Structured Discourse Representation for Factual Consistency Verification

Kun Zhang[1,2] and **Oana Balalau**[1,2] and **Ioana Manolescu**[1,2]
[1]Inria, [2]Institut Polytechnique de Paris

Abstract

Analysing the differences in how events are represented across texts, or verifying whether the language model generations hallucinate, requires the ability to systematically compare their content. To support such comparison, structured representation that captures fine-grained information plays a vital role. In particular, identifying distinct atomic facts and the discourse relations connecting them enables deeper semantic comparison. Our proposed approach combines structured discourse information extraction with a classifier, **FDSpotter**, for factual consistency verification. We show that adversarial discourse relations pose challenges for language models, but fine-tuning on our annotated data, **DiscInfer**, achieves competitive performance. Our proposed approach advances factual consistency verification by grounding in linguistic structure and decomposing it into interpretable components. We demonstrate the effectiveness of our method on the evaluation of two tasks: data-to-text generation and text summarisation. Our code and dataset will be publicly available on GitHub.

1 Introduction

The analysis of public discourse plays a vital role in social sciences, in particular for sociology (Wodak and Meyer, 2015), history (Rule et al., 2015), and computational journalism (Cazalens et al., 2018). A key task in this context is to quantify whether two texts convey the same information through their expressed content. Such comparisons enable the detection of media bias: for instance, the outlets with different political orientations usually tend to selectively report different subsets of information units, which has been a phenomenon known as omission bias (Baker et al., 1994). More recently, similar techniques have been utilised to assess the factual consistency of text generation, verifying whether the content in the output can be grounded in the reference or source text (Tang et al., 2024).

Extracting RDF-style triples from two texts and then assessing their overlap appears to be a straightforward strategy for comparing texts. However, prior works that extract structured contents for factual consistency verification (Joty et al., 2017; Goodrich et al., 2019; Goyal and Durrett, 2020) exhibit a performance gap compared with state-of-the-art approaches based on sentence-level entailment (Scirè et al., 2024) or large language model prompting (Luo et al., 2023). The gap is primarily due to *natural language expresses meaning in more rich and subtle ways than RDF triples*. Firstly, the contents in adverbials and complements are difficult to encode in the standard subject–predicate–object format. For example, in the text "Barack Obama was elected *in 2008* as the President of the USA", "*in 2008*" conveys crucial temporal information, yet it is hard to incorporate into a triple structure. Secondly, triple-based representations fail to encode discourse-level semantics, such as contrastive or causal relations. For instance, the causal link in "Lu Xun realised that being a doctor could not save Chinese people's minds, *so* he gave up his medical career and became a writer" cannot be captured through RDF triples. Beyond the limitations of the triple format, both large language models (Dubey et al., 2024) and smaller classifiers (He et al., 2021) often fail to detect contradictions introduced by adversarial discourse connectives (Miao et al., 2024).

Our contributions in the paper are as follows:

- We propose a structured representation of textual information that extends RDF-style triples by incorporating *complements and adverbials* into the atomic facts, and by preserving the *discourse relations* between them. We show that few-shot LLMs can extract this structure with high quality.
- We introduce a novel method for evaluating *cross-textual factual consistency*. A classifier,

FDSpotter, is trained to verify the presence of atomic facts and discourse relations across texts. We show that detecting whether discourse relations are preserved is challenging, but we achieve strong performance after fine-tuning on our annotated DiscInfer dataset.

- Empirical results demonstrate the effectiveness of our approach in evaluating data-to-text generation and text summarisation.

2 Related Work

Open Information Extraction (OIE) Open information extraction (Etzioni et al., 2008; Upadhyay et al., 2023; Pai et al., 2024) tools extract triples of the form ⟨`subject, predicate, object`⟩ from unstructured texts. In OIE, the predicates are not known beforehand, and the subjects and objects are not necessary to be named entities. While the triple format is one of the most common used, n-tuples have also been proposed (Niklaus et al., 2018). In some approaches (Del Corro and Gemulla, 2013), the elements of the tuples are defined as syntax elements, such such as subject, verb, object, complement, etc. In other works (Dong et al., 2023) the elements of the tuples are phrases. Open information extraction can be performed by semantic role labelling (Chen et al., 2025a), sequence tagging (Stanovsky et al., 2018), and large language models (Xu et al., 2023).

Meaning Representation and Parsing Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a semantic representation framework designed to abstract away from surface syntax and encode the meaning of the text in the form of *rooted, directed, acyclic graphs (DAGs)* that capture the relation between the elements of the text. One key challenge of AMR parsing lies in the annotation process, which requires high linguistic expertise. A well-trained annotator typically spends around 10 minutes to produce a single AMR graph (Sadeddine et al., 2024). AMR parsers also struggles with out-of-distribution inputs and new domains (Lee et al., 2022). Discourse parsing (Braud et al., 2023) identifies the structure of texts by linking spans of discourses through semantic and pragmatic discourse relations. can be *explicitly* or *implicitly* expressed. State-of-the-art parsing methods achieve strong performance in discourse unit segmentation (Metheniti et al., 2023), but Liu et al. (2023a) reports discourse connective detection remains challenging. Shallow discourse parsing identifies discourse argument spans and the corresponding discourse connectives jointly (Wang and Lan, 2015; Oepen et al., 2016), but state-of-the-art models on the task still exhibit room for improvement (Xue et al., 2015, 2016; Knaebel, 2021).

Faithfulness Evaluation Metrics. Metrics for comparing two texts have been used to evaluate text generation tasks, such as summarisation, data-to-text generation, and machine translation. Well known metrics include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019), and BLEURT (Sellam et al., 2020). Overlap measures based on n-*grams*, such as BLEU, ROUGE, and METEOR, have been widely used in the literature, however *word embeddings* metrics, such as BERTScore, BLEURT, MoverScore are becoming standard, as they correlate better with human evaluation than the surface based metrics. Recently, explainable metrics have been developed, for example, FActScore (Min et al., 2023), FactSpotter (Zhang et al., 2023), and FENICE (Scirè et al., 2024), which allow a more fine-grained understanding of differences between texts. FactSpotter propose a metric for factual faithfulness evaluation on data-to-text generation, which gives a score for how well each triple from the structured data is represented in the textual generation. Similarly, FActScore (Min et al., 2023) metric allows detecting hallucinations in text generation. This metric has two building blocks: decomposing the first text into atomic facts, and testing if each atomic fact is present in the second text. In Nawrath et al. (2024), the authors propose splitting the text into sub-sentences using AMR, however, the authors do not deal with maintaining the connection between these sub-sentences. A recent approach that achieves state-of-art results on measuring factual faitfulness in textual summarization FENICE (Scirè et al., 2024) proposes to split a text in atomic facts and then use natural language inference to decide if the facts are expressed in a new text. LLMs, such as ChatGPT, have been also proposed as factually evaluators in zero shot or chain-of-thought setting (Luo et al., 2023). By the time of paper submission, we saw no metric that account explicitly for the discourse relation. However, by the time of paper acceptance, we have seen a concurrent work on discourse-level evaluation for summarisation (Zhong and Litman, 2025)

3 Problem statement

To evaluate the quality of the text generation task, a common method is to calculate the similarity between the generated text and the ground-truth text. Taking graph-to-text generation and text summary as examples, we aim to quantify the similarity of two texts from two perspectives:

1. Do the two texts state the same atomic facts?

2. Are the discourse relations between the atomic facts the same in both texts?

To achieve this goal, we firstly focus on how to represent the atomic facts and the discourse relations between them. Inspired by ClausIE (Del Corro and Gemulla, 2013) and previous research in discourse relation recognition (Prasad et al., 2019), we propose a format for representing text information in Section 4. Then we introduce our method for estimating whether an atomic fact or a discourse relation is preserved in two texts in Section 5 with the model trained in Section 6.

4 Structured Information Representation

Given a text $\mathbf{T}$, our goal is to extract the atomic facts $\mathbf{A} = \{A_1, ..., A_m\}$ and discourse relations between the atomic facts $\mathbf{D} = \{D_1, ..., D_p\}$. The combination of the two structured representations captures rich information in the text, as an example illustrated in Table 1.

4.1 Atomic Fact Representation

A clause is the smallest grammatical unit in English for expressing a complete proposition[1]. We consider an atomic fact to be a clause, structured as a tuple: ⟨`subject, predicate, direct object, indirect object, short adverbial, short complement`⟩. Note that simpler triples of the form ⟨`subject, predicate, object`⟩ have been found insufficient to represent varied atomic facts (Suchanek, 2020; Nawrath et al., 2024; Sadeddine et al., 2024); the short adverbial and the short complement bring important information about the current atomic fact. The short adverbial describes *how, when* and *where* about the predicate, while the short complement has extra information about the subject or the object. Each atomic fact should at least have a predicate, and usually has a subject. The object, adverbial, or complement may be absent, and we leave them blank in the extractions for such cases.

If an element of the atomic fact tuple contains another fact in its non-finite verb, relative clause, or appositive [2]), we model this as an extra atomic fact. For instance, the text *"Amy gave a gift to her best friend, Kate"* should be represented as two atomic facts: ⟨`Amy, gave, a gift, to Amy's best friend Kate`⟩ and ⟨`Kate, is the best friend of, Amy`⟩.

4.2 Discourse Relation Representation

Discourse relations connecting atomic facts encapsulate an important part of the text information. The discourse connectives are syntactically expressed by conjunctives, relative adverbs or other transitional phrases. The detailed examples are illustrated in the Appendix A. We model a discourse relation D_j with format ⟨`fact1, connective, fact2`⟩.

We focus on the following groups of discourse relations, as defined in the PDTB dataset (Prasad et al., 2019), and their respective connectives:

- **temporal :** *precedence* (e.g., before, till), *succession* (e.g., after, subsequently), *synchronous* (e.g., when, at the same time);

- **comparison :** *concession* (e.g., although, even if), *contrast* (e.g., in contrast, however), *similarity* (e.g., similarly, in the same way);

- **contingency :** *reason* (e.g., due to, because), *result* (e.g., consequently, therefore), *condition* (e.g., provided that, in case), *negative condition* (e.g., unless).

We did not include **expansion** connectives: *simple conjunction* (e.g., and); *restatement* (e.g., in other words); *specification* (e.g., especially); *instantiation* (e.g., for example); *generalisation* (e.g., in summary). The reason is that their presence does not affect important textual information. Thus, when comparing two texts for factual consistency, we can ignore the presence of these connectives.

4.3 Implementation of Extraction

The state-of-the-art methods have achieved high performance in discourse unit segmentation

[1] https://dictionary.cambridge.org/grammar/british-grammar/clauses-and-sentences

[2] Non-finite verbs are verbs that do not show tense, person, or number, e.g., "running" in "I saw a dog running". Relative clauses are introduced by relative pronouns to describe the antecedents (e.g., in "Paris is a city which is the capital of France", "which" introduces a relative clause). An appositive is a noun phrase (NP) after another NP to provide additional information about it (e.g., in "Paris, the capital of France, is a good city", "the capital of France" is the appositive of "Paris".

Subject	Predicate	Object Direct	Object Indirect	Short Adverbial and Complement
Amy	gave	a gift	to Amy's best friend, Kate	last week
Amy	was thankful about	the exam's results	-	-
Kate	is best friend of	Amy	-	-
Kate	said	Amy did not need to give a gift	-	-
Kate	helped	Amy	-	to revise all the course materials
Amy and Kate	held	a party	-	until midnight
Amy and Kate	didn't get up	-	-	early on the next day

(a) Atomic facts extracted from the given text.

Fact 1	Connective	Fact 2
Amy gave a gift to Amy's best friend, Kate last week	because of	being thankful about the exam's results
Kate said Amy did not need to give a gift to Kate	even if	Kate helped Amy revise all the course materials
Amy would have failed the exam	if	Kate were not helping
Amy and Kate held a party until midnight	As a result	Amy and Kate didn't get up early on the next day

(b) Discourse relations extracted from the given text.

Table 1: Atomic facts and discourse relations representation, given the text: "*Being thankful about the exam's results, Amy gave a gift to her best friend, Kate, last week. Kate said that Amy did not need to do this, even if Kate helped Amy revise all the course materials. Amy would have failed the exam if Kate were not helping. They held a party until midnight. As a result, they didn't get up early on the next day.*"

	Atomic Fact											
	Strict			LCS			SBERT Elem			SBERT Full		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Human	77.21	69.17	72.97	81.45	72.97	76.98	86.20	77.22	81.46	90.15	80.76	85.20
GPT4-Turbo	68.47	**52.78**	**59.61**	73.84	56.92	**64.29**	79.36	61.17	**69.09**	88.76	68.42	**77.27**
GPT4o	**68.54**	49.03	57.17	**75.89**	54.28	63.29	**82.68**	59.14	68.95	**92.61**	66.24	77.24
Llama3.1 8B	49.54	22.36	30.81	57.32	25.87	35.65	68.79	31.05	42.79	76.40	34.49	47.52
Llama3.1 70B	57.08	51.53	54.16	63.17	**57.03**	59.94	69.59	**62.83**	66.04	80.49	**72.67**	76.38
Discourse Relation												
Human	81.94	81.94	81.94	84.93	84.93	84.93	88.10	88.10	88.10	90.32	90.32	90.32
GPT4-Turbo	75.00	**76.60**	**75.79**	77.45	79.09	**78.26**	79.33	81.02	80.17	80.26	81.97	81.11
GPT4o	61.81	70.63	65.93	71.60	**81.82**	76.37	79.83	**91.23**	**85.15**	86.06	**98.35**	**91.80**
Llama3.1 8B	**92.93**	60.13	73.02	**94.41**	61.09	74.18	**95.56**	61.83	75.08	**96.71**	62.58	75.99
Llama3.1 70B	59.85	64.23	61.96	66.57	71.44	68.92	72.08	77.36	74.63	75.50	81.02	78.16

Table 2: Quality of extraction for different large language models.

(Metheniti et al., 2023), so combining discourse unit segmentation with semantic role labelling (Zhao and Penn, 2024; Chen et al., 2025b) is an effective approach for obtaining structured atomic facts. However, the task of discourse connective detection (Braud et al., 2023) remains challenging, and the state-of-the-art work on shallow discourse parsing (Knaebel, 2021) which jointly identifies argument spans and their corresponding discourse connectives has not been advancing recently.

Based on the status quo of the research, in this work, we use large language models with few-shot prompting for jointly extracting atomic facts and discourse relations. We also evaluate the quality of the extractions on a test set of 50 samples from the Causal News Corpus (Tan et al., 2022). Two human annotators provided the gold extractions. We call any one of the fields of an atomic fact ⟨`subject, predicate, direct object, indirect object, short adverbial and complement`⟩ or a discourse relation ⟨`fact1, connective, fact2`⟩ as an *element*.

Given an input text $\mathbf{T}$, we have the golden atomic facts $\mathbf{A}^g$ and discourse relations $\mathbf{D}^g$. The outputs of the atomic fact and discourse relation extraction are $\mathbf{A}^m$ and $\mathbf{D}^m$. Similar to Castro Ferreira et al. (2020), we compare both atomic fact and discourse relation extractions with the ground-truth annotations. For each pair of corresponding tuples in the extraction and the ground truth, we compute several metrics to assess the extraction quality:

- **Strict match**: Compute whether all the tokens in each element of the extracted tuple strictly match with ground truth, e.g., whether the subjects of the atomic facts are the same.
- **LCS ratio**: Find the longest common subsequence (LCS) (Wagner and Fischer, 1974) between each element in the extraction and the corresponding element in the ground-truth. Compute the average (over all elements in the extracted atomic fact or discourse relation) of the LCS length divided by the length of the corresponding ground truth element.

- **SBERT element**: Compute the average (over all elements) of SBERT (Reimers and Gurevych, 2019) similarity for each element between the extractions and the annotations;
- **SBERT full**: Concatenate all the elements in each tuple for both of the extractions and the ground-truth annotations, then use SBERT to compute the similarity.

To compute the metrics, we assign each extracted atomic fact A_j^m or discourse relation D_j^m, to exactly one atomic fact A_i^g or discourse relation D_i^g using Hungarian Algorithm, similarly as Yang et al. (2023). We evaluate the performance of Llama3.1 (Dubey et al., 2024), GPT4o and GPT4-Turbo (OpenAI et al., 2024) in Table 2 for both extractions. We observe that two humans achieve the highest annotation agreement. For open-source models, Llama3.1 8B is good on discourse relation extraction, but struggles on atomic fact extraction. Llama3.1 70B is competitive on both extractions. GPT4 models perform best across all metrics. We present more complete results in Appendix E.

5 Factual Inclusion and Factual Overlap

Given two texts $\mathbf{T}^1$ and $\mathbf{T}^2$, which are composed of the atomic fact sets $\mathbf{A}^1 = \{A_1^1, \ldots, A_m^1\}$, $\mathbf{A}^2 = \{A_1^2, \ldots, A_n^2\}$ and discourse relation sets $\mathbf{D}^1 = \{D_1^1, \ldots, D_o^1\}$, $\mathbf{D}^2 = \{D_1^2, \ldots, D_p^2\}$. m, n are the numbers of the atomic facts, and o, p are the numbers of the discourse relations in the texts. The Factual Inclusion (FI) and Factual Overlap (FO) scores between the texts are computed as follows to verify the factual consistency between the texts.

5.1 Intrinsic Confidence of Extraction

Let $\mathbf{FD}$ be a model computing the probability of a structured atomic fact A_i or a discourse relation D_j to be expressed in the given text $\mathbf{T}$. In practice, $\mathbf{FD}$ is a Transformer-based classifier that estimates *how well the atomic facts* $\mathbf{A}$ *and discourse relations* $\mathbf{D}$ *represent the information in the text*, and the output is referred to as the **intrinsic confidence**. The intrinsic confidence C of each atomic fact A_i and discourse relation D_j for text $\mathbf{T}$ is computed as:

$$C(A_i|T) = \mathbf{FD}(A_i, \mathbf{T}),$$
$$C(D_j|T) = \mathbf{FD}(D_j, \mathbf{T}).$$

The intrinsic confidence can serve two key purposes in dealing with the uncertainty inherent in the extraction. First, the intrinsic confidence can be used for selecting the most faithful extraction candidate on top of beam search. Second, within each extraction, the intrinsic confidence provides a principled mechanism to filter out the hallucinated atomic facts and discourse relations that are not sufficiently supported by the text.

5.2 Cross-Textual Fact Corroboration

To compute the content overlap between $\mathbf{T}^1$ and $\mathbf{T}^2$, the second step is to verify *whether each atomic fact and discourse relation from one text is also expressed in the other text*; We call this **extrinsic confidence** computation. We use the same model $\mathbf{FD}$ as for intrinsic confidence computation, but the atomic facts and the discourse relations are extracted from another text for comparison. The extrinsic confidence of the atomic fact A_i^1 from $\mathbf{T}^1$ being expressed in $\mathbf{T}^2$ is computed as:

$$C(A_i^1|\mathbf{T}^2) = \mathbf{FD}(A_i^1, \mathbf{T}^2),$$

and similarly, the extrinsic confidence of the discourse relation D_j^1 from $\mathbf{T}^1$ expressed in $\mathbf{T}^2$ is:

$$C(D_j^1|\mathbf{T}^2) = \mathbf{FD}(D_j^1, \mathbf{T}^2).$$

We also compute the extrinsic confidence of each atomic fact A_i^2 and discourse relation D_j^2 from $\mathbf{T}^2$ represented in the text $\mathbf{T}^1$, i.e., $C(A_i^2|\mathbf{T}^1) = \mathbf{FD}(A_i^2, \mathbf{T}^1)$ and $C(D_j^2|\mathbf{T}^1) = \mathbf{FD}(D_j^2, \mathbf{T}^1)$.

5.3 Factual Inclusion Score

We define a score to assess whether all atomic facts $\mathbf{A}^1$ and discourse relations $\mathbf{D}^1$ extracted from $\mathbf{T}^1$ are included in $\mathbf{T}^2$. The Factual Inclusion Score ($\mathbf{FI}$) is computed based on two components: the inclusion of atomic facts ($\mathbf{FI}_A$) and the inclusion of discourse relations ($\mathbf{FI}_D$).

To mitigate the effect of potential hallucinated extractions, we apply the threshold $\theta = 0.5$ to the intrinsic confidence, i.e., only the atomic facts and discourse relations with their intrinsic confidence above θ in $\mathbf{T}^1$ are considered for Factual Inclusion. We define the filtering function as:

$$\delta_\theta(x) = \begin{cases} 1, & \text{if } x > \theta, \\ 0, & \text{otherwise.} \end{cases}$$

We then score the inclusion of the atomic facts from $\mathbf{T}^1$ into $\mathbf{T}^2$ as follows:

$$\mathbf{FI}_A(\mathbf{T}^1 \subset \mathbf{T}^2) = \sum_{i=1}^{m} \delta_\theta(C(A_i^1|\mathbf{T}^1))C(A_i^1|\mathbf{T}^2).$$

Similarly, the inclusion of the discourse relations from $\mathbf{T}^1$ within $\mathbf{T}^2$ is:

$$\mathbf{FI}_D(\mathbf{T}^1 \subset \mathbf{T}^2) = \sum_{j=1}^{o} \delta_\theta(C(D_j^1|\mathbf{T}^1))C(D_j^1|\mathbf{T}^2).$$

The overall Factual Inclusion Score (**FI**) combines the inclusion score of atomic facts ($\mathbf{FI}_A$) and discourse relations ($\mathbf{FI}_D$), i.e.,

$$\mathbf{FI}(\mathbf{T}^1 \subset \mathbf{T}^2) = \frac{1}{Z_I}(\mathbf{FI}_A(\mathbf{T}^1 \subset \mathbf{T}^2) + \mathbf{FI}_D(\mathbf{T}^1 \subset \mathbf{T}^2)).$$

Z_I is a normalisation factor that scales the score according to the number of faithfully extracted atomic facts and discourse relations from $\mathbf{T}^1$, i.e.,

$$Z_I = \sum_{i=1}^{m} \delta_\theta(C(A_i^1|\mathbf{T}^1)) + \sum_{k=1}^{o} \delta_\theta(C(D_k^1|\mathbf{T}^1)).$$

5.4 Factual Overlap Score

To symmetrically evaluate if two texts contain the same contents, we compute the Factual Overlap Score (**FO**) between $\mathbf{T}^1$ and $\mathbf{T}^2$, as follows:

$$\mathbf{FO}(\mathbf{T}^1, \mathbf{T}^2) = \frac{1}{Z_D}(\mathbf{FI}_A(\mathbf{T}^1 \subset \mathbf{T}^2) + \mathbf{FI}_D(\mathbf{T}^1 \subset \mathbf{T}^2) + \mathbf{FI}_A(\mathbf{T}^2 \subset \mathbf{T}^1) + \mathbf{FI}_D(\mathbf{T}^2 \subset \mathbf{T}^1))$$

Z_D is a normalisation factor that scales the result based on the total number of faithful atomic fact and discourse relation extractions from both texts:

$$Z_D = \sum_{i=1}^{m} \delta_\theta(C(A_i^1|\mathbf{T}^1)) + \sum_{j=1}^{n} \delta_\theta(C(A_j^2|\mathbf{T}^2)) + \sum_{k=1}^{o} \delta_\theta(C(D_k^1|\mathbf{T}^1)) + \sum_{l=1}^{p} \delta_\theta(C(D_l^2|\mathbf{T}^2)).$$

6 FDSpotter: Atomic Fact and Discourse Relation Entailment from the Text

To obtain the Factual Inclusion and Factual Overlap scores in Section 5, we train an entailment model, FDSpotter, to compute the probability of whether a structured atomic fact A_i or a discourse relation D_j is expressed in the text $\mathbf{T}$. We collect atomic fact entailment and discourse relation entailment data for training and testing.

6.1 Atomic Fact Entailment Data

To compute the probability of a structured atomic fact A_i being present in a given text $\mathbf{T}$, the model is trained on two groups of datasets.

1. Derived from FactSpotter: FactSpotter (Zhang et al., 2023) is a model for evaluating if an atomic fact from a knowledge graph (KG) is stated in a text generated from the KG. Its training data were generated using a self-supervised method. The positive samples were taken directly from datasets with pairs of atomic facts and their corresponding ground-truth texts, such as WebNLG (Castro Ferreira et al., 2020) or GrailQA (Gu et al., 2021). The negative samples were generated by perturbing the atomic facts or the corresponding descriptive texts, such that the atomic fact can no longer be entailed from the text. We use the same method to generate the training and testing samples from the WebNLG dataset.

2. Derived from Text Entailment Data: Text entailment corpus contain hypothesis-premise pairs and aim to determine if each hypothesis is entailed, contradictory, or neutral for the given premise text. The task is also known as Natural Language Inference (NLI). The use of NLI data is to introduce a degree of semantic flexibility in the training signal and to mitigate false negatives, cases where atomic facts are inferable across longer or more implicit spans of text but are not explicitly stated. We used the spaCy (Honnibal and Montani, 2017) model to extract the subject, predicate, object, adverbial and complement from a hypothesis, and add delimiters between them to structure the text into an atomic fact.
We generate pairs (natural language premises, hypotheses represented as structured atomic facts) from the following datasets: SNLI (MacCartney and Manning, 2008), MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018), ANLI (Nie et al., 2020), LingNLI (Parrish et al., 2021), WANLI (Liu et al., 2022), and CNC (Tan et al., 2022).

6.2 Discourse Relation Entailment Data

The models trained on the existing entailment datasets are good at identifying whether an atomic fact in the discourse relation is stated in the premise text. However, when classifying whether a discourse relation as a hypothesis is represented in the

premise, the accuracy decreases, especially when the discourse connective is adversarial.

For instance, consider the text: "Germany can be the birthplace of Heidegger, Hegel, Leibniz, Bach, Beethoven, Brecht, and Martin Luther, but they started one world war". The discourse relation ⟨⟨Germany, is the birthplace of, many great minds⟩, so, ⟨Germany, started, the world war⟩⟩ is not entailed from this text: both atomic facts are in the text, but the its connective "so" means the opposite to the connective "but" in the text. However, the models trained without adversarial samples consider this relation to be entailed from the premise.

To enable the classification of adversarial discourse relations from the premise, the model needs to be trained on data with pairs of texts and discourse relations, but especially include discourse connectives contradicting the ones from the premise. To the best of our knowledge, no such dataset exists. We leverage the NLI datasets (MNLI, ANLI, LingNLI, and WANLI) with large quantity of discourse connectives in the hypotheses to generate our **DiscInfer (i.e., Discourse Relation Inference)** dataset by the following steps.

1. *Sample Selection.* Automatically select pairs of hypotheses and premises from NLI data. Each hypothesis should contain at least one type of discourse connective in Section 5.4.

2. *Connective Replacing.* Replace one discourse connective in the hypothesis of the selected sample with another discourse connective that would make the hypothesis potentially contradict the premise, e.g., replacing "if" with "unless". Potential contradictory discourse connective pairs are listed in the Appendix B.

3. *Human Verification.* Human annotators check whether the relation between each new hypothesis-premise pair is entailing, neutral, or contradicting. The annotators correct the labels and add the created pairs to the dataset. The verification is due to changing the discourse connective in the hypothesis does not necessarily contradict the premise[3] and the new hypothesis is sometimes not coherent.

The training split and the test split of DiscInfer are both created from the training and testing splits of MNLI, LingNLI, ANLI, and WANLI. The dataset has 920 annotated (hypothesis, premise) pairs.

[3]For example, the meaning of the discourse relation ⟨⟨He, did not finish, homework⟩, when, ⟨the class, started⟩⟩ is equivalent to ⟨⟨He, did not finish, homework⟩, before, ⟨the class, started⟩⟩ after replacing the discourse connective from "when" to be "before".

6.3 Training and Evaluation

We fine-tune 304M DeBERTa V3 Large (He et al., 2021), pretrained on the tasksource dataset (Sileo, 2024). The model is finetuned on three parts of data: (1) original text entailment; (2) synthetic atomic fact entailment; (3) DiscInfer. The model is finetuned on NVIDIA Tesla V100 with learning rate 1e-5, batch size 16, and AdamW Optimiser (Loshchilov and Hutter, 2019) for 3 epochs.

Model/Split	Temp.	Cont.	Comp.
DeBERTa w/o DiscInfer	58.2	48.3	23.1
DeBERTa w DiscInfer	**76.4**	**84.3**	**59.6**
GPT4 zero-shot	73.0	53.8	48.1

Table 3: Accuracy of DeBERTa and GPT4 on DiscInfer.

In Table 3, we present the results on the test set of DiscInfer of DeBERTa and GPT4. For DeBERTa, we have one checkpoint that did not see the training split of DiscInfer, and another checkpoint finetuned along with DiscInfer. We observe that DeBERTa trained without DiscInfer does not perform well on the test split of discourse relation entailment, and there is a remarkable improvement after finetuning on DiscInfer. We have also tested zero-shot GPT4 performance on DiscInfer, with the prompt asking whether the hypotheses are entailed, neutral, or contradictory to the premises. GPT4 performs better than (much smaller) DeBERTa trained without DiscInfer, but underperforms compared with DeBERTa finetuned on DiscInfer.

7 Applications on Generation Evaluation

As explained in Section 5.4, the Factual Overlap Score is a symmetric metric to measure the content alignments between two texts, such as the ground-truth reference and the generated output. It symmetrically aggregates the Factual Inclusion Scores in both directions, evaluating whether both texts exactly contain the same atomic facts and discourse relations. This makes it suitable for the general text generation evaluation, e.g., graph-to-text, where bidirectional content equivalence is expected.

Reference-free faithfulness evaluation of some tasks requires an asymmetric approach, for which the Factual Inclusion in Section 5.3 can be applied in a single direction. For example, for graph-to-text generation faithfulness, we can use Factual Inclusion to determine whether all the triples $\mathbf{F} =$

$\{F_1, \ldots, F_m\}$, which can be considered as atomic facts, are correctly verbalised in the text $\mathbf{T}$, i.e.,

$$\mathbf{FI}_A(\mathbf{F} \subset \mathbf{T}) = \sum_{i=1}^{m} C(F_i|\mathbf{T}).$$

For text summary faithfulness classification, the generated summaries should be faithful to the input, but do not need to cover all of the input document. It is reflected in the consistency measurement of text summary benchmarks. To classify whether all the atomic facts and discourse relations in the generated summary $\mathbf{S}$ are included by the input document $\mathbf{I}$, we modify Factual Inclusion with the sum of log probabilities, i.e.,

$$\mathbf{FI}_A(\mathbf{S} \subset \mathbf{I}) = \sum_{i=1}^{m} \delta_\theta(C(A_i|\mathbf{S})) \log C(A_i|\mathbf{I})$$

$$\mathbf{FI}_D(\mathbf{S} \subset \mathbf{I}) = \sum_{j=1}^{o} \delta_\theta(C(D_j|\mathbf{S})) \log C(D_j|\mathbf{I})$$

To compare with existing metrics, we compute *system-level* and *sample-level* correlations between the automatic metrics and human judgements. We report three correlation coefficients: Spearman ρ, Pearson r, and Kendall's Tau τ.

7.1 Natural Language Generation

WebNLG (Castro Ferreira et al., 2020) is a graph-to-text generation benchmark (i.e. triples or sets of triples of the form ⟨`subject, predicate, object`⟩ are paired with their equivalent texts) on which 16 systems have been evaluated by humans on data coverage (whether the text includes descriptions of all predicates in the data), correctness (whether the predicates in the data are correctly mentioned with the subject and object), and relevance (whether the text describes only the predicates in the data with related subjects and objects). Fluency and grammar evaluations were also reported by annotators, but they are not related to factual consistency. We report the performance of existing and our proposed metrics in Table 4, where the Factual Overlap and Factual Inclusion have the highest correlations on the dimensions related to factual consistency.

7.2 Text Summarisation

SummEval (Fabbri et al., 2020) benchmark has summaries from CNN and Daily Mail generated by 23 models with human judgements. We evaluate on the dimension of consistency, which reflects the factual alignment between the summary and the input, where the hallucinations are penalised by annotators. We compute the correlations of Factual Inclusion with human annotated consistency, and compare against BERTScore, BARTScore, as well as state-of-the-art methods such as FENICE (Scirè et al., 2024), ChatGPT-DA (Wan et al., 2023) and G-EVAL-4 (Liu et al., 2023b). We observe in Table 5 that Factual Inclusion has a strong performance.

AggreFact (Tang et al., 2023) contains summaries from CNN, Daily Mail and XSUM (Narayan et al., 2018). It focuses on factual errors in text summarisation and gives summaries a binary label: factual or non-factual. It evaluated three groups of models and pairs them with human evaluations: FTSOTA, EXFORMER, and OLD. FTSOTA has more recent models on which factuality metrics struggle. We compare with the following competitors: DAE (Goyal and Durrett, 2020), QuestEval (Scialom et al., 2021), SummaC (Laban et al., 2022), QAFactEval (Fabbri et al., 2022), TrueTeacher (Gekhman et al., 2023), MENLI (Chen and Eger, 2023), AlignScore (Zha et al., 2023), ChatGPT-ZS, ChatGPT-CoT (Luo et al., 2023), ChatGPT-DA, ChatGPT-Star, and FENICE. These metrics have been proposed as competitors in Scirè et al. (2024). We present in Table 6 the balanced accuracy on FTSOTA, and in the Appendix C on all the splits. We observe that our proposed Factual Inclusion method performs on par with the top-performing methods.

To evaluate the performance on diverse long-form summarisation faithfulness classification, we conduct experiments on DiverSumm (Zhang et al., 2024), a comprehensive benchmark that incorporates faithfulness annotations across five distinct domains: ChemSumm (CSM), QMSUM (QMS), ArXiv (AXV), GovReport (GOV), and MultiNews (MNW). DiverSumm presents unique challenges for faithfulness evaluation as it contains longer documents and summaries. We compare our proposed Factual Inclusion against several state-of-the-art faithfulness evaluation methods, including Full-Doc, SummaC, SentLI (Schuster et al., 2022), and INFUSE (Zhang et al., 2024). Following Laban et al. (2022) and Zhang et al. (2024), we report ROC-AUC (Bradley, 1997) scores for faithfulness classification across all DiverSumm subsets. As shown in Table 7, Factual Inclusion achieves the highest average performance, substantially outperforming the previous state-of-the-art methods with consistent improvements across most domains.

	Text-level									System-level								
	Correctness			D. Coverage			Relevance			Correctness			D. Coverage			Relevance		
Metric	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
Human	67.2	57.3	45.3	68.3	61.2	48.6	65.1	50.2	40.8	96.0	80.0	65.0	93.0	83.0	68.0	96.0	74.0	59.0
BLEU	60.8	57.1	42.5	56.9	60.0	44.5	58.1	48.2	35.4	59.6	64.4	48.5	54.0	53.6	40.5	57.0	60.0	45.2
METEOR	65.2	59.5	42.9	66.2	57.1	43.3	68.1	55.2	41.8	72.8	75.7	60.3	65.7	58.3	45.0	70.5	64.7	50.0
BERTScore	65.3	60.8	46.6	67.1	61.4	46.5	66.1	58.0	43.7	83.1	77.7	60.5	74.8	58.3	43.7	81.0	65.9	50.8
BARTScore	71.5	59.6	45.4	69.6	61.9	46.9	70.4	61.8	47.1	90.6	**83.2**	**67.6**	87.0	71.5	53.4	88.7	71.7	56.8
BLEURT	72.0	63.3	48.0	68.8	60.5	46.0	71.7	**63.3**	**47.6**	93.1	82.9	**67.6**	87.0	65.6	50.6	91.1	70.0	55.8
FactSpotter	68.9	59.0	45.0	71.2	**64.0**	48.8	69.8	59.4	45.8	94.7	80.2	64.2	91.4	87.1	71.5	96.2	**80.0**	**64.9**
FactInclusion	70.7	63.1	48.7	72.3	63.6	**49.0**	**74.3**	61.1	46.9	**97.3**	81.6	67.4	96.8	**92.8**	81.1	**96.6**	79.5	64.7
FactOverlap	**74.2**	**65.1**	**50.3**	**72.9**	62.7	48.5	73.9	60.1	46.2	96.9	78.6	64.0	**97.2**	**92.8**	**81.6**	95.4	74.3	59.2

Table 4: Text-level and System-level correlations between metrics and human annotations of factuality dimensions on WebNLG2020. We **highlight** the best result and underline the other top 3 values in tables here and below.

	System-level			Sample-level		
	r	ρ	τ	r	ρ	τ
BERTScore	17.9	-7.6	-3.3	10.9	15.2	9.0
BARTScore	81.0	77.4	60.0	40.1	48.7	33.2
G-EVAL-4	92.3	76.5	60.0	**50.1**	59.1	**42.0**
ChatGPT-DA	95.0	81.2	68.3	41.9	51.7	38.9
FENICE-GPT	91.8	82.4	68.3	43.5	**65.9**	36.0
FactInclusion	**96.5**	**85.9**	**70.0**	40.8	64.5	34.0

Table 5: System-level and sample-level correlations between metrics and human annotations of consistency on SummEval benchmark.

Metric	CNN/DM	XSUM	Average
DAE	65.4±4.4	70.2±2.3	67.8
QuestEval	70.2±3.2	59.5±2.7	64.9
SummaC-ZS	64.0±3.8	56.4±1.2	60.2
SummaC-Conv	61.0±3.9	65.0±2.2	63.0
QAFactEval	67.8±4.1	63.9±2.4	65.9
TrueTeacher	62.0±1.3	**74.9±1.2**	68.5
MENLI	65.0±2.8	57.0±1.8	61.0
AlignScore	67.0±3.1	60.3±1.9	63.7
ChatGPT-ZS	56.3±2.9	62.7±1.7	59.5
ChatGPT-CoT	52.5±3.3	55.9±2.1	54.2
ChatGPT-DA	53.7±3.5	54.9±1.9	54.3
ChatGPT-Star	56.3±3.1	57.8±0.2	57.1
FENICE-GPT	**70.5**±1.6	72.8±0.3	71.6
FactInclusion	69.6±3.4	73.8±1.7	**71.7**

Table 6: Balanced accuracy of on FTSOTA split of AggreFact, with single-threshold setting.

System	CSM	QMS	AXV	GOV	MNW	AVG
FullDoc	50.15	37.12	62.78	79.19	44.76	54.80
SummaC-Conv	53.14	51.13	61.22	65.34	53.05	56.78
SummaC-ZS	54.41	48.21	69.44	79.37	50.17	60.32
SentLI	50.13	47.56	64.49	79.68	46.61	57.69
INFUSE	54.11	52.16	**71.38**	80.45	**53.16**	62.25
FactInclusion	**56.79**	**70.30**	70.56	**83.44**	44.88	**65.20**

Table 7: ROC-AUC scores on DiverSumm benchmark for long form summary faithfulness evaluation.

8 Conclusion

We explore a novel method for computing factual consistency in this work. The work is based on the literature of structured discourse information representation and we propose to incorporate discourse relations between atomic facts in structured information representation for factual consistency verification. We validate this approach on data-to-text generation and text summarisation evaluation benchmarks, and our method achieved competitive performance when compared with a variety of other approaches. We also introduced a text entailment dataset with adversarial discourse connectives, i.e., DiscInfer, which is challenging for the language models trained with existing data. While factual consistency verification is a rapidly evolving field, we believe it is important to leverage fundamental linguistic knowledge so that the advances remain explainable and hence more trustworthy.

Acknoledgement.

The authors were partially funded by the ANR-20-CHIA-0015 project and the ANRT. This work was performed using HPC resources from GENCI-IDRIS (Grant AD011014244R1). The authors thank Huajian Zhang and Yang Zhong for baseline reproduction. We are also grateful to Claire Gardent and Simon Razniewski for their valuable suggestion on improving the paper. Kun Zhang thanks his high school teachers for the solid instruction in English grammar and linguistics.

9 Limitations

Our work has the following limitations:

- We have tested our method only on English text, and in particular the structured sentence representation is based on English syntax. We believe it is very important to address this task in other language. However, due to time and knowledge limitations we could not include it in our work.

- We took advantage of NLI data to train the classifier. However, the hypotheses of NLI may include facts which are semantically plausible but not explicitly grounded in the premise. There exists a conceptual gap between entailment and strict fact inclusion. Training with such data is a trade-off between minimising false negatives, where legitimate facts are mistakenly excluded due to the lack of exact match, and avoiding false positives, where inferred but unstated facts are accepted.

- While our metric is explainable in the sense we can obtain very fine-grained information on how a factuality score was given, nowadays explanations in textual format given by LLMs have gained a lot of attention; we believe that this is a future step that is attainable with current techniques.

- We are currently completing a comprehensive error analysis for our technique and for competitors. Understanding to what type of errors each method is sensitive can be useful for further improving the SOTA. Another question that previous work have tried to answer using automated techniques and human annotation (Zhang et al., 2023) is if the benchmarks are still challenging or if the remaining model errors are actually due to human errors in annotations or divergence of annotations.

References

Brent H Baker, Tim Graham, and Steve Kaminsky. 1994. *How to identify, expose & correct liberal media bias*. Media Research Center Alexandria, VA.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Andrew P. Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. 2018. A content management perspective on fact-checking. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 565–574, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Huiyao Chen, Meishan Zhang, Jing Li, Min Zhang, Lilja Øvrelid, Jan Hajič, and Hao Fei. 2025a. Semantic role labeling: A systematical survey.

Huiyao Chen, Meishan Zhang, Jing Li, Min Zhang, Lilja Øvrelid, Jan Hajič, and Hao Fei. 2025b. Semantic role labeling: A systematical survey.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366.

Kuicai Dong, Aixin Sun, Jung-jae Kim, and Xiaoli Li. 2023. Open information extraction via chunks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15390–15404, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey et al. 2024. The llama 3 herd of models.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, Richard Socher, and Dragomir R. Radev. 2020. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 166–175, New York, NY, USA. Association for Computing Machinery.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond i.i.d.: Three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, WWW '21, page 3477–3488, New York, NY, USA. Association for Computing Machinery.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2017. Discourse structure in machine translation evaluation. *Computational Linguistics*, 43(4):683–722.

René Knaebel. 2021. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Young-Suk Lee, Ramón Astudillo, Hoang Thanh Lam, Tahira Naseem, Radu Florian, and Salim Roukos. 2022. Maximum Bayes Smatch ensemble distillation for AMR parsing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5379–5392, Seattle, United States. Association for Computational Linguistics.

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

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei Liu, Yi Fan, and Michael Strube. 2023a. HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. *arXiv preprint arXiv:2303.15621*.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The Association for Computational Linguistics.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Marcel Nawrath, Agnieszka Nowak, Tristan Ratz, Danilo Walenta, Juri Opitz, Leonardo Ribeiro, João Sedoc, Daniel Deutsch, Simon Mille, Yixin Liu, Sebastian Gehrmann, Lining Zhang, Saad Mahamood, Miruna Clinciu, Khyathi Chandu, and Yufang Hou. 2024. On the role of summary content units in text summarization evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 272–281, Mexico City, Mexico. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stephan Oepen, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfred Stede, Erik Velldal, and Lilja Øvrelid. 2016. OPT: Oslo–Potsdam–Teesside. pipelining rules, rankers, and classifier ensembles for shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 20–26, Berlin, Germany. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, and Sandhini Agarwal et al. 2024. Gpt-4 technical report.

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

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*, pages 311–318.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. *Penn Discourse Treebank Version 3.0*. Abacus Data Network.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Alix Rule, Jean-Philippe Cointet, and Peter S Bearman. 2015. Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112:10837 – 10844.

Zacchary Sadeddine, Juri Opitz, and Fabian Suchanek. 2024. A survey of meaning representations – from theory to practical utility. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2877–2892, Mexico City, Mexico. Association for Computational Linguistics.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14148–14161, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.

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, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Fabian M. Suchanek. 2020. The need to move beyond triples. In *Text2Story@ECIR*.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *J. ACM*, 21(1):168–173.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24, Beijing, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ruth Wodak and Michael Meyer. 2015. Methods of critical discourse studies.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *Frontiers Comput. Sci.*, 18:186357.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

Zinong Yang, Feng Xu, Jianfei Yu, and Rui Xia. 2023. UniCOQE: Unified comparative opinion quintuple extraction as a set. In *Findings of the Association for*

Computational Linguistics: ACL 2023, pages 12229–12240, Toronto, Canada. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

Kun Zhang, Oana Balalau, and Ioana Manolescu. 2023. FactSpotter: Evaluating the Factual Faithfulness of Graph-to-Text Generation. In *Findings of EMNLP 2023 - Conference on Empirical Methods in Natural Language Processing*, Singapore, Singapore.

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

Jinman Zhao and Gerald Penn. 2024. LLM-supertagger: Categorial grammar supertagging via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 697–705, Miami, Florida, USA. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Yang Zhong and Diane Litman. 2025. Discourse-driven evaluation: Unveiling factual inconsistency in long document summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2050–2073, Albuquerque, New Mexico. Association for Computational Linguistics.

A Grammatical form of Discourse Connectives

Syntactically, the discourse connectives are frequently expressed by conjunctives, relative adverbs or other transitional phrases. Below are several representative grammatical forms and their corresponding structured representation.

- Two clauses connected by a conjunction. For example, the text *"Amy and Kate were both in good mood, so they went to celebrate happily"* is structured to be ⟨⟨`Amy and Kate, were both, in good mood`⟩, `so`, ⟨`Amy and Kate, went to celebrate, happily`⟩⟩.

- Two clauses in different sentences connected by a transitional word or phrase. For instance, the text *"Amy and Kate celebrated until midnight. As a result, they got up late on the next day"* is represented as ⟨⟨`Amy and Kate, celebrated, until midnight`⟩, `as a result`, ⟨`Amy and Kate, got up, late on the next day`⟩⟩.

- Relative adverbs (e.g., when, why) leading adverbial clauses. For example, the text *"The company built a new school where there was a park"* is represented as ⟨⟨`the company, built, a new school`⟩, `where`, ⟨`there, was, a park`⟩⟩.

- Complex phrases connecting adverbial clauses. For instance, *"Despite the fact that he cheated on the exam, he was not dismissed by the university"* is represented as ⟨⟨`he, was not dismissed, by the university`⟩, `despite the fact that`, ⟨`he, cheated, on the exam`⟩⟩.

B Creation of the DiscInfer Dataset

To facilitate the construction of the DiscInfer dataset for training and evaluating models on identifying contradictions induced by adversarial discourse connectives, we automatically generated adversarial examples by replacing one discourse connective in the selected hypothesis with another discourse connective that could potentially render the hypothesis contradictory to the premise. The groups of potentially conflicting discourse relations are listed in Table 8.

To ensure grammatical plausibility, a discourse connective is replaced only if the original and replaced connectives belong to the same grammatical category, such as subordinating conjunctions or adverbial phrases, to preserve syntactic coherence.

The annotators were trained on the linguistic concepts related to discourse connectives (e.g., conjunctions, transitional phrases, adverbial clauses, etc.). They were also provided with a reference

Original connective	Potential conflicting connectives
concession	condition; reason; result.
contrast	reason; result; similarity.
similarity	concession; contrast.
negative condition	condition; reason; result.
condition	concession; negative condition; result.
reason	concession; negative condition; precedence; result.
result	concession; condition; contrast; reason; succession.
succession	precedence; result; synchronous.
synchronous	precedence; succession.
precedence	concession; reason; succession; synchronous.

Table 8: Potential conflicting discourse relation groups.

list of connective groups and examples of potentially conflicting discourse relations. All annotators have completed postgraduate-level education and possess at least a C1 level of English proficiency.

For completeness, we also list examples of discourse connectives of each relation category as follows. These examples are primarily drawn from the Penn Discourse Treebank (Prasad et al., 2019).

- alternative: alternatively, as an alternative, instead of.
- concession: although, in spite of, even though, though, despite, even if, albeit, regardless of.
- condition: depending on, depending upon, if, provided that, in case.
- contrast: by/in contrast, conversely, however, on the contrary, on the other hand, rather than, whereas, by/in comparison, but.
- disjunctive: otherwise, unless, except.
- precedence: thereafter, later, then, before, subsequently, afterwards, afterward.
- reason: as a result of, because, because of, due to, as a consequence of.
- result: consequently, therefore, thus, hence, thereby, so, so that, for the purpose of, with the purpose of, in order to, as a result, as a consequence, for this/that reason.
- similarity: similarly, in a similar way, in the same way, likewise.
- succession: earlier, previously, until, till, after.
- synchronous: when, during, at the same time, meanwhile, simultaneously, meantime.

To ensure balanced coverage across discourse relation categories, we aimed to include at least 100 training examples and 50 test examples per high-level group (comparison, contingency, temporal). However, due to differences in connective frequency in NLI datasets, the final distribution in the dataset varied. Table 10 illustrates the category distribution of Discinfer training and test samples.

C Detailed result on AggreFact

Table 9 reports balanced accuracy on AggreFact, with threshold-per-split settings. The accuracy is computed with the same method as Tang et al., and we report 95% confidence intervals for our method and the baselines with publicly available outputs. Following Scirè et al., the performance of DAE is excluded on the EXF (Exformer) and OLD splits of XSum because its training data (Goyal and Durrett, 2021) covers these splits.

We compare the effectiveness of structured representations for factual consistency verification. Open Information Extraction (Open IE) systems typically extract subject-predicate-object (SPO) triples. However, the triple representations have semantic incompleteness, as they omit adverbials and complements. We propose to enhance atomic facts by including adverbials and complements, and to incorporate discourse relations to represent connections between facts. As shown in Table 11, using atomic facts outperforms the triple format by including more complete content units. Also, incorporating discourse relations further improves the classification performance. We further assess the effect of training with DiscInfer dataset, which introduces adversarial discourse connectives. As is shown in Table 12, DiscInfer improves the ability of models for factual consistency verification.

D Complexity Comparison for Factual Consistency Verification Methods

We compare the computational complexity of the approaches for factual consistency verification. We categorize existing methods into three groups.

1. Simple Text Entailment. These methods use a classification model (e.g., DeBERTa (He et al., 2021) for NLI) once per instance, taking one text (e.g. text summarisation output) as the hypothesis and another text (e.g., the input document of text summarisation) as the premise.

	AggreFact-CNN				AggreFact-XSum			
	FTS	EXF	OLD	AVG	FTS	EXF	OLD	AVG
DAE	59.4±3.1	67.9±3.1	69.7±1.1	65.6	73.1±1.8	-	-	-
QuestEval	63.7±3.2	64.3±3.1	65.2±1.1	64.4	61.6±1.9	60.1±4.5	59.7±3.4	60.5
SummaC-ZS	63.3±3.0	**76.5**±2.8	76.3±1.0	72.0	56.1±2.0	51.4±4.6	53.3±3.8	53.6
SummaC-Conv	**70.3±2.5**	69.8±3.2	78.9±1.1	**73.0**	67.0±2.0	64.6±4.3	67.5±4.1	66.4
QAFactEval	61.6±3.6	69.1±3.0	80.3±1.0	70.3	65.9±2.0	59.6±4.7	60.5±5.4	62.0
TrueTeacher-11B	65.7	57.7	81.9	68.4	75.2	68.4	52.8	65.5
MENLI	51.7	52.8	68.4	57.6	58.3	59.7	**73.9**	64.0
AlignScore	53.5	73.9	78.0	68.5	**80.2**	**79.9**	63.7	**74.6**
ChatGPT-ZS	66.2	64.5	74.3	68.3	62.6	69.2	60.1	64.0
ChatGPT-CoT	49.7	60.4	66.7	59.0	56.0	60.9	50.1	55.7
ChatGPT-DA	48.0	63.6	71.0	60.9	53.6	65.6	61.5	60.2
ChatGPT-Star	55.8	65.8	71.2	64.3	57.7	70.6	53.8	60.7
FENICE$_{gpt}$	68.2	68.8	**82.1**	**73.0**	73.9	73.5	69.9	72.4
FactOverlap$_{gpt}$	63.7±3.3	69.6±3.0	79.0±1.0	70.7	74.7±1.8	73.4±3.7	73.0±3.6	73.7

Table 9: Balanced accuracy on the test splits of AggreFact, with threshold-per-split settings.

Category	Size	Details
Train		
Temporal	186	Succession: 55, Precedence: 72, Synchronous: 59
Contingency	237	Condition: 100, Reason: 68, Result: 69
Comparison	117	Concession: 62, Contrast: 55
Test		
Temporal	123	Succession: 49, Precedence: 37, Synchronous: 37
Contingency	248	Condition: 136, Reason: 85, Result: 19
Comparison	52	Concession: 34, Contrast: 18

Table 10: Distribution of DiscInfer training and test sets.

Extraction Format/Split	CNN/DM	XSUM
Subject-Predicate-Object Only	65.4	69.6
Atomic Facts Only	66.4	72.4
Atomic Facts + Discourse Relation	**69.6**	**73.8**

Table 11: Balanced accuracy of different extraction formats on AggreFact.

2. Question Answering. Methods such as such as QAFactEval (Fabbri et al., 2022)) and QuestEval (Scialom et al., 2021) generate multiple questions about the text. They require one model to propose questions and another to answer the questions. The total time cost for evaluating a single text is:

$$T_{QA} = T_{\text{propose}} + N_{\text{questions}} \times T_{\text{answering}}.$$

Some LLM-based methods (Wan et al., 2023) request a single factual consistency score from the LLM, involving only one sequence-to-sequence model invocation. However, the models based on Chain-of-Thoughts (Luo et al., 2023) still require question answering of multiple rounds.

Checkpoint	CNN-DM	XSUM
w/o DiscInfer	62.6	72.0
w/ DiscInfer	**69.6**	**73.8**

Table 12: Effect of DiscInfer on performance.

3. Information Extraction + Entailment These methods first extract content units from the input by semantic parsing (Goyal and Durrett, 2020), sentence tokenisers (Chen and Eger, 2023), or LLM atomic fact extractors (Scirè et al., 2024), and then they check the factual consistency of the extracted content units. Their computational complexity is:

$$T_{IE+E} = T_{\text{extraction}} + N_{\text{content units}} \times T_{\text{entailment}}.$$

The complexity of the content unit extractors varies: sentence tokenisers and semantic parsers are lightweight, while both FENICE (Scirè et al., 2024) and our method use LLMs. Moreover, the number of extracted units also varies, typically,

$$N_{\text{sentences}} \leq N_{\text{atomic facts}}.$$

If a classification model is introduced to score all top extractions for selecting the most faithful extraction, this would incur additional overhead:

$$N_{\text{extractions}} \times (N_{\text{content units}}) \times T_{\text{entailment}}.$$

Our proposed Factual Inclusion Score (**FI**) belongs to the third category, which combines atomic

facts with discourse relations as the content units. Therefore, the number of the content unit becomes:

$$N_{\text{content\_units}} = (N_{\text{atomic facts}} + N_{\text{discourse relations}}).$$

Then the overall complexity changes according to the number of atomic facts and discourse relations. Given the number of discourse relations is usually smaller than the atomic facts, the runtime overhead of our method is marginally higher in this step.

E Extraction evaluation

We present the prompt we used for the few shot extraction of atomic facts and discourse relations in Figure 1. In Tables 13 and 14 we present the results of the extraction using different models and temperatures. We use a 16 bit quantified Llama 3.1 8B model and 4 bit quantified 70B model. We observe that when we change the temperature, there is a fluctuation of two to three percent between the results, with no temperature being generally better across all models. For Llama 3.1 7B model with temperature 0.2 and 0.3 we observe a significant drop in the recall for atomic fact extraction. We verified the extraction results, and Llama 3.1 8B indeed extracts less atomic facts under these two temperatures, while the extraction of the discourse connectives is not affected. This highlights the complex problem of finding the best parameters.

```
Task
1. Extract all atomic facts from the given text and list them in a markdown table with columns: subject,
predicate, object direct, object indirect, short adverbial and complement.

2. Extract relations between the events, with columns: discourse1, connective, discourse2. The relations
are frequently connected by the following groups of connectives: concession (eg, although, even if); reason
(eg, due to, because); result (eg, consequently, therefore); contrast (eg., in contrast, however);
disjunctive (eg, otherwise, unless); alternative (eg., instead, as an alternative); condition (eg.,
provided that, in case); synchronous (eg, when, at the same time); precedence (eg, before, until);
succession (eg, after, subsequently); similarity (eg, similarly, in the same way); location (eg, where).
The connectives shouldn't be limited by the examples.

Example text: Being thankful about the exam's results, person A gave a gift to her best friend, person B,
last week. Person A would have failed, if person B were not helping. Person B said that her best friend did
not need to do this, even if person B helped person A revise all the course materials from time to time.
They held a party until midnight. As a result, they didn't get up early on the next day.

Expected output Format:
# Atomic Facts
| Subject | Predicate | Object Direct | Object Indirect | Short Adverbial and Complement |
| - | - | - | - | - |
| person A | gave | a gift | to person A's best friend, person B | - | last week |
| person A | was thankful about | the exam's results | - | - | - |
| person B | is best friend of | person A | - | - | - |
| person B | said | her best friend did not need to give a gift to person B | - | - | - |
| person B | helped | person A | - | from time to time | to revise all the course materials |
| person A and person B | held | a party | - | until midnight |
| person A and person B | didn't get up | -  | - | early on the next day |

# Discourse Relations
| Event1 | Connective | Event2 |
| - | - | - |
| person A gave a gift to person A's best friend, person B | because of | being thankful about the exam's
results |
| person A would have failed | if | person B were not helping |
| person B said person A did not need to give a gift to person B | even if | person B helped person A
revise all the course materials from time to time |
| person A and person B held a party until midnight | As a result | person A and person B didn't get up
early on the next day |


# Standards

Having correct tense for predicate is important.

Do not miss any detailed information, especially features, or identities of the entities.

Do not extract content that is not present in the input.

Perform co-reference resolution on entities: if an entity has a full name, put the full name, rather than a
pronoun.

If an auxiliary verb refers to another predicate, put their original predicate. For example, "did not need to
do this" in the original text is extracted to be "did not need to give a gift to person B".

Make sure the predicate links subject and object in a correct direction. For example, we only know "person B
| helped | person A" from the text, rather than "person A | helped | person B".

If the subject or object has another atomic fact in its adjective, clause, non-finite verb, or appositive,
place all related words in the corresponding column and the new fact should also be in another line of the
table. For example, "person B | is best friend of | person A" is a new fact inside the appositive "person A's
best friend, person B".

If two discourses are linked with a connective, connect them in the discourse relation table. For example,
the events "person A | gave | a gift | to person A's best friend, person B | last week" and "person A | was
thankful about | the exam's results" are connected with "because of". The discourse connective shouldn't
appear in "Short Adverbial and Complement".

If a discourse is hypothetical or counterfactual situation, such as subjunctive mood, output the discourse
relation but don't extract it as an atomic fact. For example, "person B were not helping" and "person A would
have failed" are not atomic facts, but discourse relation should contain "person A would have failed | if |
person B were not helping".

Don't output atomic facts without a predicate.

Discourse connectives should correctly lead another event. Example: "As a result" is leading event the
content "person A and person B didn't get up early the next day", rather than anything else.

Don't generate discourse relation table when there is no discourse relations.

Don't output the following connectives: simple conjunction (eg, and, in addition); specification (eg,
especially, in particular); restatement (eg, in other words); instantiation (eg, for example, for instance);
generalisation (eg, in summary, overall).

Don't output extra explanations.
```

Figure 1: The prompt for atomic fact and discourse relation extraction.

	Strict			**LCS**			**SBERT Elem**			**SBERT Full**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Human	77.21	69.17	72.97	81.45	72.97	76.98	86.20	77.22	81.46	90.15	80.76	85.20
GPT4-Turbo t=0	69.45	53.06	60.16	75.05	57.33	65.00	49.85	38.08	43.18	89.29	68.21	77.34
GPT4-Turbo t=0.1	69.11	53.75	60.47	75.41	58.65	65.99	80.89	62.91	70.78	88.71	69.00	77.63
GPT4-Turbo t=0.2	66.49	51.25	57.88	72.66	56.01	63.26	79.23	61.07	68.97	90.35	69.65	78.66
GPT4-Turbo t=0.3	68.47	52.78	59.61	73.84	56.92	64.29	79.36	61.17	69.09	88.76	68.42	77.27
GPT4o t=0	70.00	53.47	60.63	75.87	57.96	65.71	81.32	62.12	70.44	91.00	69.52	78.82
GPT4o t=0.1	68.99	52.22	59.45	75.22	56.94	64.81	81.41	61.62	70.15	88.57	67.04	76.32
GPT4o t=0.2	69.64	53.19	60.31	75.87	57.96	65.71	82.13	62.74	71.14	89.11	68.07	77.19
GPT4o t=0.3	68.54	49.03	57.17	75.89	54.28	63.29	82.68	59.14	68.95	92.61	66.24	77.24
LLAMA3 8B t=0	49.61	43.75	46.49	57.72	50.91	54.10	65.99	58.20	61.85	75.05	66.19	70.34
LLAMA3 8B t=0.1	50.81	43.75	47.01	58.79	50.62	54.40	67.60	58.21	62.56	77.48	66.72	71.69
LLAMA3 8B t=0.2	52.94	25.00	33.96	59.44	28.07	38.13	68.56	32.37	43.98	73.34	34.63	47.05
LLAMA3 8B t=0.3	49.54	22.36	30.81	57.32	25.87	35.65	68.79	31.05	42.79	76.40	34.49	47.52
LLAMA3 70B t=0	57.80	50.97	54.17	63.41	55.92	59.43	70.21	61.92	65.81	82.71	72.94	77.52
LLAMA3 70B t=0.1	55.85	50.42	52.99	62.07	56.03	58.90	69.35	62.61	65.81	80.24	72.44	76.14
LLAMA3 70B t=0.2	58.14	52.08	54.95	63.86	57.20	60.35	71.03	63.63	67.13	83.59	74.88	78.99
LLAMA3 70B t=0.3	57.08	51.53	54.16	63.17	57.03	59.94	69.59	62.83	66.04	80.49	72.67	76.38

Table 13: Quality of atomic fact extraction on annotated samples.

	Strict			LCS			SBERT Elem			SBERT Full		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Human	81.94	81.94	81.94	84.93	84.93	84.93	88.10	88.10	88.10	90.32	90.32	90.32
GPT4-Turbo t=0	74.67	82.96	78.60	78.88	87.64	83.03	83.06	92.29	87.43	85.30	94.78	89.79
GPT4-Turbo t=0.1	73.61	78.52	75.99	78.17	83.38	80.69	82.70	88.21	85.37	84.37	89.99	87.09
GPT4-Turbo t=0.2	66.67	71.32	68.91	70.74	75.68	73.13	74.50	79.70	77.01	75.05	80.29	77.58
GPT4-Turbo t=0.3	75.00	76.60	75.79	77.45	79.09	78.26	79.33	81.02	80.17	80.26	81.97	81.11
GPT4o t=0	61.81	68.99	65.20	70.02	78.16	73.87	77.62	86.65	81.89	83.38	93.07	87.96
GPT4o t=0.1	60.28	67.46	63.67	68.72	76.90	72.58	77.50	86.72	81.85	84.84	94.94	89.61
GPT4o t=0.2	59.86	71.54	65.19	67.32	80.46	73.31	74.75	89.33	81.39	79.92	95.51	87.02
GPT4o t=0.3	61.81	70.63	65.93	71.60	81.82	76.37	79.83	91.23	85.15	86.06	98.35	91.80
LLAMA3 8B t=0	91.92	59.48	72.22	93.78	60.68	73.69	95.37	61.71	74.93	96.43	62.40	75.77
LLAMA3 8B t=0.1	94.79	59.48	73.09	96.23	60.38	74.20	97.91	61.43	75.49	99.40	62.37	76.60
LLAMA3 8B t=0.2	93.94	60.78	73.81	95.58	61.85	75.10	97.87	63.33	76.90	99.34	64.28	78.05
LLAMA3 8B t=0.3	92.93	60.13	73.02	94.41	61.09	74.18	95.56	61.83	75.08	96.71	62.58	75.99
LLAMA3 70B t=0	45.65	60.00	51.85	53.14	69.84	60.35	60.54	79.56	68.76	63.44	83.38	72.06
LLAMA3 70B t=0.1	47.83	59.46	53.01	56.00	69.63	62.08	63.57	79.03	70.46	67.37	83.75	74.67
LLAMA3 70B t=0.2	48.89	57.89	53.01	55.73	65.99	60.43	62.32	73.80	67.57	68.35	80.94	74.12
LLAMA3 70B t=0.3	59.85	64.23	61.96	66.57	71.44	68.92	72.08	77.36	74.63	75.50	81.02	78.16

Table 14: Quality of discourse extraction on annotated samples.