NAACL 2024

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

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han
University of Illinois at Urbana-Champaign

Overview

- **Background**: from Zero-shot RE to Generative RE (GRE)
- **Introduction**: Why should we care about GRE's evaluation?
- **Method**: GenRES (Generative Relation Extraction Scoring)
- **Results**:

  (1) Why not traditional metrics but GenRES?

  (2) Our evaluation of the leading LLMs' GRE capabilities

Background: Traditional Relation Extraction

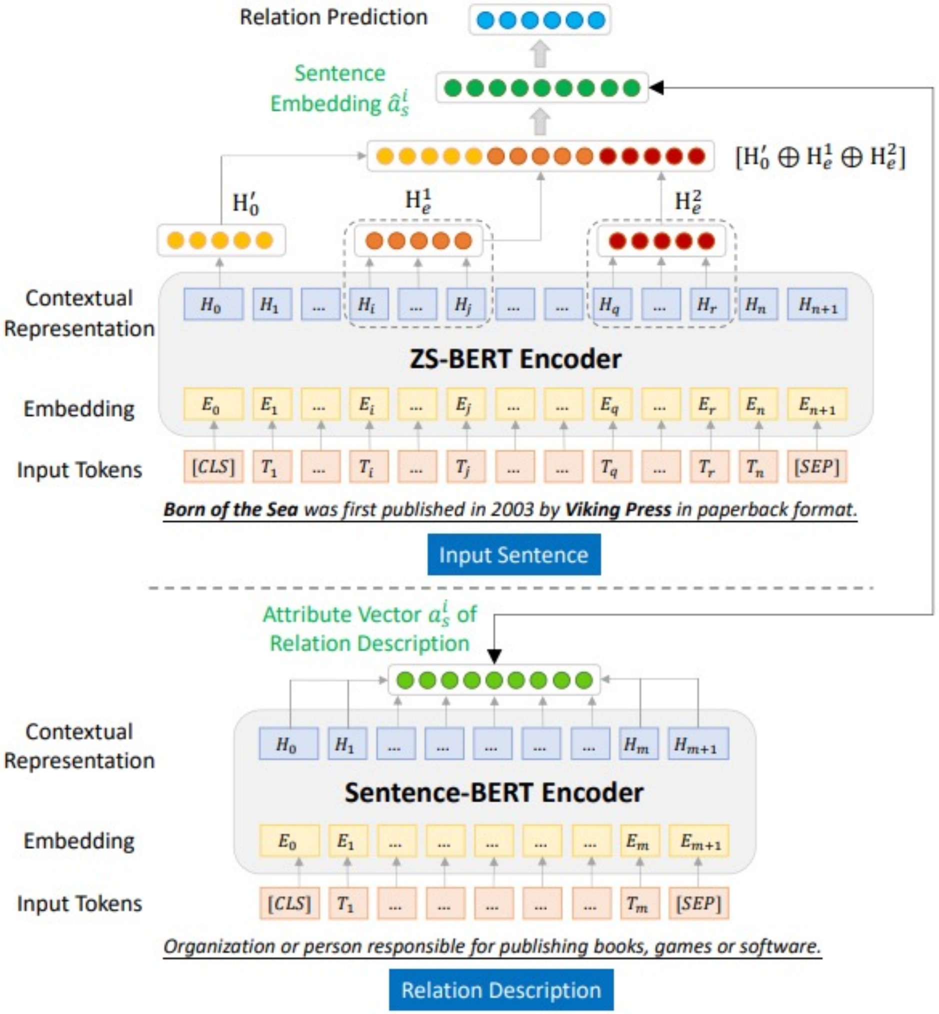
- Relation extraction is a major task in the field of information extraction
- **Task definition 1**: Given a sentence with two annotated entities, classify their relation (or no relation)
- **Task definition 2**: Given a sentence, detect entities and all the relations between them
  - NER is required first
  - Entities can be pronouns, requiring coreference resolution
  - Relations can be pre-defined or discovered

Citing high fuel prices, [$_{\text{ORG}}$ **United Airlines**] said [$_{\text{TIME}}$ **Friday**] it has increased fares by [$_{\text{MONEY}}$ **$6**] per round trip on flights to some cities also served by lower-cost carriers. [$_{\text{ORG}}$ **American Airlines**], a unit of [$_{\text{ORG}}$ **AMR Corp.**], immediately matched the move, spokesman [$_{\text{PER}}$ **Tim Wagner**] said. [$_{\text{ORG}}$ **United**], a unit of [$_{\text{ORG}}$ **UAL Corp.**], said the increase took effect [$_{\text{TIME}}$ **Thursday**] and applies to most routes where it competes against discount carriers, such as [$_{\text{LOC}}$ **Chicago**] to [$_{\text{LOC}}$ **Dallas**] and [$_{\text{LOC}}$ **Denver**] to [$_{\text{LOC}}$ **San Francisco**].

Entity 1	Relation	Entity 2
United	PartOf	UAL Corp.
Tim Wagner	OrgAff	American Airlines
...	...	...

Background: Zero-shot Relation Extraction

ZS-BERT [1]



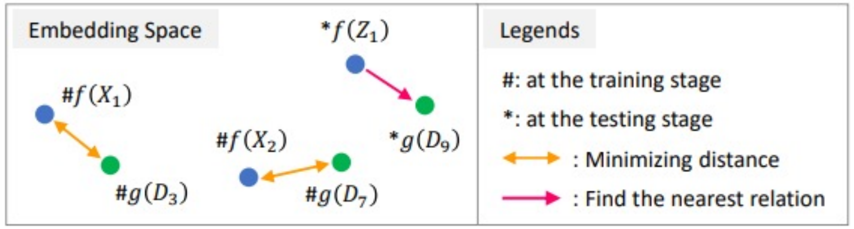
Two training objectives:

(1) Aligning Sentence Embedding and Attribute Vector of Relation Description

(2) Maximize the accuracy of Relation Classification

Zero-shot Prediction:

- Nearest Neighbor Search

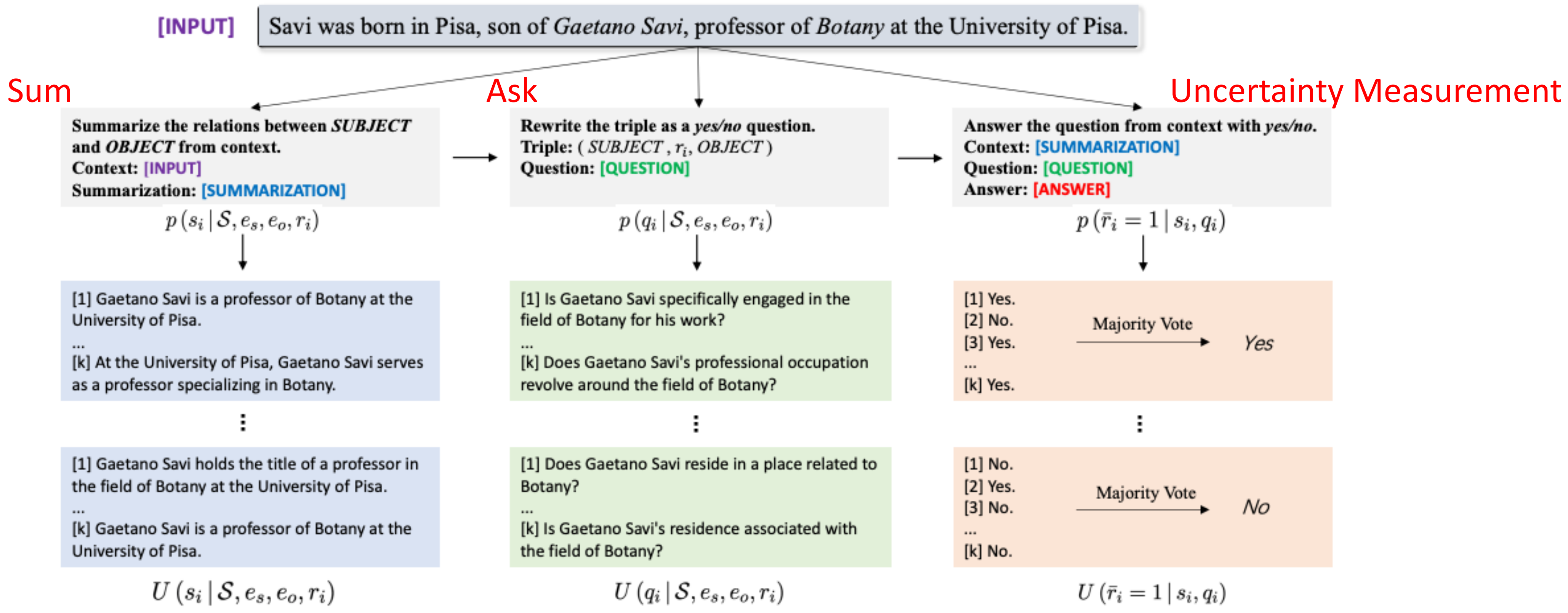


	Input Sentence	Relation	Relation Description
Training	X_1: In 1997, Dennis Crouch and Hester put together a western swing band called "The Time Jumpers"	y_3: **member of** (seen)	D_3: organization, musical group, or club to which the subject belongs
Training	X_2: He had roles in two 2008 films: the sci-fi film "Jumper" and the World War II drama "Defiance"	y_7: **main subject** (seen)	D_7: primary topic of a work
Testing	Z_1: During the Philippine–American War, Mark Twain wrote a short pacifist story titled "The War Prayer"	y_9: **author** (unseen, ground truth)	D_9: main creator(s) of a written work

[1] Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards Zero-Shot Relation Extraction with Attribute Representation Learning. *NAACL 2021*

Background: Generative Relation Extraction

SumAsk [2]



$$
\begin{aligned}
p(\textit{field of work}) &= p(s \mid S, e_s, e_o, r_{field\_of\_work})\, p(q \mid S, e_s, e_o, r_{field\_of\_work})\, p(\bar{r}_{other\_relation} = 0 \mid \bar{r}_{field\_of\_work} = 1)\, p(\bar{r}_{field\_of\_work} = 1 \mid s, q) \\
&= p(s \mid S, e_s, e_o, r_{field\_of\_work})\, p(q \mid S, e_s, e_o, r_{field\_of\_work})\, p(\bar{r}_{field\_of\_work} = 1 \mid s, q) \\
&\ 1/\!\propto U(s \mid S, e_s, e_o, r_{field\_of\_work})\, U(q \mid S, e_s, e_o, r_{field\_of\_work})\, U(\bar{r}_{field\_of\_work} = 1 \mid s, q)
\end{aligned}
$$

$$
p(\textit{residence}) = p(s \mid S, e_s, e_o, r_{residence})\, p(q \mid S, e_s, e_o, r_{residence})\, p(\bar{r}_{other\_relation} = 0 \mid \bar{r}_{residence} = 1)\, p(\bar{r}_{residence} = 1 \mid s, q) = 0
$$

[2] Li et al., Revisiting large language models as zero-shot relation extractors. *EMNLP'23-Findings*

Background: Generative Relation Extraction

SumAsk [2]

We call such works as

Closed GRE

Given Relations: (*member of, award won, work location, ..., father, spouse*)
What are the relations between the subject entity and the object entity expressed by the sentence?
Sentence: "Marie Curie won her first Nobel Prize in Physics for her work on radioactivity with her husband, Pierre."
Subject: Marie Curie
Object: Pierre
Identified Relation: **spouse**

"LLMs as zero-shot relation ~~extractors~~ classifiers"

[2] Li et al., Revisiting large language models as zero-shot relation extractors. *EMNLP'23-Findings*

Background: Generative Relation Extraction

Wadhwa et al. [2]

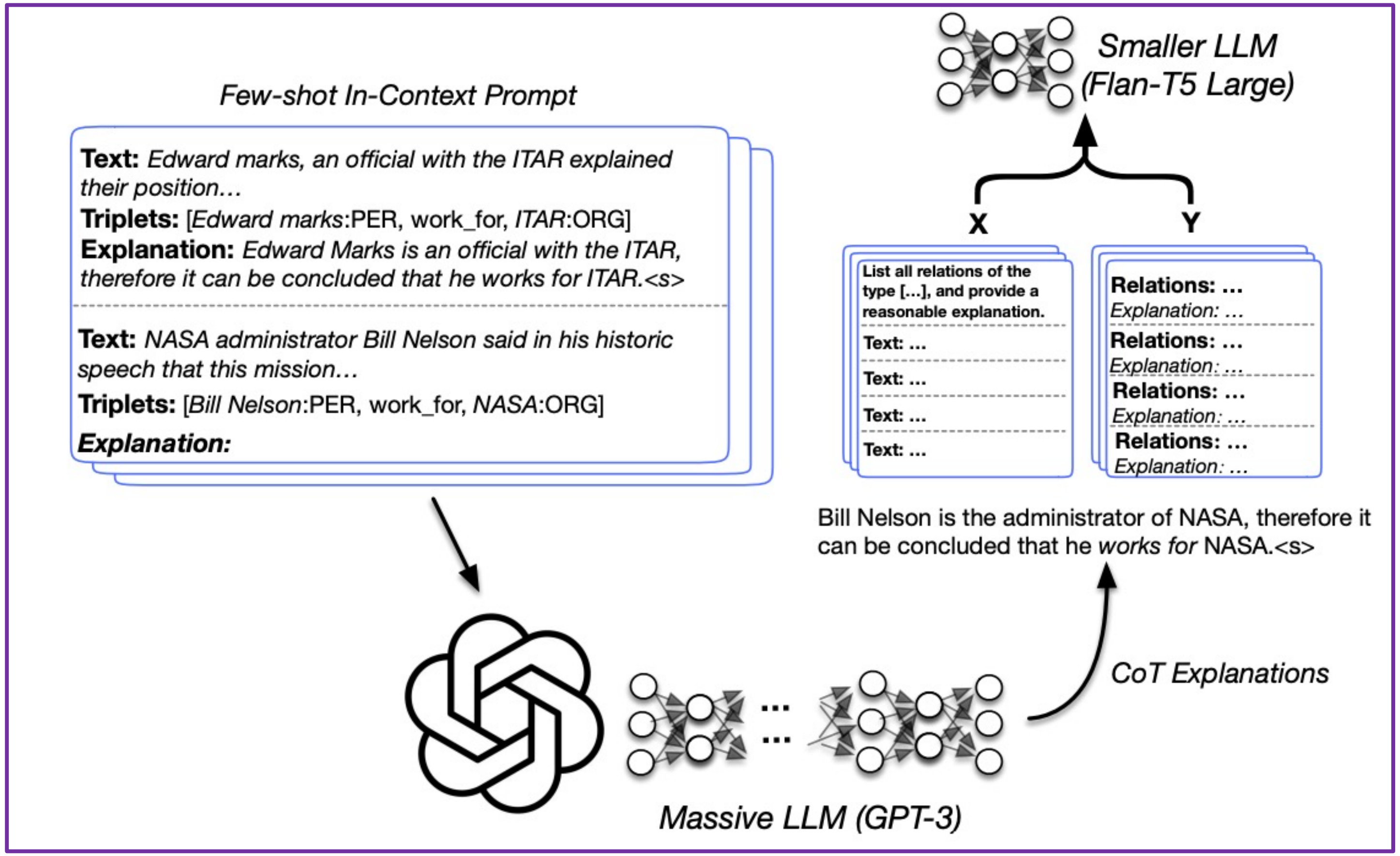
They tested two settings:

(1) GPT few-shot reasoning

we use: *List the entities of the types [LOCATION, ORGANIZATION, PERSON] and relations of types [Organization Based In, Work For, Located In, Live In, Kill] among the entities in the given text.* Since

Predefined sets of entity types and relation types

(2) Flan-T5 Large trained with GPT CoT



[3] Wadhwa, S., Amir, S. and Wallace, B.C., 2023. Revisiting relation extraction in the era of large language models. *ACL 2023*

Background: Generative Relation Extraction

Wadhwa et al. [2]

Why manual evaluation? Too many misclassified predictions as they keep entity types open!

ADE

Four days after the initial injection of 3.6 mg of goserelin acetate, severe dyspnea developed due to worsening pleuritis carcinomatosa, which was considered as a flare-up.

Reference
`[('goserelin acetate','flare')]`
Wrong, but counted as a false negative

Generated
`[('goserelin acetate','severe dyspnea')]`
Correct, but counted as false positives

NYT

Some have called for a memorial to the lynched youth to join the many other shrines here in Waco, a city of 113,000 neighboring President Bush's ranch in Crawford, and home to Baylor University, founded in 1845, the first institution of higher learning in Texas and the largest baptist university in the world.

Reference
`[('texas', '/location/contains', 'waco')]`

Generated
`[('texas', '/location/contains', 'waco'),`
`('texas', '/location/contains', 'crawford')]`
Correct, but counted as a false positive

CoNLL04

On Friday, U.S. Ambassador Vernon A. Walters displayed photographs of one Libyan jet showing shapes resembling missile pods on its wings and fuselage.

Reference
`[('Vernon A. Walters', 'Live_In', 'U.S.')]`
Wrong, but counted as a false negative

Generated
`[('Amb. Vernon A. Walters', 'Work_For', 'U.S')]`
Correct, but counted as a false positive

Out-of-Domain (CoNLL04)

In 1881 , President James A. Garfield was shot by Charles J. Guiteau, a disappointed office-seeker, at the Washington railroad station.

Reference
`[('Charles J. Guiteau', 'Kill', 'President James A. Garfield')]`

Generated
`[('James A. Garfield', 'Shot_By', 'Charles J. Guiteau')]`

Figure 2: Examples of misclassified FPs and FNs from GPT-3 (generated under few-shot in-context prompting scheme) under traditional evaluation of generative output. In each instance, the entity-type of `subject` and `object` was correctly identified.

Future directions We have left several avenues open for further exploration. For example, evaluating LLMs like GPT-3 for RE required collecting manual annotations to identify ostensible "false positive" and "false negative" model outputs which were in fact accurate. Designing models to automate this evaluation might provide similar reliability without the accompanying costs; we provide preliminary work in this direction through the use of simple BERT-style classifiers in Appendix D.

Automated multi-aspect evaluation metrics are needed.

[3] Wadhwa, S., Amir, S. and Wallace, B.C., 2023. Revisiting relation extraction in the era of large language models. *ACL 2023*

Background: Generative Relation Extraction

Wadhwa et al. [2]

We call such works as

Semi-open GRE

List the relation of the types (*member of, award won, work location, ..., father, spouse*) among the entity types (*PERSON, WORK_FIELD, AWARD*)
<EXAMPLE>
Sentence: "Marie Curie won her first Nobel Prize in Physics for her work on radioactivity with her husband, Pierre."
Relations: **[[Marie Curie, spouse, Pierre], [Marie Curie, award won, Nobel Prize], [Marie Curie, work on, Physics]]**

"LLMs as zero-shot entity extractors and relation classifiers"

[3] Wadhwa, S., Amir, S. and Wallace, B.C., 2023. Revisiting relation extraction in the era of large language models. *ACL 2023*

Introduction: Open Generative Relation Extraction

There is a third type of GRE without any limitations of entity types and relation types

Open GRE

Given a sentence, identify and list the relationships between entities within the text.
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.
<EXAMPLE>
Sentence: "Marie Curie won her first Nobel Prize in Physics for her work on radioactivity with her husband, Pierre."
Relations: **[[Marie Curie, won, Nobel Prize in Physics], [Marie Curie, worked on, radioactivity], [Marie Curie, worked with, Pierre], [Radioactivity, researched by, Marie Curie and Pierre], [Marie Curie, was awarded for, work on radioactivity], [Marie Curie, is married to, Pierre], [Pierre, is the husband of, Marie Curie], [Marie and Pierre, collaborated on, radioactivity research], [Nobel Prize in Physics, awarded for, work on radioactivity], …**

"LLMs as zero-shot relationship (both entity and relation) extractors"

Based on extremely strong text understanding capabilities of LLMs. We believe that RE method in the LLM era should be revolutionized:

We should transfer from the strategy

"manually defining a set of relation types" → "finding matches between entities"

to

exploring as many relations and entities as possible without constraints → gathering and sorting relationships (e.g., clustering)

Introduction: GenRES (Genarative Relation Extraction Scoring)

We believe hard matching Precision/Recall/F1 metrics are no longer adequate to evaluate GRE

Good metrics for GRE should be able to evaluate :

1. How much content of the source text is covered by the relationships extracted (by comparing triples* to the source text)

2. How many unique relationships are extracted (by comparing similarity within the extracted triples)

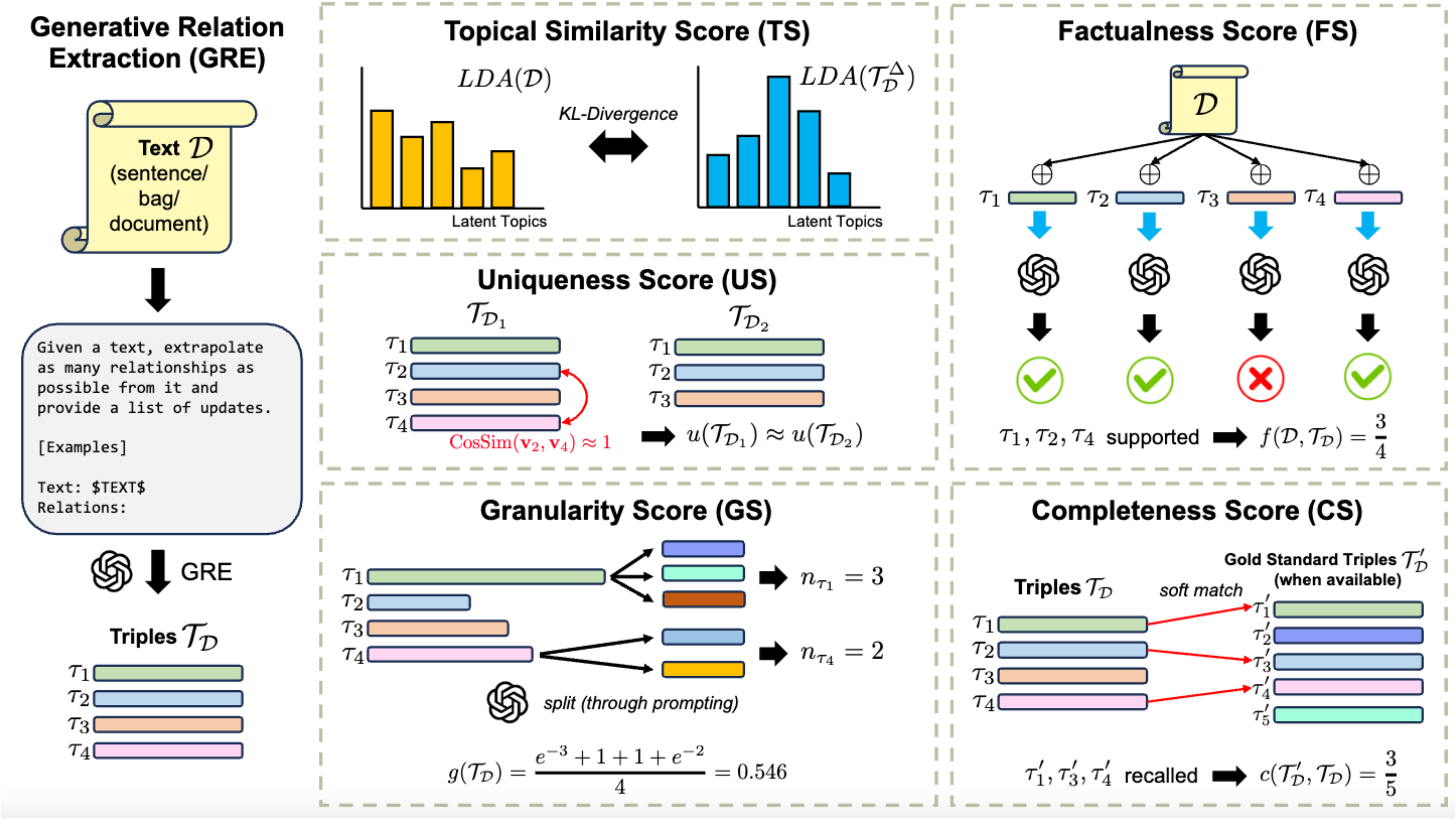
3. How factual the extracted triples are, referring to the source text (by factualness verification treating source text as the “knowledge base”)

4. How atomic the extracted triples are (by asking LLM to split each triple)

5. How many ground truth relations are predicted (by computing soft matching recall)

* We refer relationships as triples in the format of <s, r, o> where s is subject entity, r is relation, and o is object.

Method: GenRES – Overview



Method: GenRES

Topical Similarity Score (TS)

“How much content of the source text are covered by the relationships extracted (by comparing triples* to the source text)”

Text $\mathcal{D}$

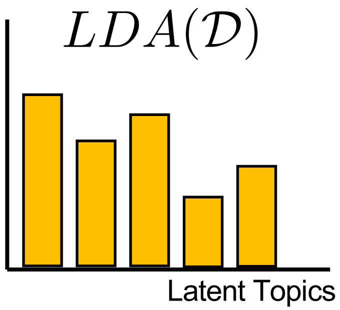
Four workers died in a massive oil rig fire that raged for hours off the coast of Mexico Wednesday. Mexican state oil company Pemex said 45 workers were injured in the blaze, which began early Wednesday morning. Two of them are in serious condition, the company said. Authorities evacuated about 300 people from the Abkatun Permanente platform after the fire started, Pemex said. At least 10 boats worked to battle the blaze for hours. The fire had been extinguished by Wednesday night, Pemex said in a Twitter post. The company denied rumors that the platform had collapsed and said there was no oil spill as a result of the fire. The state oil company hasn't said what caused the fire on the platform, which is located in the Gulf of Mexico's Campeche Sound. The fire began in the platform's dehydration and pumping area, Pemex said. CNN's Mayra Cuevas contributed to this report.

Generative Relation Extraction

Triples $\mathcal{T}_{\mathcal{D}}$

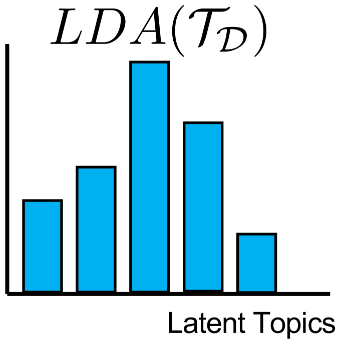
[Four workers | were died in | oil rig fire],
[45 workers | were injured in | the blaze],
[Two workers | are in | serious condition],
[300 people | were evacuated from | the platform],
[The fire | had been extinguished by | Wednesday night],
[The fire | did not result in | oil spill].

Topical Distribution



KL-Divergence

Topical Similarity Score

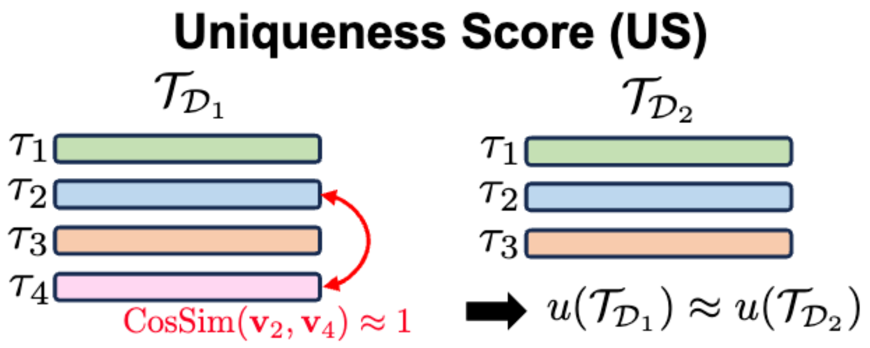


$$t(\mathcal{D}, \mathcal{T}_{\mathcal{D}}^{\Delta}) = e^{-\sum_{i=1}^{K} LDA(\mathcal{D})_i \cdot \log\left(\frac{LDA(\mathcal{D})_i}{LDA(\mathcal{T}_{\mathcal{D}}^{\Delta})_i}\right)}$$

Method: GenRES

Uniqueness Score (US)

"How many unique relationships are extracted (by comparing similarity within the extracted triples)"



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

threshold

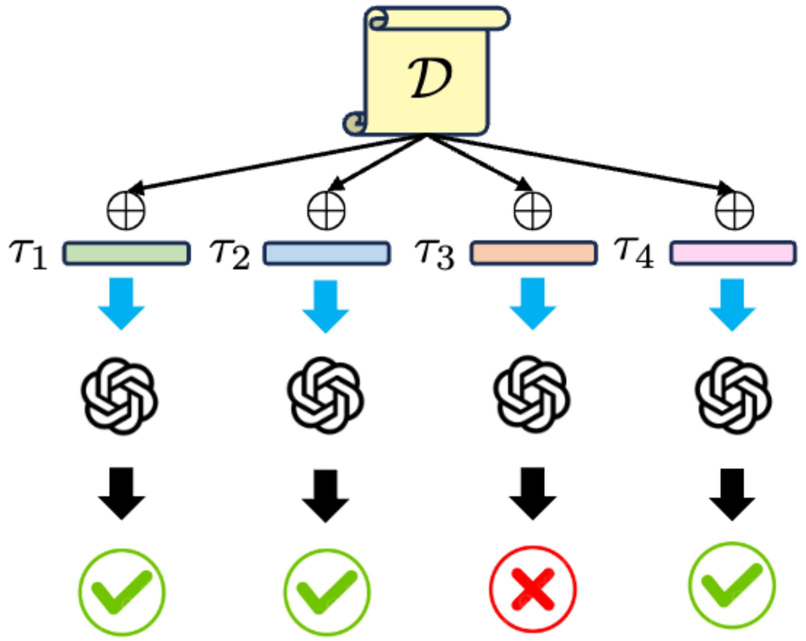
Different triples with similar semantic meaning should be regarded as redundant, this score check whether a model is extracting repeated relationships or not.

Method: GenRES

Factualness Score (FS)

"How factual the extracted triples are, referring to the source text (by factualness verification treating source text as the "knowledge base")"

Factualness Score (FS)



τ_1, τ_2, τ_4 supported ➡ $f(\mathcal{D}, \mathcal{T}_{\mathcal{D}}) = \frac{3}{4}$

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

Fack-checking prompt:

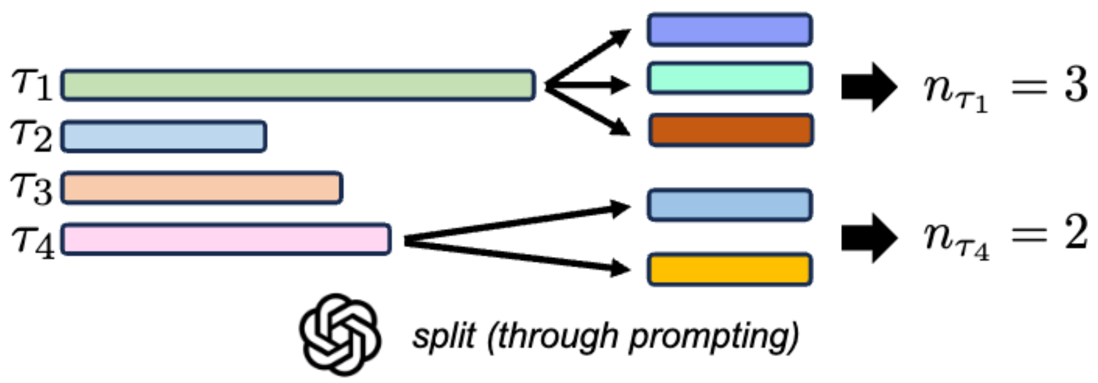
```
Evaluate the factualness of an extracted relationship (triplet) based on
the given source text. Indicate whether the relationship accurately
reflects the information in the source text by responding with "true" or
"false".
You should only output "true" or "false" with no additional information.

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:
```

Method: GenRES

Granularity Score (GS)

“How atomic the extracted triples are (by asking LLM to split each triple)”

Granularity Score (GS)



$$g(\mathcal{T}_\mathcal{D}) = \frac{e^{-3} + 1 + 1 + e^{-2}}{4} = 0.546$$

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

Granularity-checking prompt:

```
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
```

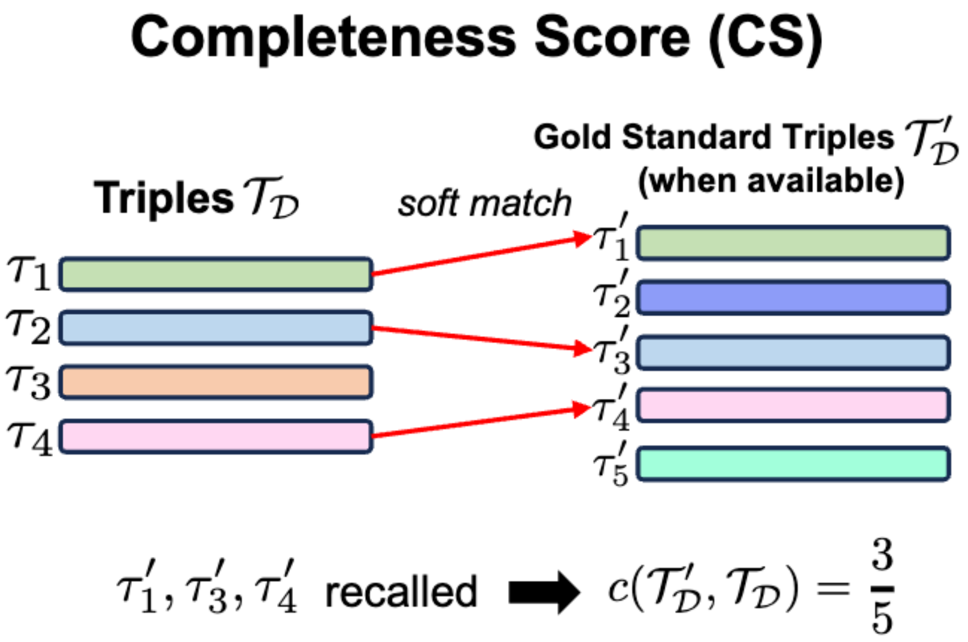
⋮ (9 examples)

```
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:
```

Method: GenRES

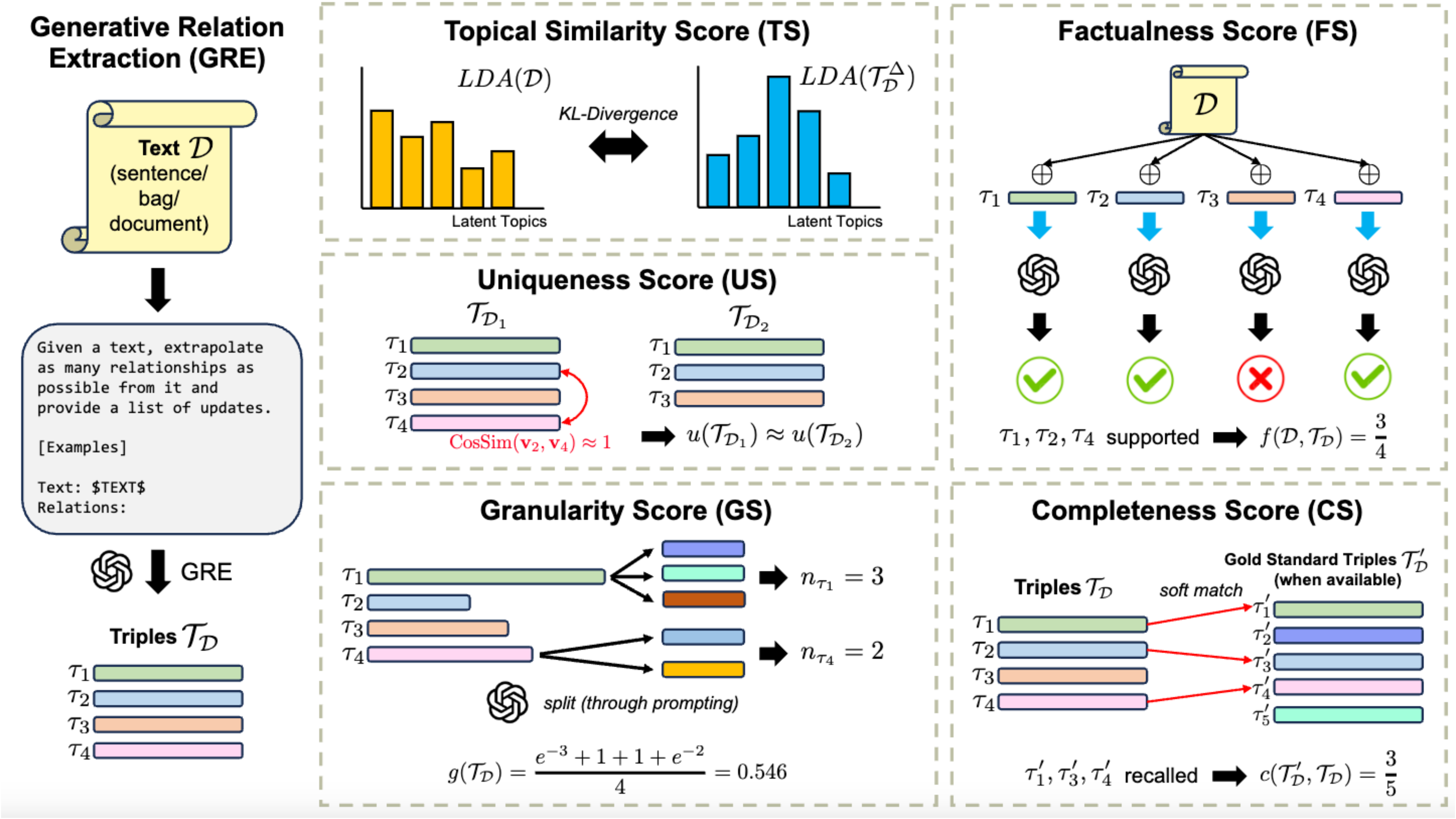
Completeness Score (CS)

"How many ground truth relations are predicted (by computing soft matching recall)"



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

Method: GenRES – Overview



Datasets

We test on 6 datasets:

2 document-level datasets:

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.

2 bag-level datasets:

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.

2 sentence-level datasets:

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.

Results – why not Precision/Recall/F1 metrics?

	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** he highest GREScores for each dataset.

We found that those hard matching-based metrics do not work for both semi-open and open GRE methods

While our Factualness Score (soft precision) and Completeness Score (soft recall) can well indicate the quality of the extract triples

Results – why not Precision/Recall/F1 metrics and why Open GRE?

Comparative Analysis on an example of NYT10m dataset

Inaccurate labels (pure recall is not reliable)

I. Text "Peter Munk , founder and chairman of Barrick Gold in Toronto , has warned that an exodus of head offices to other countries will cause , among other things , lower levels of charitable donations and fewer opportunities for skilled workers ."

II. Ground Truth [Peter Munk, place lived, Toronto], [Barrick Gold, advisors, Peter Munk], [Barrick Gold, location, Toronto], [Barrick Gold, company, Peter Munk], [Barrick Gold, founders, Peter Munk], [Peter Munk, company, Barrick Gold], [Barrick Gold, place lived, Toronto]

III. Predefined Relation Types: (*administrative_divisions, advisors, capital, children, company, contains, country, county_seat, ethnicity, featured_film_locations, founders, geographic_distribution, location, locations, majorshareholders, nationality, neighborhood_of, place_founded, place_lived, place_of_birth, place_of_burial, place_of_death, religion*)

IV. Predefined Entity Types: (*business, company, country, deceasedperson, ethnicity, event, film, location, neighborhood, people, person, region, time, us_county*)

Closed GRE

Input: I, III, and entity pairs in II.

Output:
[Peter Munk, place founded, Toronto]
[Barrick Gold, founders, Peter Munk] (*FS*, *CS*)
[Barrick Gold, location, Toronto] (*FS*, *CS*)
[Barrick Gold, founders, Peter Munk] (*FS*, *CS*)
[Barrick Gold, founders, Peter Munk] (*FS*, *CS*)
[Peter Munk, founder of, Barrick Gold] (*FS*, *CS*)
[Barrick Gold, location, Toronto] (*FS*, *CS*)

Evaluation:
Tranditional: *P*: 71.4, *R*: 28.6, *F1*: 40.8
GREScores:
TS: 3.6, *US*: 66.7, *FS*: 85.7, *GS*: 100, *CS*: 57.1

Semi-open GRE

Input: I, III, and IV.

Output:
[Peter Munk, advisors, Barrick Gold] (*CS*)
[Peter Munk, founders, Barrick Gold] (*FS*, *CS*)
[Barrick Gold, location, Toronto] (*FS*, *CS*)
[Peter Munk, warning, exodus [head offices location, other countries],
[exodus, cause, lower levels of charitable donations and fewer opportunities for skilled workers] (*FS*, *GS*)

Evaluation:
Tranditional: *P*: 16.7, *R*: 14.2, *F1*: 15.4
GREScores:
TS: 22.1, *US*: 100.0, *FS*: 50.0, *GS*: 85.6, *CS*: 71.4

Open GRE

Input: I (text only).

Output:
[Peter Munk, founder of, Barrick Gold] (*FS*, *CS*)
[Peter Munk, chairman of, Barrick Gold] (*FS*, *CS*)
[Barrick Gold, located in, Toronto] (*FS*, *CS*)
[Peter Munk, based in, Toronto] (*FS*)
[Peter Munk, warn, effects of exodus of head offices] (*FS*)
[exodus of head offices, will cause, lower levels of charitable donations] (*FS*)
[exodus of head offices, will cause, fewer opportunities for skilled workers] (*FS*)

Evaluation:
Tranditional: *P*: 0, *R*: 0, *F1*: 0
GREScores:
TS: 44.9, *US*: 80.0, *FS*: 100.0, *GS*: 100.0, *CS*: 57.1

Good extraction gets all zeros by P/R/F1

Inaccurate prediction given a fixed set of relation types

Inaccurate entity recognition given a fixed set of entity types

The generation is the best among the three

Results – Testing leading LLMs' Open GRE Capabilities

On **CDR** and **DocRED** – two document-level datasets:

		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.

Results – Testing leading LLMs' Open GRE Capabilities

On **NYT10m** and **Wiki20m** – two bag-level datasets:

		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.

Results – Testing leading LLMs' Open GRE Capabilities

On **TACRED** and **Wiki80** – two sentence-level datasets:

		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.

Results – Testing leading LLMs' Open GRE Capabilities

		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	[illegible]
	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.

	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
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**

	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
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

Observations:

(1) LLaMA-2-70B, GPT-4-Turbo, and OpenChat-3.5 notably lead in performance. Small LLM OpenChat-3.5 (7B) achieves comparable or even better performance than large LLMs.

(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.

(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) No clear correlation between the number of triples, Topical Similarity (TS), and Uniqueness Score (US), indicating the distinct significance of each metric.

(5) GPT-4-Turbo outperforms human labels on factualness.

Results – Robustness of GenRES and Its Alignment with Human Evaluation

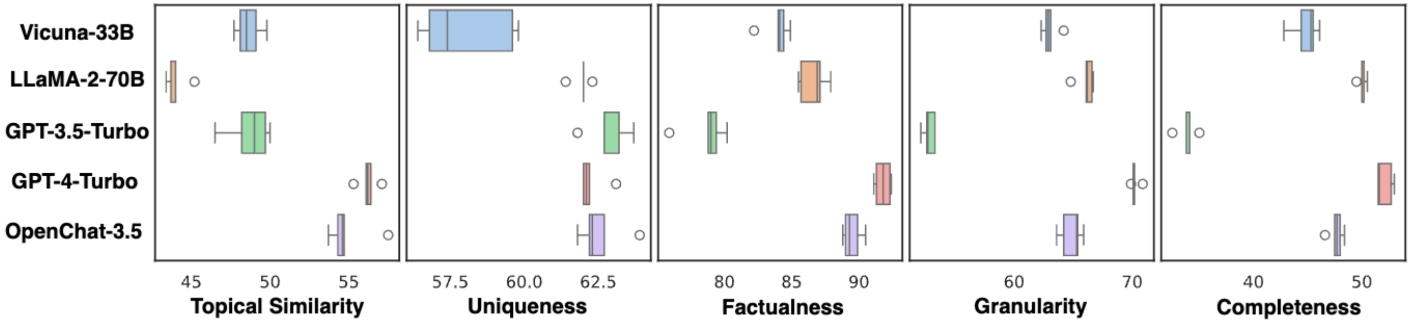


Figure 4: **GRE performance of five LLMs on Wiki20m, each with five runs with random seeds.**

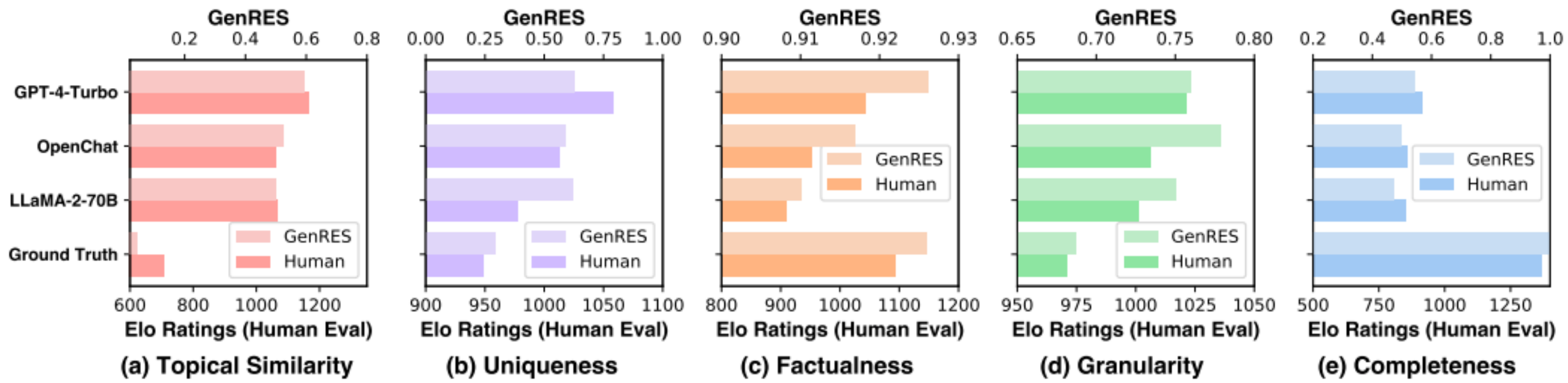


Figure 5: **Human Preference Evaluation (Elo Ratings) vs GenRES Evaluation on 100 Wiki20m samples.**

Observations:

(1) The robustness of GenRES as an evaluation framework across different metrics

(2) In most cases, GenRES aligns well with human evaluation of generative relation extraction.

Thank you!

Code: https://github.com/pat-jj/GenRES

Patrick (Pengcheng) Jiang
pj20@illinois.edu