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

Anonymous ACL submission

Abstract

In this study, we present GENRES, a novel evaluation framework specifically designed for assessing Large Language Models (LLMs) in Generative Relation Extraction (GRE). While GRE represents an advanced setting where LLMs are utilized to extract a broader spectrum of relationships from text, our primary contribution lies in the development of GENRES. This suite of metrics addresses the critical need for a more nuanced evaluation of LLMs in GRE, focusing on factualness, topical similarity, uniqueness, granularity, and completeness of the knowledge extracted. Our empirical analysis across various datasets highlights the effectiveness of GENRES in capturing the intricacies of relational data generated by LLMs. This framework not only provides a comprehensive tool for evaluating LLM performance in GRE but also marks a significant advancement in relation extraction research, underscoring the transformative potential of LLMs in deriving deep insights from text.

1 Introduction

The digital era is characterized by an unprecedented surge in textual data, presenting unique challenges and opportunities for Natural Language Processing (NLP). Central to NLP's utility is Relation Extraction (RE), a technique vital for transforming unstructured text into structured, actionable knowledge. While traditional RE approaches have relied on predefined patterns and statistical models, effective in certain situations, they often lack the flexibility to fully capture the complexity of natural language. The advent of Generative Relation Extraction (GRE), powered by Large Language Models (LLMs) like the GPT series, signifies a notable evolution in NLP. These models, capable of intuitively understanding and generating text, have shifted the paradigm of RE by identifying complex relationships without the constraints of predefined patterns or extensive training datasets. However, existing applications of LLMs in GRE are either essentially performing simple binary classification tasks (Li et al., 2023a) given entity pairs and a set of predefined relation types, or relatively open to the entity extraction (Wadhwa et al., 2023a), overlooking extensive novel and potentially important relationships beneath the text. We advocate a transformative approach in RE with LLMs: moving from **"defining a set of relation types"** → **"finding matches between entities"** to **"exploring as**

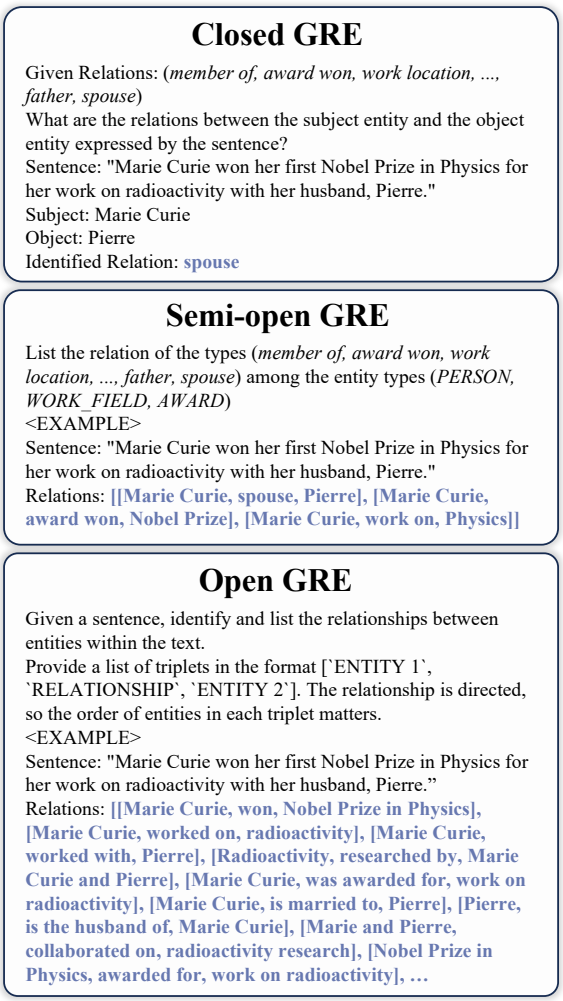

Figure 1: **Closed, Semi-open, and Open GRE.**

many relations and entities as possible without limitation" → **"gathering and sorting them (e.g., clustering)"**. This strategy leverages LLMs' capabilities to their fullest, allowing for the discovery of a wider array of relationships, which we define as "Open GRE". In Figure 1, we showcase three GRE strategies on the same text.

The versatility of Open GRE, however, poses significant challenges in evaluation (Wadhwa et al., 2023a). Traditional metrics fall short as they rely on comparing outputs with "ground truth" triples, confined to a limited set of relationships. To address this, we propose a reimagined evaluation framework. 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 comprehensive framework tailored for evaluating GRE. GENRES assess LLMs' performance through various lenses (1) *Topical Similarity*: Assessing topical similarity between extracted knowledge and the source text. (2) *Uniqueness*: Evaluating diversity and novelty in the extracted knowledge. (3) *Factualness*: Gauging the accuracy of the extracted knowledge referring to the source text. (4) *Granularity*: Examining the specificity and level of detail in the extracted knowledge. (5) *Completeness*: Optionally, measuring the soft matching recall of the extracted knowledge to the "gold standard".

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 (semi-) open GRE methods as the LLM generations are flexible. To fill in this gap, we introduce GENRES, a series of automated multi-aspect evaluation metrics for generative relation extraction. GENRES are composed of the following sub-scores: (1) Distribution Similarity Score, (2) Uniqueness Score, (3) Factualness Score, (4) Granularity Score, and (5) Completeness Score,. Each of these scores will be elaborated upon in the subsequent subsections.

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 $\mathcal{T}_\mathcal{D}^\Delta = [\tau_1', \tau_2', ...] = [s_1 \oplus r_1 \oplus o_1, s_2 \oplus r_2 \oplus o_2, ...]$. TS is computed as:

$$t(\mathcal{D}, \mathcal{T}_\mathcal{D}^\Delta) = e^{-\sum_{i=1}^{K} LDA(\mathcal{D})_i \cdot \log\left(\frac{LDA(\mathcal{D})_i}{LDA(\mathcal{T}_\mathcal{D}^\Delta)_i}\right)} \quad (1)$$

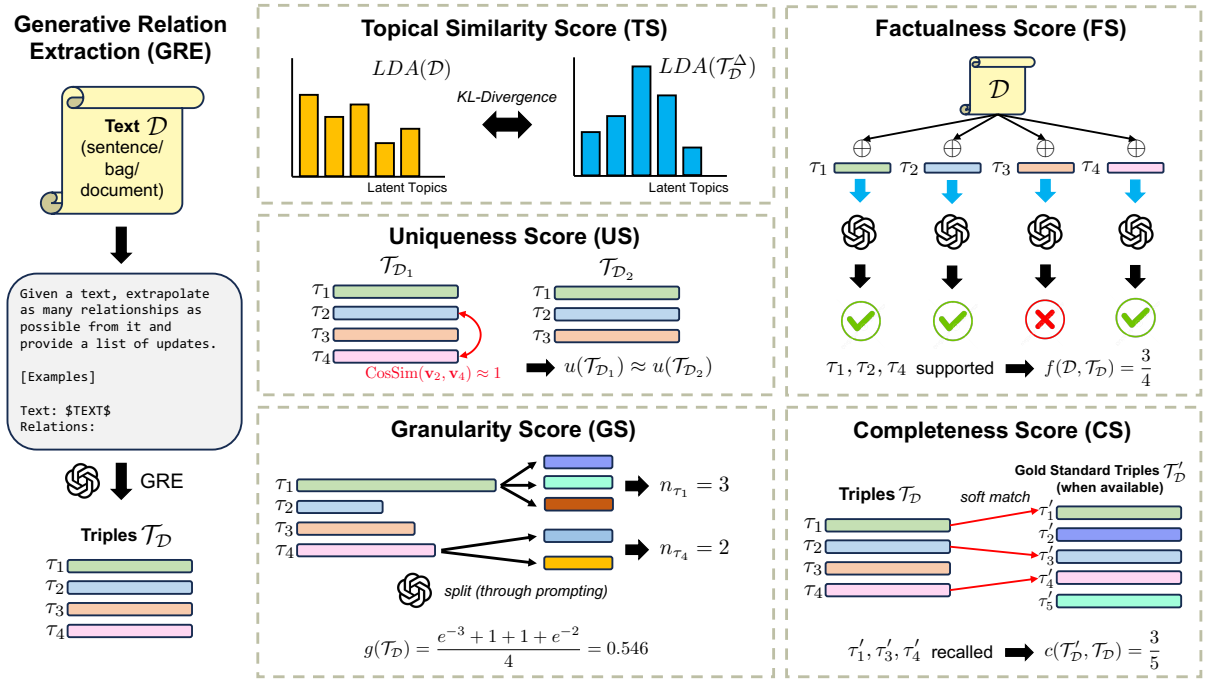

Figure 2: **GENRES for the evaluation of generative relation extraction (GRE).** ***Left***: An example showing the GRE process to extract triples $\mathcal{T}_\mathcal{D}$ from a source text $\mathcal{D}$ through prompting generative large language model. ***Right***: illustration of sub-scores contained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

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 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

The 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:

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

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

The 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 addressing the issue of hallucinations (Zhang et al., 2023), a phenomenon where Language Model systems (LLMs) produce content not present or suggested in the source text. Building on the foundations laid by prior research (Min et al., 2023; Jiang et al., 2021), the 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 $\mathcal{D}$:

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

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 A.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:

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

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:

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

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

We evaluate the following RE datasets, focusing on GRE performance using their test sets, which feature meticulous human annotations.

CDR (Li et al., 2016). A *document-level* RE dataset comprising 1,500 PubMed abstracts. The dataset is divided evenly for training, development, and testing. Each abstract has been meticulously annotated to mark the binary interactions between chemical compounds and disease entities.

DocRED (Yao et al., 2019). A *document-level* RE dataset derived from Wikipedia and Wikidata, featuring 5,053 Wikipedia documents with 132,375 entities and 56,354 relational facts. It includes human annotations for entity mentions, coreferences, and intra- and inter-sentence relations, along with supporting evidence.

NYT10m & **Wiki20m** (Han et al., 2019). Two *bag-level*[1] RE datasets sourced from The New York Times and Wikipedia, respectively. Both datasets have manually annotated test sets.

TACRED (Zhang et al., 2017) & **Wiki80** (Han et al., 2019): Two *sentence-level* RE datasets. TACRED includes 106,264 examples from newswire and web texts, covering 41 relation types, using TAC KBP challenge data and crowdsourcing. Wiki80, sourced from FewRel (Han et al., 2018), contains 80 relations with 56,000 instances from Wikipedia and Wikidata.

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].

[1] A "bag" of information that share the same entity pair. https://opennre-docs.readthedocs.io

[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,

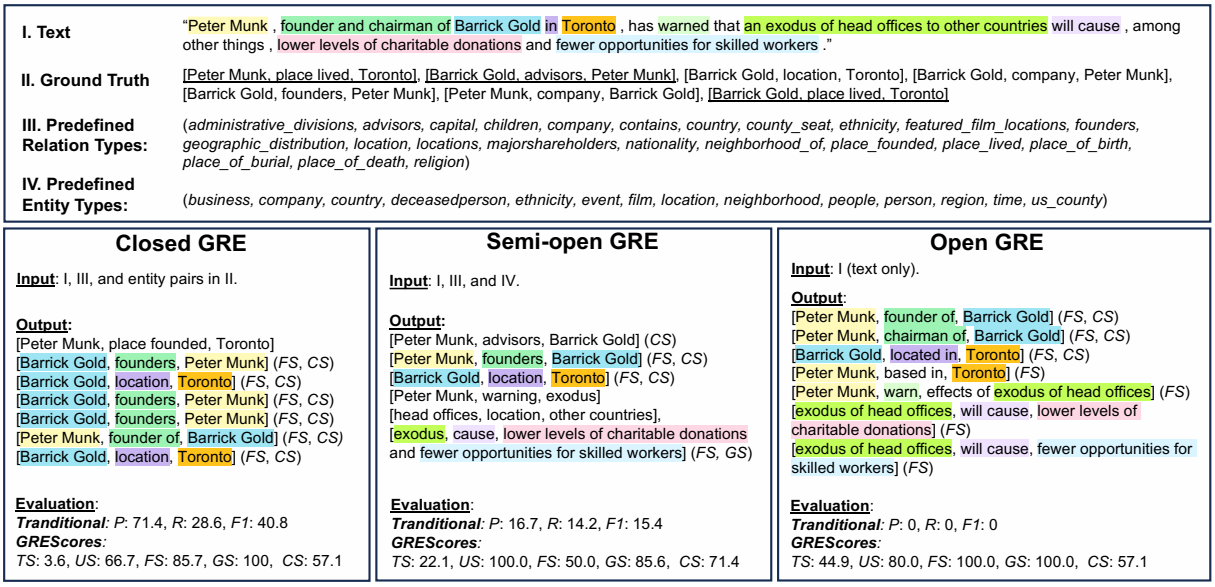

Figure 3: **Comparative Analysis of GRE Methods and Evaluation Metrics using the NYT10m Dataset.** The diagram showcases the outcomes of closed, semi-open, and open Generative Relation Extraction (GRE) strategies. The distinct entity and relation spans are color-coded, with factual triples specifically highlighted. The extracted triples that affect FS (soft precision), CS (soft recall), and GS are listed with the corresponding labels. We further 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	**84.6**	58.5	43.1	88.2	**84.2**	30.8	62.5	85.6
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 GREScores.** "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.

4.2 Implementation Details

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 (hidden dimension: 1536) for each entity and relation in the triple using `text-embedding-ada-002`[3], and perform element-wise addition to obtain the triple embedding. Based on our tests, we set the similarity threshold ϕ at 0.95. All local LLMs are run on 8 NVIDIA A100 GPUs. All templates we used for prompting are detailed in Appendix A.

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.

Precision, Recall, and F1 are not ideal metrics for evaluating GRE methods, particularly for semi-open and open GRE. Figure 3 illustrates that despite open GRE's high-quality extractions based on FS and CS, they score zero across these metrics.

500 samples were randomly selected.

[3] https://openai.com/blog/new-and-improved-embedding-model

		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	88.2	100	12.4	6.0	8.4	64.0	94.4	72.4	100
LLaMA	Vicuna-7B	6.8	8.4	57.8	86.9	84.7	31.8	30.7	7.4	9.9	23.1	81.9	93.4	37.7	28.3
	Vicuna-33B	6.4	10.5	73.0	89.2	**97.3**	30.5	32.0	10.8	9.8	34.7	82.8	97.2	42.0	36.9
	LLaMA-2-7B	5.6	6.7	48.6	92.0	62.0	29.5	25.7	2.7	3.2	12.8	93.3	34.0	20.7	12.1
	LLaMA-2-70B	10.8	8.1	**74.8**	87.6	96.6	**48.9**	**51.0**	13.8	8.7	**39.2**	82.6	**97.3**	**51.8**	**39.2**
	WizardLM-70B	10.2	7.8	65.4	**94.1**	76.4	29.2	32.6	5.8	3.6	24.3	**94.9**	37.9	18.3	12.8
GPT	text-davinci-003	12.7	8.3	76.7	87.2	96.8	**44.1**	44.3	15.3	8.5	40.1	84.2	97.6	49.5	46.2
	GPT-3.5-Turbo-Inst.	16.1	8.3	77.6	89.6	96.8	43.1	47.8	17.8	8.9	47.8	85.6	98.1	46.3	44.7
	GPT-3.5-Turbo	11.2	11.4	81.7	89.2	**98.2**	33.0	30.2	15.0	9.9	**50.4**	84.0	98.5	42.1	36.5
	GPT-4	14.3	9.3	81.7	91.0	97.9	39.6	46.3	17.8	8.7	48.6	82.8	**98.6**	50.5	47.3
	GPT-4-Turbo	18.6	8.5	**82.1**	**91.9**	96.8	43.4	**48.8**	21.5	8.7	50.0	**87.4**	97.6	**52.4**	**49.3**
others	Mistral-7B-Inst.	14.2	9.1	69.0	74.9	93.5	42.0	**40.0**	11.3	9.6	30.2	76.4	94.1	46.0	27.5
	Zephyr-7B-Beta	25.9	8.8	49.1	79.5	70.1	**47.4**	29.3	18.6	8.6	27.9	79.4	94.7	**54.6**	37.1
	Galactica-30B	0.2	0.3	4.1	1.1	0.9	0.8	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**	30.9	31.8	15.4	8.9	**39.7**	**82.1**	**98.1**	51.3	**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	85.6	100	2.0	6.3	4.4	21.2	85.7	66.1	100
LLaMA	Vicuna-7B	3.1	7.8	42.0	86.4	80.0	49.4	38.9	3.0	7.5	48.3	67.8	50.0	55.8	37.3
	Vicuna-33B	4.7	7.2	**47.8**	80.1	75.1	55.2	46.5	4.1	7.0	**49.8**	56.4	84.4	62.7	46.1
	LLaMA-2-7B	3.1	6.0	35.4	82.2	78.9	52.1	38.4	3.1	6.3	37.9	**73.8**	73.4	58.6	36.0
	LLaMA-2-70B	5.0	6.9	45.4	83.0	**81.7**	**63.5**	**52.4**	4.1	6.9	45.2	62.0	**87.1**	**66.1**	**50.2**
	WizardLM-70B	4.4	4.2	30.5	**88.9**	43.9	32.7	27.6	3.6	5.6	43.1	67.8	67.3	47.9	40.9
GPT	text-davinci-003	4.9	7.1	50.6	81.4	85.8	60.0	52.6	3.7	8.2	51.8	56.9	91.3	62.3	43.5
	GPT-3.5-Turbo-Inst.	5.8	7.0	54.2	83.0	84.0	**62.5**	53.4	4.8	7.7	54.0	60.3	90.1	65.1	43.8
	GPT-3.5-Turbo	4.1	6.2	43.3	82.3	68.2	42.4	29.8	3.6	7.7	48.2	61.8	80.2	52.7	32.5
	GPT-4	5.1	7.4	56.2	81.8	89.0	60.9	52.6	3.8	8.1	**59.0**	56.2	**93.2**	66.4	40.0
	GPT-4-Turbo	5.3	7.8	**58.1**	**84.2**	**89.6**	61.1	**53.7**	4.2	7.6	56.4	**62.0**	92.4	**69.9**	**52.7**
others	Mistral-7B-Inst.	5.7	7.4	40.6	77.6	75.4	53.3	36.5	4.0	6.9	43.3	57.0	83.6	58.5	40.1
	Zephyr-7B-Beta	7.8	7.2	36.5	80.8	64.9	**64.5**	47.0	5.2	6.8	40.3	**65.5**	75.5	**67.9**	45.9
	Galactica-30B	8.3	8.7	29.7	48.4	52.4	49.3	37.0	6.0	8.4	35.3	49.4	65.2	57.1	38.6
	OpenChat-3.5	5.2	7.2	**54.0**	**84.7**	**84.3**	61.5	**55.3**	4.3	7.0	**57.5**	61.8	**90.5**	63.6	**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.

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 are sometimes 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 Precision/Recall/F1 metrics which 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 enforce LLMs into generating inaccurate 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

		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	88.5	100	1.0	5.8	5.9	100	90.1	70.3	100
LLaMA	Vicuna-7B	2.6	8.7	40.4	85.0	**75.6**	50.3	36.2	2.4	7.9	41.3	76.8	81.0	51.2	36.6
	Vicuna-33B	4.3	7.3	**44.3**	75.5	71.0	58.5	47.2	3.8	7.2	**47.3**	62.1	79.9	60.2	46.8
	LLaMA-2-7B	2.8	6.3	36.7	85.3	66.9	57.2	37.8	2.4	5.8	25.8	69.8	60.4	53.2	31.4
	LLaMA-2-70B	4.1	6.4	40.8	79.3	74.5	**67.2**	**56.4**	3.7	6.6	41.5	64.8	**82.4**	**65.6**	**49.4**
	WizardLM-70B	2.1	2.9	23.3	**90.7**	28.0	24.7	9.8	2.1	3.2	25.6	**84.9**	36.6	27.3	21.4
GPT	text-davinci-003	4.4	7.1	56.1	79.8	84.0	63.4	58.6	4.0	6.8	59.2	65.3	89.2	64.0	51.9
	GPT-3.5-Turbo-Inst.	5.0	7.0	58.6	80.5	81.6	**63.8**	58.6	4.4	6.9	60.2	69.3	88.7	63.9	**54.8**
	GPT-3.5-Turbo	3.9	6.8	52.7	81.1	76.4	52.1	39.7	3.4	6.3	50.9	**69.5**	75.6	48.1	36.0
	GPT-4	4.3	7.5	**59.1**	80.4	87.6	60.5	57.8	4.0	7.1	**65.4**	66.2	92.3	**64.2**	47.8
	GPT-4-Turbo	4.4	7.8	58.5	**82.6**	**88.6**	61.9	**63.4**	4.0	7.6	61.9	69.4	**92.8**	63.9	47.1
others	Mistral-7B-Inst.	4.7	7.1	43.9	78.6	71.0	53.5	41.2	3.6	7.8	44.6	67.8	83.9	57.6	38.5
	Zephyr-7B-Beta	5.4	7.6	36.4	78.6	65.8	62.9	44.9	4.5	7.8	43.2	68.1	77.8	63.0	42.6
	Galactica-30B	8.5	8.9	33.4	43.9	57.5	54.1	30.9	5.6	7.2	35.0	47.9	63.1	59.8	38.4
	OpenChat-3.5	4.3	7.1	**50.7**	**80.8**	**80.4**	**63.6**	**60.0**	4.0	7.0	**53.8**	**69.7**	**88.7**	**64.5**	**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.

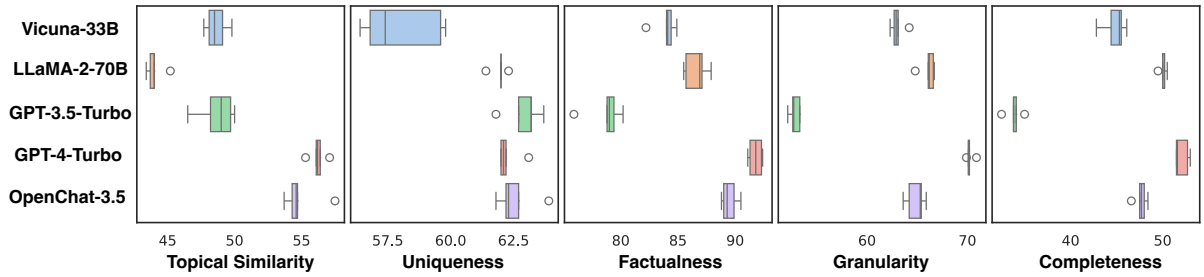

Figure 4: **GRE performance of five LLMs on Wiki20m, each with five runs with random seeds.**

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.

4.4 Open GRE Performance of LLMs

Due to the aforementioned strengths of Open GRE, we further test the capabilities of the leading LLMs (as of December 6, 2023) 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. **Others**: Mistral-7B-Instruct (Jiang et al., 2023), Zephyr-7B-Beta (Tunstall et al., 2023), GALACTICA (Taylor et al., 2022), and OpenChat-3.5 (Wang et al., 2023). We select these models majorly based on their performance on Chatbot Arena (Zheng et al., 2023). Our evaluation results on docoment-level, bag-level, and sentence-level datasets are showcased in Tables 2, 3, and 4, respectively.

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 in-

terpret 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 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 presents the results of our GRE task performance test using five leading LLMs, each with five random seeds on the Wiki20m dataset. The outcomes clearly demonstrate the model's high-quality generation capabilities and the robustness of our comprehensive multi-dimensional evaluation framework. This robustness is particularly noteworthy, given the model's consistent performance across different runs, underscoring the reliability of GPT-4-Turbo in generating high-quality triples. The results further cement the validity of our evaluation metrics, which are designed to capture a nuanced view of GRE model performance.

5 Related Works

Open RE. Open RE uncovers new relation types in unsupervised open-domain corpora, primarily through tagging-based and clustering-based approaches. Tagging-based Open RE treats the task as sequence labeling, extracting relational phrases from sentences (Jia et al., 2019; Cui et al., 2018; Stanovsky et al., 2018), while clustering-based methods utilize external linguistic tools to feature-rich relations and cluster them into distinct types (Zhou et al., 2023; Marcheggiani and Titov, 2016; ElSahar et al., 2017).

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). Before the era of LLMs, researchers such as Ni et al. (2022); Paolini et al. (2021); Cabot and Navigli (2021) utilized sequence-to-sequence models like BART (Lewis et al., 2020) for extracting relation triplets. With the rise of LLMs, Wadhwa et al. (2023b) have demonstrated the remarkable capabilities of LLMs such as GPT-3 (Brown et al., 2020) and FLAN-T5 (Chung et al., 2022) in RE tasks using generative LLMs. GRE employing LLMs consistently outperforms traditional RE methods by a significant margin.

Evaluation for Text Generation. In the era of LLMs, evaluating text generation has become increasingly significant. This evaluation is crucial not only for understanding the capabilities of different LLMs but also for laying the groundwork for future research advancements. Traditional metrics like BERTScore (Zhang et al., 2019) fall short in capturing the multifaceted nature of generated text, as they primarily focus on semantic similarity. FActScore (Min et al., 2023) enhances evaluation by integrating retrieval techniques and LLM-based factual verification. UniEval (Zhong et al., 2022) employs aspect-specific questions answered by a fine-tuned Seq2Seq model, producing binary predictions. GPTScore (Fu et al., 2023) suggests leveraging pre-trained LLMs for token-level probability analysis with flexibly designed questions. Other approaches include prompting LLMs for explicit evaluation answers, as explored in (Liu et al., 2023; Gao et al., 2023; Li et al., 2023b), further diversifying the methods for text generation evaluation.

6 Conclusions

In this paper, we introduced GENRES, some innovative metrics 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

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.

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.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 407–413. Association for Computational Linguistics.

Hady ElSahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frédérique Laforest. 2017. Unsupervised open relation extraction. In *The Semantic Web: ESWC 2017 Satellite Events - ESWC 2017 Satellite Events, Portorož, Slovenia, May 28 - June 1, 2017, Revised Selected Papers*, volume 10577 of *Lecture Notes in Computer Science*, pages 12–16. Springer.

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

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

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*.

Shengbin Jia, Shijia E, and Yang Xiang. 2019. Supervised neural models revitalize the open relation extraction. *CoRR*, abs/1908.01761.

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. 2023. 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.

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*.

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*.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Trans. Assoc. Comput. Linguistics*, 4:231–244.

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*.

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

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.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 885–895. Association for Computational Linguistics.

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.

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*.

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*.

Jie Zhou, Shenpo Dong, Yunxin Huang, Meihan Wu, Haili Li, Jingnan Wang, Hongkui Tu, and Xiaodong Wang. 2023. U-CORE: A Unified Deep Cluster-wise Contrastive Framework for Open Relation Extraction. *Transactions of the Association for Computational Linguistics*, 11:1301–1315.

A Templates for Prompting LLMs

A.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.
- 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.

A.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 6. 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.

A.3 Template for Granularity Checking

For granularity checking, we employ the template shown in Figure 7. 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).

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`]. The relationship
is directed, so the order of entities in each triplet matters.
```

✚

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 5: **Templates used for Open Generative Relation Extraction.**

```
Evaluate the factualness of an extracted relationship (triplet) based on the its
source text. Responding "true" or "false" to indicate whether the relationship can be
extracted from the source text.
You should only output "true" or "false" with no additional information.

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

Figure 6: **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 7: **Template for Granularity Checking.**