaMA	Vicuna-7B	2.6	8.7	40.4	85.0	**75.6**	58.9	36.2	2.4	7.9	41.3	76.8	81.0	61.7	36.6
	Vicuna-33B	4.3	7.3	**44.3**	75.5	71.0	69.2	47.2	3.8	7.2	**47.3**	62.1	79.9	73.8	46.8
	LLaMA-2-7B	2.8	6.3	36.7	85.3	66.9	71.2	37.8	2.4	5.8	25.8	69.8	60.4	**76.9**	31.4
	LLaMA-2-70B	4.1	6.4	40.8	79.3	74.5	**76.8**	**56.4**	3.7	6.6	41.5	64.8	**82.4**	**76.9**	**49.4**
	WizardLM-70B	2.1	2.9	23.3	**90.7**	28.0	72.1	9.8	2.1	3.2	25.6	**84.9**	36.6	74.4	21.4
GPT	text-davinci-003	4.4	7.1	56.1	79.8	84.0	72.8	58.6	4.0	6.8	59.2	65.3	89.2	74.9	51.9
	GPT-3.5-Turbo-Inst.	5.0	7.0	58.6	80.5	81.6	72.6	58.6	4.4	6.9	60.2	69.3	88.7	**75.4**	**54.8**
	GPT-3.5-Turbo	3.9	6.8	52.7	81.1	76.4	67.5	39.7	3.4	6.3	50.9	**69.5**	75.6	68.9	36.0
	GPT-4	4.3	7.5	**59.1**	80.4	87.6	69.1	57.8	4.0	7.1	**65.4**	66.2	92.3	74.2	47.8
	GPT-4-Turbo	4.4	7.8	58.5	**82.6**	**88.6**	**73.2**	**63.4**	4.0	7.6	61.9	69.4	**92.8**	74.5	47.1
others	Mistral-7B-Inst.	4.7	7.1	43.9	78.6	71.0	65.5	41.2	3.6	7.8	44.6	67.8	83.9	67.7	38.5
	Zephyr-7B-Beta	5.4	7.6	36.4	78.6	65.8	72.0	44.9	4.5	7.8	43.2	68.1	77.8	74.2	42.6
	Galactica-30B	8.5	8.9	33.4	43.9	57.5	64.1	30.9	5.6	7.2	35.0	47.9	63.1	73.3	38.4
	OpenChat-3.5	4.3	7.1	**50.7**	**80.8**	**80.4**	**72.1**	**60.0**	4.0	7.0	**53.8**	**69.7**	**88.7**	**74.9**	**50.6**

Table 4: **GENRES evaluation of Open GRE on *sentence-level* datasets.** Scores (%) are averaged across sentences. *#tri* and *#tok* denote the number of triples per sentence and the number of tokens per triple, respectively. We **highlight** the highest within-group scores.

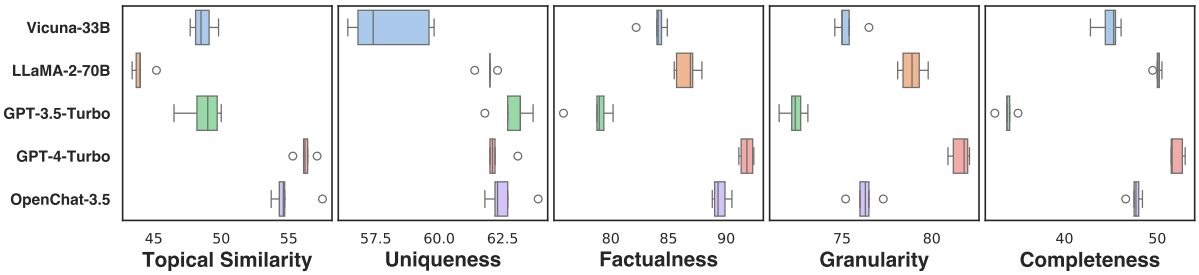

Figure 4: **GRE performance of five LLMs on Wiki20m, each with five runs with random seeds.**

4.4 Open GRE Performance of LLMs

Due to the aforementioned advantages of Open GRE, we further test the capabilities of the leading LLMs to perform this task, which includes **LLaMA Family** (Touvron et al., 2023a,b): LLaMA-2-7B, LLaMA-2-70B, Vicuna-1.5-7B, Vicuna-1.3-33B, and WizardLM-70B (Xu et al., 2023). **GPT Family** (Brown et al., 2020): text-davinci-003, GPT-3.5-Turbo (1106), GPT-3.5-Turbo-Instruct, GPT-4, and GPT-4-Turbo (OpenAI, 2023). **Others**: Mistral-7B-Instruct (Jiang et al., 2023a), Zephyr-7B-Beta (Tunstall et al., 2023), GALACTICA (Taylor et al., 2022), and OpenChat-3.5 (Wang et al., 2023). Models are selected majorly based on their performance on Chatbot Arena (Zheng et al., 2023). Our evaluation results are shown in Tables 2, 3, and 4.

We summarize our findings as follows.

(1) Within individual datasets, LLaMA-2-70B, GPT-4-Turbo, and OpenChat emerge as the top performers in their respective categories based on the highest scores obtained across six datasets. Inter-dataset comparisons reveal that the GPT family consistently outperforms others in Topical Similarity (TS), likely due to their supreme capability to interpret the full content of the text unit. Surprisingly, a light model - OpenChat-3.5 (7B) ourperforms heavier LLMs like Galactica-30B, Vicuna-33B, LLaMA-2-70B, WizardLM-70B, text-davinci-003, and GPT-3.5-Turbo on most datasets.

(2) High Completeness Score (CS) can indicate high Factualness Score (FS). This means human annotations are still valuable to evaluate GRE with our soft matching recall. However, high FS does not indicate high CS, as Open GRE is not limited to the fixed relation/entity types. We also observe that the factualness of GPT-4 and GPT-4-Turbo are consistently higher than that of ground truth.

(3) A greater number of tokens per triple does not inherently result in a lower Granularity Score (GS). This suggests that the GS metric can encourage models to identify more atomic relationships rather than merely focusing on brevity.

(4) We observed no clear correlation between the number of triples, Topical Similarity (TS), and Uniqueness Similarity (US), indicating the distinct significance of each metric. For instance, on the

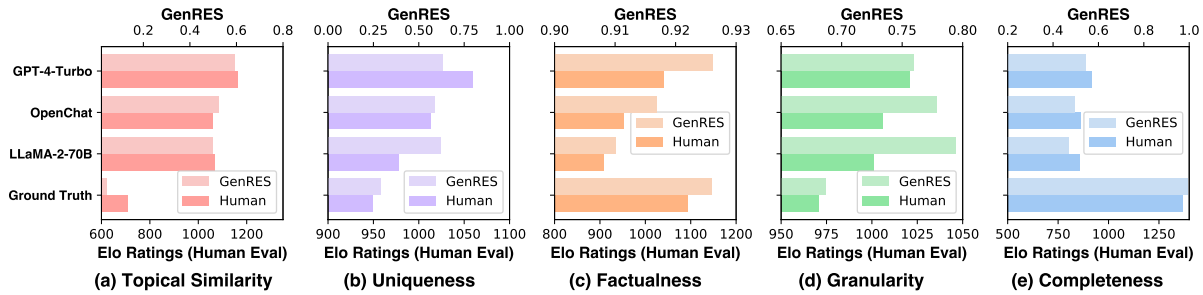

Figure 5: **Human Preference Evaluation (Elo Ratings) vs GenRES Evaluation on 100 Wiki20m samples.**

CDR dataset, Mistral-7B-Instruct and Zephyr-7B-Beta show that a larger output of triples does not necessarily equate to higher TS or lower US. While Zephyr-7B-Beta produces more off-topic triples than Mistral-7B-Instruct, it does not result in more repetitive content. This highlights the importance of evaluating each metric independently.

Figure 4 shows the GRE task performance of five leading LLMs tested with five random seeds on the Wiki20m dataset. The results demonstrate the models' high-quality generation and the effectiveness of our multi-dimensional evaluation framework. Notably, the models' consistent performance across different runs validates our nuanced evaluation metrics, highlighting their robustness in assessing GRE model performance.

Figure 5 showcases the Elo Rating (Elo and Sloan, 1978) results of 100 samples from Wiki20m dataset via human annotation and our proposed GENRES. In most cases, the model ranks by GENRES are consistent with human annotators. We also evaluate the consistency between human annotators using the tie-discounted accuracy (Gao et al., 2023a). We find the following agreement scores: Topical Similarity 81.0%, Uniqueness 93.0%, Factualness 82.7%, Granularity 92.7 %, and Completeness 88.2%. These results showcase the consistency between the human annotators. More details of human evaluation can be found in Appendix D.

5 Related Works

Generative RE. Generative models have exhibited significant promise in the field of RE (Wadhwa et al., 2023b; Wan et al., 2023; Li et al., 2023a). Sequence-to-sequence models such as BART (Lewis et al., 2020) were utilized to extract triples from input texts (Ni et al., 2022; Paolini et al., 2021; Cabot and Navigli, 2021). Then, LLMs were proved to be able to make zero-shot and few-shot generative RE without fine-tuning (Wadhwa et al., 2023b; Li et al., 2023a). Specifically, Wadhwa et al. (2023b) compared GPT-3 (Brown et al., 2020) and FLAN-T5 (Chung et al., 2022) to fully supervised RE methods and identified LLMs reach comparable performance in the zero-shot setup. However, existing GRE methods still rely on a predefined set of relations and entities similar to traditional RE. In this paper, we explore a more open setting and propose a unified evaluation framework GENRES applicable to all types of generative RE.

Evaluation for Text Generation. The evaluation of text generation quality is central to benchmarking the performance of LLMs. While traditional metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) assess surface-level word matching, they often inadequately capture the quality of the generated text. BERTScore (Zhang et al., 2019) focuses on semantic similarity, but still missing the multifaceted nature of text generation. Recently, LLMs have been utilized to evaluate text generation quality, such as FActScore (Min et al., 2023) on verifying the factualness, and UniEval (Zhong et al., 2022) on multi-aspect evaluation. In addition, GPTScore (Fu et al., 2023) utilizes LLMs for token-level probability analysis, enhancing flexibility in text assessment. Recent studies (Liu et al., 2023; Gao et al., 2023b; Li et al., 2023b) explore prompting-based multi-aspect evaluation, broadening the scope of evaluation methods. Unlike all the above works, our GENRES is the first metric designed specifically for Generative RE tasks.

6 Conclusions

In this paper, we introduced GENRES, a framework for evaluating Generative Relation Extraction using Large Language Models, marking a significant shift in the NLP field. Our findings based on extensive tests highlight the potential of LLMs to transform relation extraction and set the stage for future research, potentially revolutionizing information extraction processes and applications across various domains.

References

Elena Baralis, Luca Cagliero, Naeem Mahoto, and Alessandro Fiori. 2013. Graphsum: Discovering correlations among multiple terms for graph-based summarization. *Information Sciences*, 249:96–109.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2370–2381. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Arpad E Elo and Sam Sloan. 1978. The rating of chessplayers: Past and present. *(No Title)*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Kaiyuan Gao, Sunan He, Zhenyu He, Jiacheng Lin, QiZhi Pei, Jie Shao, and Wei Zhang. 2023a. Examining user-friendly and open-sourced large gpt models: A survey on language, multimodal, and scientific gpt models. *arXiv preprint arXiv:2308.14149*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023b. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 745–758, Suzhou, China. Association for Computational Linguistics.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. 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): System Demonstrations*, pages 169–174, Hong Kong, China. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*.

Shibo Hao, Bowen Tan, Kaiwen Tang, Bin Ni, Xiyan Shao, Hengzhe Zhang, Eric Xing, and Zhiting Hu. 2023. BertNet: Harvesting knowledge graphs with arbitrary relations from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5000–5015, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *CoRR*, abs/2310.06825.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 402–410, Online. Association for Computational Linguistics.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2023b. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy,

Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023a. Revisiting large language models as zero-shot relation extractors. *arXiv preprint arXiv:2310.05028*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Ruosen Li, Teerth Patel, and Xinya Du. 2023b. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yike Liu, Tara Safavi, Abhilash Dighe, and Danai Koutra. 2018. Graph summarization methods and applications: A survey. *ACM Comput. Surv.*, 51(3).

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*.

Sameh K Mohamed, Vít Nováček, and Aayah Nounu. 2020. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*, 36(2):603–610.

Jian Ni, Gaetano Rossiello, Alfio Gliozzo, and Radu Florian. 2022. A generative model for relation extraction and classification. *CoRR*, abs/2202.13229.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3):489–508.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023a. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023b. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15566–15589. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-RE: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data.

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1835–1844, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Xiangxiang Zeng, Xinqi Tu, Yuansheng Liu, Xiangzheng Fu, and Yansen Su. 2022. Toward better drug discovery with knowledge graph. *Current opinion in structural biology*, 72:114–126.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: less is more for alignment. *CoRR*, abs/2305.11206.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities.

A Limitations, Ethics, and Risks

A.1 Limitations

LLMs as Evaluators. Within GENRES, we employ the GPT-3.5-Turbo-Instruct large language model (LLM) for assessing the factualness and granularity of extracted relationship triples. However, challenges arise when the LLM delivers incorrect evaluations, particularly in instances where information is overly implicit, misleading, debatable (Chen et al., 2019), or when the model encounters its inherent hallucination issues (Zhang et al., 2023). To mitigate these problems, potential solutions include instructing the model to detail its reasoning process leading to a prediction (Wei et al., 2022), or applying ensemble methods (Li et al., 2023a) to determine the most likely answer. These approaches are areas of interest for our future research endeavors.

Unfocused Extraction by Open GRE. Our research champions the Open Generative Relation Extraction (Open GRE) paradigm, which motivates LLMs to harvest a broader array of relationships, unconstrained by specific relation or entity types. While this approach has demonstrated enhanced topical breadth and factual content in extractions, it also results in a less focused extraction process compared to traditional methods like closed GRE and semi-open GRE (Wadhwa et al., 2023b; Li et al., 2023a). For instance, in constructing a Knowledge Graph (KG) for medical question answering, certain extractions, such as the triple ⟨John, age, 16⟩, might be irrelevant and hence undesirable for inclusion in the KG. However, we posit that an intermediary layer, such as post-processing, should exist between Relation Extraction (RE) and downstream applications. This step would serve to refine and tailor the extracted relationships to meet specific requirements, aligning with methodologies proposed in existing literature (Paulheim, 2017; Liu et al., 2018). Moreover, our GENRES framework is versatile enough to assess all forms of GRE, with the Open GRE configuration, noted for its flexibility, serving as a particularly effective benchmark for evaluating the robustness of our approach.

A.2 Ethics and Risks

All datasets used in this study, namely CDR (Li et al., 2016), DocRED (Yao et al., 2019), NYT10m (Han et al., 2019), Wiki20m (Han et al., 2019), TACRED (Zhang et al., 2017), and Wiki80 (Han et al., 2018) are publicly available. This transparency minimizes ethical concerns related to data sourcing and usage.

Additionally, the interpretability and transparency of LLM decision-making processes are paramount, particularly in contexts involving sensitive or personal data. Recognizing the limitations and error tendencies of LLMs, including occasional information inaccuracies, we emphasize the importance of reliability in our evaluation methods. Furthermore, the integration of LLMs as evaluators impacts traditional human roles, calling for a careful examination of the ethical implications of labor displacement. Lastly, the potent capabilities of LLMs underscore the need for responsible use and measures to prevent misuse, aligning our research with high ethical standards and societal well-being. We carefully checked and ensured that there is no offensive information contained in the data we used as the input to any LLMs.

B Templates for Prompting LLMs

B.1 Templates for Generative Relation Extraction

This appendix delineates the structured prompts and demonstrations utilized in our generative relation extraction methodology. The templates are devised to prime the model for precise and contextually relevant relationship extraction from textual data across different domains and levels of granularity.

General Instruction: The model is instructed to identify relationships between entities, with the aim to extract both intra-sentence and inter-sentence relational triples. This ensures a comprehensive understanding of the text, reflecting the intricacies of document-level nuances and the succinctness of sentence-level information.

LLaMA-2 Model Instruction: An additional directive is provided to the LLaMA-2 model to maintain output stability. The goal is to have the model generate a consistent list of triples, avoiding any extraneous information that does not contribute to the relationship representation.

Demonstration Examples: Examples are tailored to the general and biomedical domains to pre-heat the model towards the target topics. This stratagem is intended to:

- Facilitate the model's adaptation to the domain-specific language and context, thus enabling more accurate and relevant extractions.

Hyperparameter	Values
LDA latent topics	
CDR	{20, 30, 40, **50**, 60, 70, 80, 90, 100}
DocRED	{30, 50, 70, **100**, 150}
NYT10m	{50, 100, **150**, 200, 250}
Wiki20m	{50, 100, **150**, 200, 250}
TACRED	{100, **150**, 200, 250, 300}
Wiki80	{100, **150**, 200, 250, 300}
Triple similarity threshold ϕ	{0.85, 0.90, 0.91, 0.92, 0.93, 0.94, **0.95**, 0.96, 0.97, 0.98}
Open-source LLMs-related	
max_new_tokens	min[#token_limit, {3, 5, 6, 7, **8**, 9, 10}*#input_tokens]
floating-point number	**float16**
GPT-related	
max_new_tokens	**800**
temperature	**0.3**

Table 5: **Hyperparameters Tuning.** We highlight the optimal ones based on our experiments in **bold**.

- Encourage the model to discern and replicate the desired output structure from the examples, which is crucial for reliable relationship extraction.

The provided demonstrations span a variety of contexts and exemplify the format in which the relationships should be presented. The clear and topic-oriented examples aim to fine-tune the model's performance, ensuring it can navigate the complexities of relation extraction with precision across both biomedical and general domains.

B.2 Template for Factualness Verification

In the context of evaluating the factual accuracy of information extracted by language models, we present our template for factualness verification in Figure 8. Utilizing GPT-3.5-Turbo-Instruct as the language model evaluator, our template is designed to solicit a binary output: "true" if the relationship (triplet) is factually correct, "false" otherwise, based solely on the information entailed in the source text.

The template is constructed with three examples, each serving a specific purpose to calibrate the model's understanding of factual correspondence:

- **Example 1** establishes the model's ability to recognize direct factual statements that are explicitly stated in the source text.
- **Example 2** tests the model's discernment of geographical facts and common knowledge, challenging it to detect misinformation.
- **Example 3** assesses the model's capacity to correctly interpret narrative contexts and character relationships, a more subtle and complex form of factual verification.

The inclusion of these examples in the template aims to ensure that the model is thoroughly vetted across a spectrum of factual verification scenarios ranging from straightforward fact-checking to the interpretation of literary works.

B.3 Template for Granularity Checking

For granularity checking, we employ the template shown in Figure 9. The template contains 9 examples, to teach the LLM (GPT-3.5-Turbo-Instruct) what triples can be further split and what are not. Explanations are required when a triple cannot be split (GS = 0).

C Hyper-parameter Tuning

We detail our hyper-paramter tuning in Table 5.

D Human Evaluation

We further conducted human evaluation experiments to verify the alignment of our proposed GenRES with human preferences. Three annotators, who are all computer science graduate students, are involved in this evaluation.

D.1 Evaluation Setup

Our setup for human evaluation follows the approach detailed in studies such as Gao et al. (2023a), Zhou et al. (2023), and Dettmers et al. (2023). We adopt a pairwise comparison method for assessing model outputs. This approach simplifies the evaluation process by requiring human annotators to choose the better result from a pair of options. The evaluation was performed using 100 samples from the Wiki20m dataset. In this process, for each score proposed in Section 3, three human annotators compared the output relationships from

Groundtruth, LLaMA-2-70b, OpenChat, and GPT-4-Turbo in pairs, leading to three possible outcomes for each pair: model A being superior, model B being superior, or a tie. Subsequently, we apply the Elo rating (Elo and Sloan, 1978) system to score the final results.

Elo Rating Elo rating, initially established as a prevalent system for assessing player skill in chess and various competitive games, has recently been adapted to evaluate LLMs[4] (Gao et al., 2023a; Zhou et al., 2023; Dettmers et al., 2023). Its adaptability, characterized by features such as scalability and incremental adjustment, makes it particularly suitable for this purpose. This innovative use of the Elo rating system offers a robust quantitative framework for comparing the performance of various LLMs. In our pairwise comparison setup, the outcome of each comparison impacts the models' scores: a tie results in no change in scores, while a victory leads to an increase in the winner's score and a decrease in the loser's score. Following the completion of all comparisons, the Elo Rating system outputs a final score for each model, thereby establishing their relative rankings based on performance.

Instructions for Annotators The instructions for annotators are shown in Figure 6. Annotators should evaluate the outputs from five aspects in Section 3. During the evaluation process, the models are anonymous for annotators. It should be noted that Completeness is measured after all other metrics have been assessed to prevent the leakage of ground truth information to annotators.

Inter-Annotator Agreement To evaluate Inter-Annotator Agreement with tie-discounted accuracy, we randomly select 50 samples from the 100 Wiki20m samples, resulting in a total of 1500 overlap pairs for two human annotators. This process aimed to assess the consistency level between annotators, anticipating a significant alignment in their evaluations. For the final scoring, we merged all the annotations. The scoring protocol for merging is as follows: (1) When both annotators' responses were in agreement, this consensus was accepted as the merged result. (2) If one annotator declared a tie, the decision of the other was taken as the final annotation. (3) If one annotator believed that 'model A wins' and the other that 'model B wins,' the models were considered tied.

[4] https://lmsys.org/blog/2023-05-03-arena/

Welcome!
As an evaluator, your expertise is pivotal in analyzing how language models interpret and extract information from source texts. This task, termed "extracting relationships," requires you to identify the connections between entities presented as a list of "triples". Each triple consists of [ENTITY1, RELATION, ENTITY2] and represents the link between two entities, collectively forming what we refer to as the "triple list".

Your critical analysis of five key aspects: Topical Similarity, Uniqueness, Factualness, Granularity, and Completeness.

General Instructions
Read Thoroughly: Begin by comprehensively understanding the paragraph to grasp the entities and their interrelations.

Assess Independently: Consider each pair of model extractions independently for each aspect. Avoid allowing judgments in one area to affect another.

Decision Making: For each aspect, determine which model (A or B) better identifies and presents the relationships, or if both are equivalent (tie).

Objective Analysis: Base your evaluations on the outlined criteria, rather than personal opinions or external information.

Aspect-Specific Guidelines
Topical Similarity
Compare the information coverage of the extraction against the source text.
High score: The extraction closely aligns with the main topics and information in the paragraph.
Low score: The extraction deviates from the key topics or includes irrelevant details.

Uniqueness
Examine the information redundancy within the extraction.
High score: The extraction provides unique, diverse perspectives or information.
Low score: The extraction repeats common ideas or lacks originality.

Factualness
Cross-reference the extraction with the source text.
High score: The extraction is factually consistent with the paragraph, with no incorrect or misleading information.
Low score: The extraction contains inaccuracies or fabrications not supported by the paragraph.

Granularity
Evaluate the detail level in the extraction versus the source text.
High score: The extraction offers detailed, specific insights, breaking down meaningful concepts effectively.
Low score: The extraction is overly broad, lacking in specific details or explanations.

Completeness
Compare the extraction relative to the "gold standard" triple list.
High score: The extracted list contains triples that are similar to the gold standard, acknowledging that similar triples convey the same information.
Low score: The extracted list omits a significant number of triples found in the gold standard or has very few similarities.

Final Remarks
Your assessments are integral to our understanding and enhancement of language model capabilities. Please dedicate the necessary time for thoughtful and precise evaluations based on the provided criteria. Your objective and detailed feedback is invaluable to the advancement of language model technology.

We are grateful for your thoroughness and the attention to detail you bring to this task.

Figure 6: **Instruction for Human Annotators.**

Common Instruction

```
Given a text, identify and list the relationships between entities within the text. (doc/bag-level)
Extract relationships both within a single sentence (intra-sentence) and across multiple sentences
(inter-sentence).
Provide a list of triplets in the format [`ENTITY 1`, `RELATIONSHIP`, `ENTITY 2`].
```

LLAMA-2-Specific Instruction

```
The output should only be a list of triplets ([[`ENTITY 1`, `RELATIONSHIP`, `ENTITY 2`], ...]) without
any additional information. Do not explain how you extract them.
```

Demonstration Doc/Bag-Level, General Domain

```
Example 1:
text: In 2020, the Nobel Peace Prize was awarded to the World Food Programme for its efforts to combat
hunger. The organization has been operational since 1961.
relations:
[["Nobel Peace Prize", "awarded in", "2020"], ["Nobel Peace Prize", "awarded to", "World Food
Programme"], ["World Food Programme", "efforts to", "combat hunger"], ["World Food Programme",
"operational since", "1961"]]

Example 2:
text: The Great Barrier Reef, located off the coast of Australia, is the world's largest coral reef
system. It has been severely affected by climate change, leading to coral bleaching.
relations:
[["Great Barrier Reef", "located at", "coast of Australia"], ["Great Barrier Reef", "is", "world's
largest coral reef system"], ["Great Barrier Reef", "affected by", "climate change"], ["Climate
change", "leads to", "coral bleaching"]]
```

or

Demonstration Sentence-Level, General Domain

```
Example:
text: The Great Barrier Reef has been severely affected by climate change, leading to coral
bleaching.
relations:
[["Great Barrier Reef", "affected by", "climate change"], ["Climate change", "leads to", "coral
bleaching"]]
```

or

Demonstration Doc/Bag – Level, Bio Domain

```
Example 1:
text: Penicillin is an antibiotic that treats bacterial infections. It was discovered by Alexander
Fleming.
relations:
[["Penicillin", "is a type of", "antibiotic"], ["Penicillin", "treats", "bacterial infections"],
["Penicillin", "discovered by", "Alexander Fleming"]]

Example 2:
text: Metformin is commonly prescribed for managing type 2 diabetes. It helps by lowering glucose
production in the liver and increasing the body's sensitivity to insulin.
relations:
[["Metformin", "is prescribed for", "managing type 2 diabetes"], ["Metformin", "helps by", "lowering
glucose production in the liver"], ["Metformin", "increases", "body\'s sensitivity to insulin"]]
```

or

Demonstration Sentence-Level, Bio Domain

```
Example:
text: Penicillin is an antibiotic that treats bacterial infections.
relations:
[["Penicillin", "is a type of", "antibiotic"], ["Penicillin", "treats", "bacterial infections"]]
```

Common Instruction

```
text: $INPUT_TEXT$
relations:
```

Figure 7: **Templates used for Open Generative Relation Extraction.**

```
Evaluate the factualness of an extracted relationship (triplet) based on the given
source text. Indicate whether the relationship accurately reflects the information in
the source text by responding with "true" or "false".
You should only output "true" or "false" with no additional information.

Example 1:
Source Text: The Great Barrier Reef, located off the coast of Australia, is the
world's largest coral reef system. It has been severely affected by climate change,
leading to coral bleaching.
Relationship: ["Great Barrier Reef", "affected by", "climate change"]
Factualness: true

Example 2:
Source Text: The Eiffel Tower was constructed in 1889 and is located in Paris, France.
It is one of the most recognizable structures in the world.
Relationship: ["Eiffel Tower", "located in", "London"]
Factualness: false

Example 3:
Source Text: The novel "Moby-Dick" by Herman Melville features a ship named Pequod.
The narrative follows the ship and its crew in their pursuit of a giant white sperm
whale.
Relationship: ["Moby-Dick", "is about", "a whale named Pequod"]
Factualness: false


Source Text: $TEXT$
Relationship: $TRIPLE$
Factualness:
```

Figure 8: **Template for Factualness Verification.**

```
Evaluate the given triple for its potential to be split into more specific sub-triples. Provide the
sub-triples in the format [e, r, o] and give the total count. If no split is necessary, explain
briefly.

Example 1:
Triple: ["text messaging", "has popularized", "the use of abbreviations"]
Sub-triples: N/A (The triple is already specific and cannot be broken down further.)
Granularity: 0

Example 2:
Triple: ["electric cars", "offer benefits like", "energy efficiency and environmental friendliness"]
Sub-triples:
["electric cars", "offer benefits like", "energy efficiency"]
["electric cars", "offer benefits like", "environmental friendliness"]
Granularity: 2

Example 3:
Triple: ["exercise", "boosts", "health"]
Sub-triples: N/A (The relationship is direct and does not need further granularity.)
Granularity: 0

Example 4:
Triple: ["trees", "provide", "oxygen, shade, and habitats"]
Sub-triples:
["trees", "provide", "oxygen"]
["trees", "provide", "shade"]
["trees", "provide", "habitats"]
Granularity: 3

Example 5:
Triple: ["healthy diet", "contributes to", "wellness"]
Sub-triples: N/A (The term 'wellness' encompasses a broad range of aspects, which are implicitly
understood.)
Granularity: 0

Example 6:
Triple: ["water", "exists as", "solid, liquid, gas"]
Sub-triples:
["water", "exists as", "solid"]
["water", "exists as", "liquid"]
["water", "exists as", "gas"]
Granularity: 3

Example 7:
Triple: ["urbanization", "leads to", "various social and environmental changes"]
Sub-triples:
["urbanization", "leads to", "social changes"]
["urbanization", "leads to", "environmental changes"]
Granularity: 2

Example 8:
Triple: ["global warming", "causes", "climate change and associated phenomena like sea-level rise"]
Sub-triples:
["global warming", "causes", "climate change"]
["global warming", "causes", "sea-level rise"]
Granularity: 2

Example 9:
Triple: ["antibiotics", "treat", "bacterial infections"]
Sub-triples: N/A (The triple is specific, conveying a singular relation between antibiotics and
bacterial infections.)
Granularity: 0

Prompt:
Triple: $TRIPLE$
Sub-triples:
```

Figure 9: **Template for Granularity Checking.**