X-PARADE: Cross-Lingual Textual Entailment and Information Divergence across Paragraphs

Juan Diego Rodriguez◇   **Katrin Erk**◇♠   **Greg Durrett**◇
◇ Department of Computer Science   ♠ Department of Linguistics
The University of Texas at Austin
juand-r@utexas.edu

Abstract

Understanding when two pieces of text convey the same information is a goal touching many subproblems in NLP, including textual entailment and fact-checking. This problem becomes more complex when those two pieces of text are in different languages. Here, we introduce X-PARADE (**C**ross-lingual **Par**agraph-level **A**nalysis of **D**ivergences and **E**ntailments), the first cross-lingual dataset of paragraph-level *information divergences*. Annotators label a paragraph in a target language at the span level and evaluate it with respect to a corresponding paragraph in a source language, indicating whether a given piece of information is the same, new, or new but can be inferred. This last notion establishes a link with cross-language NLI. Aligned paragraphs are sourced from Wikipedia pages in different languages, reflecting real information divergences observed in the wild. Armed with our dataset, we investigate a diverse set of approaches for this problem, including token alignment from machine translation, textual entailment methods that localize their decisions, and prompting LLMs. Our results show that these methods vary in their capability to handle inferable information, but they all fall short of human performance.[1]

1 Introduction

The ability to recognize differences in meaning between texts underlies many NLP tasks such as natural language inference (NLI), semantic similarity, paraphrase detection, and factuality evaluation. Less work exists on the cross-lingual variants of these tasks. However, correctly identifying semantic relations between sentences in different languages has a number of useful applications. These include estimating the quality of machine translation output (Fomicheva et al., 2020), cross-lingual fact checking (Huang et al., 2022), and helping
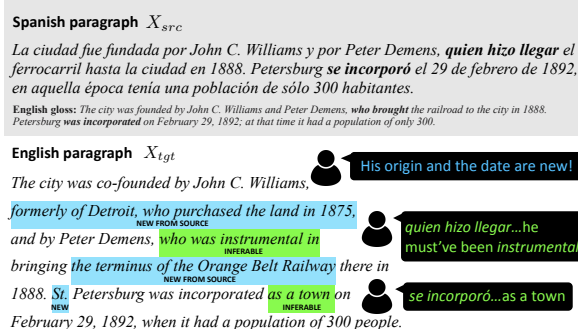

Figure 1: Wikipedia articles written in different languages often contain fine-grained differences in information, such as this paragraph pair taken from the English and Spanish articles on St. Petersburg, Florida. X-PARADE contains fine-grained span-level annotations for content in the target paragraph X_{tgt} that is *new* or *inferable* given the source paragraph X_{src}.

Wikipedia editors mitigate discrepancies in content across languages (Gottschalk and Demidova, 2017). The fact that different languages carve up the world in different ways (de Saussure, [1916] 1983; Liu et al., 2023) and have different syntactic constraints (Keenan, 1978) may also make these tasks more challenging.

Many of these tasks involve reasoning beyond the sentence level. At the level of paragraphs, it is no longer useful to have coarse labels like "entailed" or "neutral"; instead, we want to capture subtle differences in information content (Agirre et al., 2016; Briakou and Carpuat, 2020; Wein and Schneider, 2021). Thus, we focus on the problem of detecting fine-grained span-level *information divergences* between texts across languages. Notably, our notion of information divergences differentiates between new information and new information that can be inferred from the source paragraph.

This paper presents a dataset called X-PARADE : **C**ross-lingual **Par**agraph-level **A**nalysis of **D**ivergences and **E**ntailments. Figure 1 shows an example English-Spanish paragraph pair, with annotations on how the English paragraph

---

[1]Dataset available at https://github.com/juand-r/x-parade

1198

	WiCE	CLTE-2013	e-SNLI	iSTS	MS-RTE	MLQE-PE	REFRESD	X-PARADE
Cross-lingual	✗	✓	✗	✗	✗	✓	✓	✓
Multiple sentences	✓	✗	✗	✗	✗	✗	✗	✓
Fine-grained annotation	✓	✗	✓	✓	✓	✓	✓	✓
Entailment relations	✓	✓	✓	✓	✓	✗	✗	✓

Table 1: Comparison between X-PARADE and related datasets. Ours is the first dataset to provide cross-lingual, paragraph-level annotation of fine-grained entailment.

differs from the Spanish paragraph. We see a rich range of inferences being required to understand the target, including effects like *quien hizo llegar* (*who brought*) implying that someone was *instrumental* in bringing. These kinds of subtle cross-lingual divergences are anchored to individual spans in the target paragraph. Finally, unlike prior work that tackled sentence-level comparisons between languages (Briakou and Carpuat, 2020), we annotate entire paragraphs. By having larger textual units, we can capture a wider array of divergences and more appropriately model the nuances of cross-sentence context in this task.

We conduct annotation in three language pairs, yielding six directions, using trained annotators from Upwork who went through extensive qualification and feedback rounds. Our dataset is of high quality, with token-level Krippendorff α agreement scores ranging from 0.55 to 0.65, depending on the language pair.

Finally, we benchmark the performance of existing approaches on this problem. No systems in the literature are directly suitable. We compare a diverse set of techniques that solve different aspects of the problem, including token attribution of NLI models, machine translation (MT) alignment and large language models (LLMs). While GPT-4 performs the best, different approaches have different pros and cons and there remains a gap with human performance.

The main contributions of this work are:

1. We introduce X-PARADE (**C**ross-lingual **Par**agraph-level **A**nalysis of **D**ivergences and **E**ntailments), a dataset for fine-grained cross-lingual divergence detection at the paragraph level, containing four languages and six directions (ES-EN, EN-ES, EN-HI, HI-EN, ZH-EN, EN-ZH).

2. We analyze the ability of LLMs and techniques based on MT alignment and NLI to identify divergences. We show that the task is non-trivial even for state of the art models.

2 Task Setting and Related Work

2.1 Task Setting

Given pairs of paragraphs $(X_{\mathrm{src}}, X_{\mathrm{tgt}})$ with some overlapping information, we consider the problem of identifying spans in X_{tgt} (the target) containing information not present in X_{src} (the source). X_{src} and X_{tgt} are in different languages. Our dataset consists of a set of tuples $(X_{\mathrm{src}}, X_{\mathrm{tgt}}, S)$ where $S = \{(t_1, l_1), ...(t_n, l_n)\}$ is a set of labeled spans in the target paragraph X_{tgt}, and $l_i \in Y$ is a label characterizing how X_{tgt} differs from X_{src}. The task is to detect both the spans and their label for each $(X_{\mathrm{src}}, X_{\mathrm{tgt}})$. Monolingual variants of this task exist, but have mostly concerned themselves with sentence pairs; these include fine-grained textual entailment (Brockett, 2007), paraphrasing (Pavlick et al., 2015), detection of generation errors (Goyal and Durrett, 2020), including those from LLMs (Yue et al., 2023), and claim verification (Kamoi et al., 2023).

To determine an appropriate label set Y, we reviewed existing taxonomies, including taxonomies for paraphrases (Vila et al., 2014) and translations (Zhai et al., 2018). However, these were too fine-grained for our purpose (also including syntactic phenomena), and so we use the following mutually-exclusive classes for span-level annotations:[2]

1. **Same:** The span conveys information nearly identical to some part of the source paragraph.

2. **Inferable:** The span corresponds to a difference in content *inferable from background knowledge or reasoning* given the source paragraph.

3. **New:** The span corresponds to a difference in propositional content which cannot be inferred (either new or changed information).

---

[2]We initially included a fourth category for differences in connotation (e.g., "slender" vs "scrawny"). Given that there were relatively few connotation spans (less than 1% of tokens), and substantial disagreement between annotators, we decided to remove the connotation labels, and convert them to one of the other three classes, as described in Section 3.3.

We did not include a *contradiction* category as in traditional NLI tasks. Explicit contradictions were rare in the naturally-occurring data we observed. However, our taxonomy could be extended to support contradiction for future labeling efforts.

2.2 Related Tasks

Here we discuss tasks and datasets which are most closely related to our task. Table 1 compares these datasets and X-PARADE along different axes.

Semantic divergence detection The task of *semantic divergence detection*, i.e., identifying whether cross-lingual text pairs differ in meaning, was considered in Vyas et al. (2018), but not at the span-level. Wein and Schneider (2021) label semantic divergences between English and Spanish sentences based on their AMR representations, but the distinctions captured are more subtle than what we are aiming for, since some of the subtle distinctions do not affect inference. Briakou and Carpuat (2020) created a dataset, REFRESD, indicating which spans diverge in meaning between English and French sentences sampled from Wiki-Matrix (Schwenk et al., 2021). Framed in terms of our taxonomy, their dataset involves distinguishing **same** from **new or inferable** information; i.e., there is no distinction between information that can be inferred or not.

Textual entailment Several studies have considered the task of not only predicting entailment relations between sentence pairs, but also detecting which spans contribute to that decision. These tasks differ in terms of the structure and granularity of entailment relations. The MSR RTE dataset (Brockett, 2007) is the RTE-2 data (Haim et al., 2006) annotated with span alignment information. The e-SNLI dataset (Camburu et al., 2018) is annotated with spans which explain the relation (entailment, neutral or contradiction) between two sentences. Finally, the Interpretable STS (iSTS) shared task consisted in identifying and aligning spans between two sentences (Agirre et al., 2016) with labels similar to the Natural Logic entailment relations (Mac-Cartney and Manning, 2009). These studies use monolingual (English) sentences, unlike our work. Of these datasets, only iSTS distinguishes between **same** and **inferable** information.

Related to this work is fine-grained and explainable NLI. Zaman and Belinkov (2022) use MT alignment to measure the plausibility and faithfulness of token attribution methods for multilingual
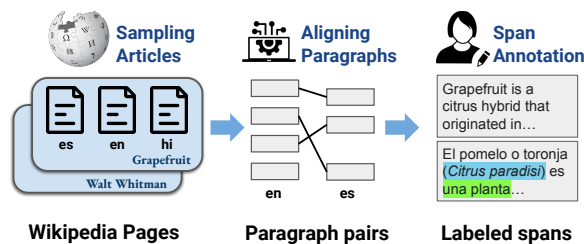

Figure 2: The dataset construction process. Sufficiently similar cross-lingual paragraph pairs are mined from Wikipedia, then annotated by experts.

NLI models. Their work builds on XNLI (Conneau et al., 2018), which uses translation and is typically handled in a monolingual setting. Stacey et al. (2022) build sentence-level NLI models by combining span-level predictions with simple rules. Finally, WiCE (Kamoi et al., 2023) consists of monolingual document-claim pairs with token-level labels for non-supported (i.e., non-entailed) tokens.

There is also a small literature on cross-language textual entailment (CLTE), mostly consisting of older techniques (Negri et al., 2012, 2013). There has been little work following in this vein, and modern neural methods enable us to pursue a more ambitious scope of changes detected.

Other tasks Two other tasks which also involve finding spans in text pairs are word-level quality estimation for MT, and factuality evaluation of generated summaries (Tang et al., 2023). MLQE-PE (Fomicheva et al., 2020) and HJQE (Yang et al., 2022) have been annotated for word-level MT quality estimation. XSumFaith (Maynez et al., 2020) and CLIFF (Cao and Wang, 2021) contain annotations of non-factual spans in generated summaries.

3 Dataset Construction

Our dataset construction pipeline is shown in Figure 2. It consists of three stages. We sample from a diverse set of Wikipedia pages, identify paragraph pairs that are sufficiently related but not identical to serve as candidates for our annotation, and present these to annotators to label.

3.1 Data Collection

Paragraph selection Wikipedia pages with versions in English, Spanish, Hindi and Chinese were sampled from the list of pages in CREAK (Onoe et al., 2021) in order to ensure a balanced distribution across topics. Paragraph alignment between

pages was performed by first computing paragraph-paragraph similarities with LaBSE (Feng et al., 2022), and selecting the set of pairs $\{(A_i, B_i)\}$ such that A_i and B_i mutually prefer each other over all other paragraphs, ensuring a 1-1 matching.

Finally, one paragraph pair was selected randomly from each article,[3] while ensuring similarity scores were distributed uniformly between 0.5 and 1. After a manual inspection, we further filtered paragraph pairs by length and similarity score. Additional details are given in Appendix B.

Annotation Process   We recruited workers with translation experience between the languages they were annotating. To ensure quality control, workers had to pass a qualification round. 210 paragraphs were annotated for each language pair in both directions (at an estimated average total time of 84 hours for each language pair). The instructions given to annotators are in Appendix G and the annotation interface is shown in Appendix E.

Both the adjudicated annotations (described in Section 3.3) and each annotator's individual annotations are made publicly available.

3.2   Inter-annotator Agreement (IAA)

Our task involves human judgements about natural language inference, which are known to be subjective (Pavlick and Kwiatkowski, 2019). There are many different reasons why annotators may disagree about whether one piece of information entails another (Jiang and de Marneffe, 2022). Here, we evaluate annotator agreement on our task, with a particular focus on the *inferable* category. Some annotators managed to identify a way to infer information in the target while others did not make such inferences and labeled tokens as *new*. In addition some inferences are quite direct, so some annotators labeled them as *same*. For example, there was disagreement over whether "changes its behavior in spring" is *new* or *inferable* in the following paragraph pair:

> Es: Las liebres son solitarias...**Tan solo se producen peleas durante la época de celo** (variable según especies)... Las liebres europeas de sexo masculino apenas comen durante **este período (primavera)**...[4]

---

[3]Given the prevalence of summary paragraphs, we resampled whenever either of the paragraphs was the first paragraph of the article.

[4]English gloss: "Hares are solitary...**Fights only occur during the mating season** (variable depending on species)... Male European hares hardly eat during **this period (spring)**..."

	Krippendorff's α	macro F1
EN-ES	0.657	62.5 ± 2.9
ES-EN	0.693	63.4 ± 3.4
EN-HI	0.605	64.9 ± 1.4
HI-EN	0.570	60.4 ± 1.0
EN-ZH	0.589	60.0 ± 3.4
ZH-EN	0.637	61.3 ± 4.2

Table 2: Inter-annotator agreement for X-PARADE. Both Krippendorff's α and macro F1 are calculated at the token level.

> En: Normally a shy animal, the European brown hare **changes its behavior in spring**...

In this case, to make the inference that these hares change their behavior in spring, one needs to to link "este período (primavera)" (spring) to "la época de celo" (mating season), and then realize that hares only fighting during mating season implies a change in their behavior in the spring. Additional examples of annotator disagreement over inferable spans are given in Appendix H.

With this context in mind, we compute two measures of inter-annotator agreement. Table 2 shows Krippendorff's α and token-level macro F1. Krippendorff's α is calculated at the token level following Goyal et al. (2022). Following Briakou and Carpuat (2020) and DeYoung et al. (2020), we report the token-level macro F1 score averaged over pairs of annotators (e.g., for three annotators, average over six F1 scores).

We also examine per-class agreement through sentence-level Krippendorff α scores and through per-class token-level F1 scores averaged over pairs of annotators (Table 3). Since we do not have sentence-level annotations, we observe whether each sentence contains a span of a given class or not in order to compute sentence-level Krippendorff α scores for each class. Our annotators strongly agree on content that is *same* or *new*, but have lower agreement about *inferable* annotations. As shown in the example above, this can be attributed to the highly subjective nature of the task of identifying natural language inferences (Pavlick and Kwiatkowski, 2019; Jiang and de Marneffe, 2022).

Handling inferable annotations   We observed that annotators were typically precise when they did select inferable tokens (i.e., they had a valid reason for why the token could be inferred). We can therefore take the union of *inferable* tokens annotated by different annotators (with some caveats,

	New		Inferable	
	α	F1	α	F1
EN-ES	0.634	84.5 ± 1.4	0.246	17.4 ± 6.9
ES-EN	0.664	86.9 ± 1.1	0.188	17.4 ± 8.3
EN-HI	0.555	77.8 ± 2.4	0.253	30.2 ± 2.3
HI-EN	0.540	75.1 ± 2.2	0.156	19.5 ± 5.4
EN-ZH	0.531	79.3 ± 3.3	0.169	16.2 ± 9.0
ZH-EN	0.572	85.7 ± 3.0	0.213	14.4 ± 11.7

Table 3: Krippendorff α for sentences and per-class token-level F1 scores over pairs of annotators.

	Paragraphs		Sentences		Tokens	
	Dev	Test	Dev	Test	Dev	Test
EN-ES	93	93	343	334	8565	8245
ES-EN	93	93	344	304	8933	8069
EN-HI	95	96	445	405	11087	10413
HI-EN	95	96	388	337	9560	8829
EN-ZH	100	99	228	204	7177	6938
ZH-EN	100	99	381	372	9903	9638

Table 4: Number of paragraphs, sentences and tokens in the X-PARADE dataset. For each pair, both paragraphs were annotated with spans indicating semantic divergence. Each row indicates the number of {paragraphs, sentences, tokens} in the target language (e.g., the Spanish language paragraphs, for EN-ES).

	Same	New	Inf	Same	New	Inf
	EN-ES			**ES-EN**		
Tokens	7797	7032	1981	7351	8183	1468
Spans	791	507	444	776	581	382
Sentences	486	464	283	451	477	242
	EN-HI			**HI-EN**		
Tokens	9779	7034	4687	9316	5858	3215
Spans	702	337	469	678	353	386
Sentences	604	402	394	577	349	315
	EN-ZH			**ZH-EN**		
Tokens	6556	4851	2708	6813	9378	3350
Spans	902	431	733	835	569	687
Sentences	351	266	295	409	541	383

Table 5: Distribution of class labels—same (**Same**), new information (**New**) and inferable (**Inf**)—over tokens, spans, and sentences in the *target* paragraph for different language pairs in the X-PARADE dataset. *Sentences* indicates the number of sentences containing at least one span in a given class.

discussed in Section 3.3) to arrive at high-precision inferable tokens for our dataset. This results in a natural interpretation for the *inferable* category: *someone* has reason to infer a given span, as exhibited by one of our annotators constructing an inference, which others possibly did not catch.

Manual inspection of 17 random Spanish-English paragraph pairs where annotators disagreed (given in Appendix H) supports this strategy. Of the 41 *inferable* spans that were disputed, we judged that 29 of them (71%) were inferable, 5 (12%) belonged to the *same* class, 4 (10%) belonged to the *new* class, and 3 (7%) could have been *inferable* or *new* depending on how much domain-specific background knowledge one has in order to judge the span as inferable. Here we accepted a range of inferences as valid, from more direct inferences such as "*las últimas décadas de la vida*" ⇒"*it is the end of the human life cycle*", to more indirect inferences such as the example of the European brown hare discussed above.

3.3 Adjudication

First, we removed any paragraph pairs whenever two annotators rejected the pair as being too dissimilar, or when at least two annotators selected over 95% of tokens as new. This left 186 paragraph pairs for English-Spanish (11% removed), 191 paragraph pairs for English-Hindi (9% removed) and 199 paragraph pairs for English-Chinese (5% removed).

We then adjudicate using majority vote at the token level, except when some annotator used the *inferable* label, where we always adjudicate the token as inferable, following the discussion in Section 3.2.[5] If *new* and *same* are tied, we break the tie

in favor of *new*, with similar logic as to why *inferable* is preferred. Connotation labels (less than 1% of the data; see footnote 2) are treated as inferable, since manual inspection revealed this class seemed most appropriate for most of them.

3.4 Dataset Statistics

X-PARADE consists of 576 paragraph pairs across three language pairs, with judgments on over 106,035 individual tokens. We split the pairs evenly between development and test sets. The number of paragraphs for each language pair are given in Table 4, and examples of annotated paragraphs can be found in Appendix D.

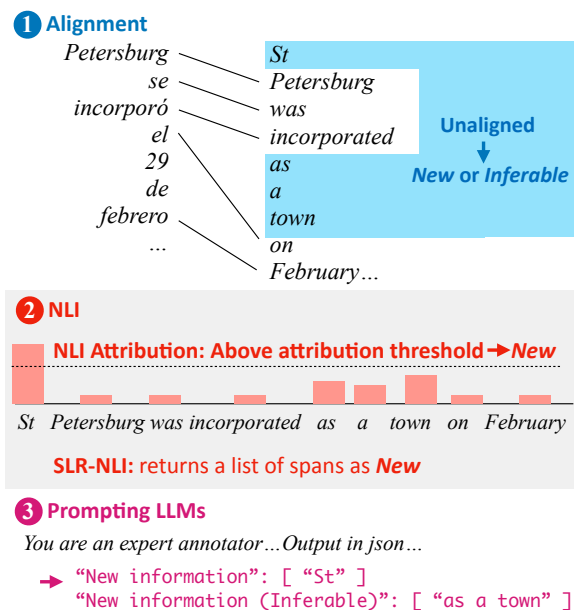The distribution of labels over tokens and spans is given in Table 5.

---

[5]The only exception to this rule is if only one annotator labeled a token as *inferable* while all the others labeled it as *same*; in this case we adjudicate it as *same*, since these are usually near-translations.

① Alignment

Petersburg
se
incorporó
el
29
de
febrero
...

St
Petersburg
was
incorporated
as
a
town
on
February...

Unaligned
↓
New or *Inferable*

② NLI

NLI Attribution: Above attribution threshold → *New*

St Petersburg was incorporated as a town on February

SLR-NLI: returns a list of spans as *New*

③ Prompting LLMs

You are an expert annotator...Output in json...

→ "New information": ["St"]
  "New information (Inferable)": ["as a town"]

Figure 3: Three of the methods illustrated schematically: (1) the MT-alignment based method attempts to align tokens across texts; tokens which can be aligned are *same*. (2) NLI can be used to either provide attribution scores or spans, identifying tokens which are non-inferable (*new*). (3) LLMs can be prompted to return any desired type of span.

4 Methods

While the task of detecting new and inferable information in paragraphs across languages is novel, it relates to ideas from machine translation and textual entailment. Here we describe how to adapt baselines from these areas to assess their performance on this task, as well as prompting LLMs to produce spans (Figure 3). Implementation details can be found in Appendix C.

Alignment MT word alignment predicts which words should be aligned across translations; thus words which do not easily align are more likely to present new content not given in the source paragraph. By way of approximation, we will assume in these experiments that unaligned tokens fall into the *new* category.

SLR-NLI SLR-NLI (Stacey et al., 2022) builds on the idea that a *neutral* or *contradiction* relation holds between two sentences only when there is at least one span in the "hypothesis" (target) that is not inferable from the premise. Since these spans are exactly the ones containing new information, we use SLR-NLI to predict which spans in the target paragraph are *new*.

	Outputs	Align	Translate
Alignment	token set	✓ (all)	×
SLR-NLI	phrase scores	✓ (EN-*)	✓
NLI Attribution	token scores	✓ (EN-*)	✓
LLM	generated text	×	×

Table 6: Summary of the methods compared. *Align* and *Translate* indicate whether MT alignment and translation are required. Translation is required for the NLI methods since we rely on English-language models.

NLI Attribution Rather than using the inherently interpretable method of Stacey et al. (2022), we can instead use a standard NLI system equipped with a post-hoc interpretation method. We use token attribution methods for NLI models to score the tokens most responsible for a *neutral* classification decision. We compute an attribution score for each token; higher-scoring tokens should be new and not inferable.

LLMs We use one-shot prompting of three state-of-the-art LLMs, **GPT-3.5-turbo**, **GPT-4**, and **Llama-2-chat** (Touvron et al., 2023), and two explicitly multilingual LLMs, **BLOOMZ** (Muennighoff et al., 2023) and **XGLM** (Lin et al., 2022). BLOOMZ is an instruction-tuned model, while XGLM is a non-instruction tuned autoregressive LM. We used prompts that specify the annotation task, given in Appendix F.

The four different methods are compared and summarized in Table 6. Alignment outputs a set of unaligned tokens, while SLR-NLI and NLI token attribution methods produce scores for phrases and tokens, respectively. The LLM generates strings which are then matched to the target paragraph.

5 Results

5.1 New information detection (N v. S+I)

Here we discuss results on the binary task of new information detection, i.e., grouping together the classes *same* and *inferable*. Performance on the EN-ES and ES-EN test sets are shown in Table 7. We omit scores for BLOOMZ since it substantially underperformed XGLM on every language pair. F1 scores are compared across language pairs in Figure 4, and full results for the dev set and other language pairs are in Appendix A. **Human*** denotes an estimate of human performance on the task, given by evaluating every annotator against the majority vote of the other annotators, and breaking ties in favor of *new*.
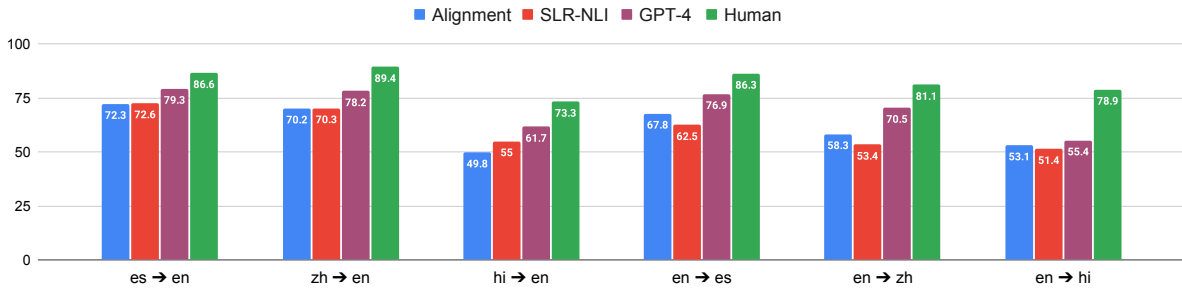
Figure 4: F1 scores for Alignment, SLR-NLI, GPT-4 and human performance on the new information detection task, evaluated on the test set.

	ES → EN			**EN → ES**		
	P	R	F1	P	R	F1
Majority baseline	44.6	100.0	61.7	39.8	100.0	57.0
Alignment	62.3	86.1	72.3	55.4	87.4	67.8
NLI Attr. (IG)	64.3	78.4	70.7	*51.7*	*80.8*	*63.1*
SLR-NLI	67.9	78.1	72.6	*60.5*	*64.6*	*62.5*
XGLM (7.5B)	45.4	30.9	36.8	42.1	21.4	28.3
Llama-2-chat (7B)	52.4	33.2	40.7	50.0	25.9	34.2
GPT-3.5-turbo	57.4	80.6	67.1	50.9	88.7	64.6
GPT-4	70.4	90.6	79.3	66.3	91.4	76.9
w/ Translation to English						
Llama-2-chat (T)	52.3	32.3	40.0	*50.8*	*28.5*	*36.5*
GPT-3.5-turbo (T)	61.0	82.1	70.0	*54.7*	*75.2*	*63.3*
GPT-4 (T)	72.0	89.7	79.9	*63.0*	*80.4*	*70.6*
Human*	86.8	86.5	**86.6**	85.7	87.0	**86.3**

Table 7: Precision, recall and F1 scores for new information detection on the English-Spanish test set. Scores in italics indicate methods where both translation and MT alignment was used on the target paragraph.

The EN-HI and HI-EN subsets are harder than EN-ES and ES-EN; one possible explanation for this is the relative scarcity of Hindi web text, which affects all the NLP components we use (alignment, translation, language models). For every language pair, GPT-4 achieved the highest F1-scores, but there is still a gap in performance compared to humans.[6] GPT-3.5-turbo struggles at the task, with scores similar to or worse than the non-LLM methods. XGLM (7.5B) and Llama-2-chat (7B) do worse than the majority-vote baseline. This is due to poor instruction-following capacity: we found they often copy from both paragraphs, and sometimes translate them. Both behaviors result in spans that cannot be matched with text in the target. Alignment is surprisingly effective, perform-

[6]Since GPT-4 is a closed, proprietary model, we believe there is substantial room to improve performance on this benchmark from the perspective of open models.

ing similarly to SLR-NLI for ES-EN. On the other hand, for HI-EN, SLR-NLI outperforms Alignment by 5 points.

Does translating into English improve LLM performance? When the source language was Spanish (ES-EN), we observed a small improvement when giving GPT-3.5-turbo translations of the source paragraph (67.1 to 70.0); for HI-EN the improvement was more substantial (43.4 to 53.0), and for ZH-EN using translations had almost no effect. For the EN-* language pairs, translating the target paragraph to English did not help GPT-3.5-turbo in most cases; this is likely due to errors in mapping the tokens back to the target language with the MT aligner. Translating to English did not help GPT-4, which already seems to have strong multilingual capabilities, or Llama-2-chat, which struggled to follow instructions regardless of the language.

5.2 Inferable Spans

Both Alignment and NLI Attribution methods only return binary predictions, and so we cannot use them to distinguish inferable spans from *new* or *same*. Where do inferable spans fall? Intuitively, the perfect NLI classifier should fail to distinguish between *same* and *inferable* (predicting the negative class for both) since both lead to entailment. Alignment, on the other hand, should group *new* and *inferable* together in the positive class, since only tokens which are near-perfect translations of each other should align.

Unfortunately, this straightforward picture is not reflected in the system behavior. For the ES-EN dev set, we noticed that Alignment predicts the positive class for 80.6% of inferable tokens. However, NLI Attribution predicts the negative class for only 33.4% of inferable tokens. If both Alignment and

	Same	New	Inf	Total
Same	2941	498	241	3680
New	405	3087	108	3600
Inf	153	515	121	789
Total	3499	4100	470	

Table 8: Confusion matrix for GPT-4's predictions on the three-way task, on the ES-EN test set. Rows are the true class labels and columns are predicted labels.

NLI Attribution were working perfectly according to our intuitions we would expect all Alignment predictions to be Positive, and all of the NLI attributions to be Negative.

5.3 Three-way Divergence Classification with GPT-4

Here we present results on the full divergence taxonomy by prompting GPT-4 with one example including both *inferable* and *new* spans.

We first analyze the raw predictions made by GPT-4 (Table 8). We note that GPT-4 predicts the inferable label far less frequently than its frequency in our dataset (470 vs 789), and that many predictions are actually same (50%) or new (23%). However, it is able to follow the task format and achieves strong performance on *same* and *new* tokens, as suggested by our results in Section 5.1.

One example of an incorrectly assigned *inferable* label, is shown below, with GPT-4's prediction highlighted in green:

Es: Los elementos geológicos de Fobos se han nombrado en memoria de astrónomos relacionados con el satélite[7]

En: Geological features on Phobos are named after astronomers who studied Phobos

"Geological...astronomers" should have been labeled *same* as it closely matches the Spanish.

Comparison to human performance Next, we compare GPT-4 against human performance (Human*), which is estimated similarly to Section 5.1 (except that since it was three-way classification we used the same adjudication procedure as in Section 3.3). For the three-way task, overall performance is slightly lower than human performance (Table 9) for all language pairs.

Finally, Table 10 compares GPT-4 and estimated human performance at classifying *inferable* tokens. GPT-4 performs worse than Human*, mainly due

---

7Gloss: "The geological features of Phobos have been named in memory of astronomers associated with the satellite"

		P	R	F1
EN-ES	GPT-4	62.1	59.5	58.9
	Human*	$69.2_{\pm3.5}$	$64.7_{\pm2.8}$	$64.6_{\pm2.4}$
ES-EN	GPT-4	61.7	60.3	60.4
	Human*	$70.3_{\pm3.7}$	$65.3_{\pm3.2}$	$65.1_{\pm2.8}$
EN-HI	GPT-4	51.3	52.4	49.4
	Human*	$66.2_{\pm0.6}$	$65.8_{\pm1.2}$	$65.6_{\pm1.0}$
HI-EN	GPT-4	52.8	55.2.	50.6
	Human*	$61.8_{\pm0.9}$	$61.5_{\pm0.3}$	$61.3_{\pm0.8}$
EN-ZH	GPT-4	51.8	54.0	51.4
	Human*	$62.6_{\pm0.9}$	$61.1_{\pm1.0}$	$59.5_{\pm2.4}$
ZH-EN	GPT-4	57.7	56.0	55.4
	Human*	$67.3_{\pm2.0}$	$65.0_{\pm3.1}$	$62.8_{\pm3.2}$

Table 9: GPT-4 vs estimated human performance on the three-way classification task; the scores are macro precision, recall and F1 scores on the test set.

		P	R	F1
EN-ES	GPT-4	33.2	13.1	18.8
	Human*	$41.2_{\pm13.9}$	$22.9_{\pm10.6}$	$25.7_{\pm5.5}$
ES-EN	GPT-4	25.7	15.3	19.2
	Human*	$42.0_{\pm15.3}$	$22.7_{\pm11.6}$	$24.9_{\pm6.9}$
EN-HI	GPT-4	29.1	8.9	13.7
	Human*	$33.0_{\pm3.3}$	$33.1_{\pm7.3}$	$32.2_{\pm2.7}$
HI-EN	GPT-4	23.4	9.5	13.5
	Human*	$19.4_{\pm0.6}$	$19.8_{\pm7.3}$	$18.9_{\pm4.3}$
EN-ZH	GPT-4	19.9	8.0	11.4
	Human*	$28.8_{\pm6.4}$	$20.1_{\pm13.9}$	$18.9_{\pm9.2}$
ZH-EN	GPT-4	30.3	13.3	18.5
	Human*	$36.2_{\pm10.7}$	$21.4_{\pm17.4}$	$19.5_{\pm10.5}$

Table 10: Performance on *inferable* tokens (GPT-4 vs estimated human performance) on the test set.

to low recall. Due to the subjectivity mentioned earlier in Section 3.2, it is difficult to obtain an accurate measure of human performance. However, given our analysis of the adjudicated results in Section 3.2 and Appendix H, we believe that achieving high *precision* of inferable tokens should be possible, even if recall is low, and GPT-4 is far below human performance at this aspect of the task.

6 Conclusion

We present X-PARADE , a new dataset of cross-lingual paragraph pairs (English-Spanish, English-Hindi, English-Chinese), annotated for semantic divergences at the span-level. Although the task features subjectivity, the analysis of our annotation shows that decisions by the annotators were well-justified. We show that while some of these fine-grained differences can be detected by GPT-4, there is still a gap with human performance. We believe that this dataset can be useful for benchmarking

the inferential capabilities of multilingual LLMs and analyzing how textual entailment systems can identify information divergences cross-lingually.

Limitations

We only compared languages from two different language families (Indo-European and Sino-Tibetan); future work could surface different kinds of differences, reflective either of cultural or typological differences (for an example in Malagasy, see Keenan (1978)). Our focus was also on locating inferable or new information, but further work could expand on this to include other aspects such as structuring of information (e.g., discourse markers) and whether information is contradictory rather than merely new. Further, we noted that inferences annotated in X-PARADE are sometimes subjective and can take many different forms. Future work could try to further understand the kinds of inferences being made, building on prior work such as Joshi et al. (2020) and Jiang and de Marneffe (2022).

We explored several baselines for the task, but the methods (e.g., Alignment, NLI Attribution) were not well-suited to distinguish *inferable* from *new* or *same* spans. We hope to see the development of new methods designed explicitly for this task; we believe that better trained cross-lingual NLI systems could potentially be effective here.

Finally, future work could seek to understand why LLMs classify spans as *inferable*. To what extent is it drawing from its parametric knowledge? Given that GPT-4 has seen all of Wikipedia, what constitutes "background knowledge" for LLMs and for people is very different. Future work could consider forcing GPT-4 to explain itself (as in chain-of-thought prompting), or explore different structures for how it should generate the data (e.g., forcing it to generate the text spans relevant to the inference).

Acknowledgments

---

[8] https://goodsystems.utexas.edu/

responsible AI technologies.

References

Eneko Agirre, Aitor Gonzalez-Agirre, Iñigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. 2016. SemEval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 512–524, San Diego, California. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Online. Association for Computational Linguistics.

Chris Brockett. 2007. Aligning the RTE 2006 corpus. *Microsoft Research*, 57.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9560–9572.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ferdinand de Saussure. [1916] 1983. *Course in General Linguistics*. Duckworth, London. (trans. Roy Harris).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and

Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick R. Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020. MLQE-PE: A multilingual quality estimation and post-editing dataset. *CoRR*, abs/2010.04480.

Simon Gottschalk and Elena Demidova. 2017. Multiwiki: Interlingual text passage alignment in wikipedia. *ACM Trans. Web*, 11(1):6:1–6:30.

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, Junyi Jessy Li, and Greg Durrett. 2022. SNaC: Coherence error detection for narrative summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 444–463, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7.

Kung-Hsiang Huang, ChengXiang Zhai, and Heng Ji. 2022. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Trans. Assoc. Comput. Linguistics*, 10:1357–1374.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. *CoRR*, abs/2303.01432.

E. L. Keenan. 1978. Some logical problems in translation. In F. Guenthner and M. Guenthner-Reutter, editors, *Meaning and Translation: Philosophical and Linguistic Approaches*, pages 157–189. New York University Press, New York, NY.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schütze. 2023. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12969–13000, Toronto, Canada. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ines Montani and Matthew Honnibal. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 399–407, Montréal, Canada. Association for Computational Linguistics.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2013. Semeval-2013 task 8: Cross-lingual textual entailment for content synchronization. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 25–33, Atlanta, Georgia, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A dataset for commonsense reasoning over entity knowledge. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *Prague Bull. Math. Linguistics*, 106:125–146.

Ellie Pavlick, Johan Bos, Malvina Nissim, Charley Beller, Benjamin Van Durme, and Chris Callison-Burch. 2015. Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1512–1522, Beijing, China. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, and Marek Rei. 2022. Logical reasoning with span-level predictions for interpretable and robust NLI models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3809–3823, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin F. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11626–11644. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Marta Vila, M Antònia Martí, Horacio Rodríguez, et al. 2014. Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(01):205.

Yogarshi Vyas, Xing Niu, and Marine Carpuat. 2018. Identifying semantic divergences in parallel text without annotations. 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 1503–1515, New Orleans, Louisiana. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2021. Classifying divergences in cross-lingual AMR pairs. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 56–65, Punta Cana, Dominican Republic. 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.

Zhen Yang, Fandong Meng, Yuanmeng Yan, and Jie Zhou. 2022. Rethink about the word-level quality estimation for machine translation from human judgement. *arXiv preprint arXiv:2209.05695*.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *CoRR*, abs/2305.06311.

Kerem Zaman and Yonatan Belinkov. 2022. A multilingual perspective towards the evaluation of attribution methods in natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1556–1576, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuming Zhai, Aurélien Max, and Anne Vilnat. 2018. Construction of a multilingual corpus annotated with translation relations. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A   Additional Results

Additional results on new information detection are given below in Tables 11, 12, and 13.

	ES → EN			EN → ES		
	P	R	F1	P	R	F1
Majority baseline	51.3	100.0	67.8	43.7	100.0	60.9
Alignment	67.4	87.6	76.1	58.2	87.7	70.0
NLI Attr. (IG)	66.2	78.0	71.6	*53.9*	*80.9*	*64.7*
SLR-NLI	69.3	76.6	72.8	*62.9*	*70.6*	*66.5*
XGLM (7.5B)	58.3	33.8	42.8	44.5	18.5	26.1
Llama-2-chat (7B)	57.5	33.5	42.4	49.0	24.0	32.2
GPT-3.5-turbo	63.6	77.7	69.9	53.7	82.9	65.2
GPT-4	75.3	91.4	82.6	72.2	89.4	79.9
w/ Translation to English						
Llama-2-chat (T)	63.4	34.1	44.3	*55.4*	*28.0*	*37.2*
GPT-3.5-turbo (T)	63.8	82.3	71.9	*56.2*	*73.7*	*63.7*
GPT-4 (T)	73.8	89.2	80.7	*66.6*	*79.3*	*72.4*
Human*	90.5	90.8	**90.7**	87.1	88.1	**87.6**

Table 11: Precision, recall and F1 scores for new information detection on the English-Spanish dev set.

	HI → EN			EN → HI		
	P	R	F1	P	R	F1
Majority baseline	35.9	100.0	52.9	34.9	100.0	51.7
Alignment	45.9	82.5	59.0	42.7	85.0	56.8
NLI Attr. (IG)	49.1	88.4	63.1	*45.7*	*61.5*	*52.4*
SLR-NLI	58.4	79.8	67.5	*50.4*	*57.9*	*53.9*
XGLM (7.5B)	31.5	20.9	25.1	41.2	24.2	30.5
Llama-2-chat (7B)	37.2	39.1	38.1	38.2	33.1	35.5
GPT-3.5-turbo	42.1	72.2	53.2	41.2	64.7	50.4
GPT-4	57.7	95.6	69.6	52.4	68.9	59.5
w/ Translation to English						
Llama-2-chat (T)	46.8	36.2	40.8	*45.1*	*26.4*	*33.3*
GPT-3.5-turbo (T)	51.4	83.8	63.7	*45.3*	*58.3*	*51.0*
GPT-4 (T)	56.0	94.4	70.3	*50.9*	*61.5*	*55.7*
Human*	64.5	85.1	**72.6**	66.3	84.4	**73.5**

Table 12: Precision, recall and F1 scores for new information detection on the English-Hindi dev set.

B   Dataset Construction

Wikipedia paragraph selection Pywikibot[9] was used to download articles which had versions in English, Spanish, Chinese and Hindi.[10] These were split into sections and paragraphs with wikitextparser.[11]

---

[9]v. 8.0.1, https://pypi.org/project/pywikibot/
[10]Download date: March 22, 2023.
[11]v. 0.51.1, https://pypi.org/project/wikitextparser/

	HI → EN			EN → HI		
	P	R	F1	P	R	F1
Majority baseline	27.4	100.0	43.1	30.4	100.0	46.6
Alignment	35.8	81.8	49.8	38.3	86.5	53.1
NLI Attr. (IG)	37.8	88.5	52.9	*43.0*	*64.9*	*51.8*
SLR-NLI	46.2	67.9	55.0	*46.3*	*57.8*	*51.4*
XGLM (7.5B)	27.7	24.7	26.1	43.1	29.0	34.7
Llama-2-chat (7B)	28.8	37.7	32.6	31.9	32.2	32.1
GPT-3.5-turbo	31.6	69.5	43.4	33.6	62.5	43.7
GPT-4	46.4	92.1	61.7	47.4	66.7	55.4
w/ Translation to English						
Llama-2-chat (T)	34.1	31.9	33.0	*39.2*	*24.6*	*30.2*
GPT-3.5-turbo (T)	38.7	84.0	53.0	*40.4*	*55.7*	*46.9*
GPT-4 (T)	45.9	91.5	61.1	*50.3*	*62.5*	*55.7*
Human*	65.6	85.8	**73.9**	66.5	86.4	**74.8**

Table 13: Precision, recall and F1 scores for new information detection on the English-Hindi test set.

	ZH → EN			EN → ZH		
	P	R	F1	P	R	F1
Majority baseline	47.4	100.0	64.3	33.2	100.0	49.8
Alignment	58.1	86.2	69.4	44.2	81.8	57.4
NLI Attr. (IG)	57.9	91.1	70.8	*43.0*	*71.8*	*53.8*
SLR-NLI	65.4	79.0	71.5	*53.1*	*53.7*	*53.3*
XGLM (7.5B)	49.8	34.8	41.0	33.7	33.6	33.7
Llama-2-chat (7B)	52.3	34.5	41.6	33.7	38.5	36.0
GPT-3.5-turbo	58.6	76.8	66.4	39.5	77.0	52.2
GPT-4	69.4	90.0	78.3	56.0	89.3	68.8
w/ Translation to English						
Llama-2-chat (T)	56.8	34.8	43.1	*39.0*	*25.9*	*31.1*
GPT-3.5-turbo (T)	55.2	81.4	65.8	*41.3*	*68.3*	*51.5*
GPT-4 (T)	63.6	88.1	73.9	*47.8*	*72.0*	*57.4*
Human*	89.5	88.1	**88.6**	83.0	83.9	**82.9**

Table 14: Precision, recall and F1 scores for new information detection on the English-Chinese dev set.

	ZH → EN			EN → ZH		
	P	R	F1	P	R	F1
Majority baseline	48.6	100.0	65.4	35.6	100.0	52.5
Alignment	59.1	86.5	70.2	45.9	79.9	58.3
NLI Attr. (IG)	58.4	91.6	71.3	*44.5*	*73.6*	*55.5*
SLR-NLI	66.2	75.0	70.3	*54.6*	*51.5*	*53.0*
XGLM (7.5B)	51.1	35.1	41.6	39.1	35.2	37.0
Llama-2-chat (7B)	51.9	36.7	43.0	36.8	40.6	38.7
GPT-3.5-turbo	58.3	80.8	67.7	40.6	73.0	52.2
GPT-4	68.5	91.1	78.2	58.3	88.3	70.5
w/ Translation to English						
Llama-2-chat (T)	57.1	34.3	42.9	*43.0*	*31.2*	*36.2*
GPT-3.5-turbo (T)	58.4	82.6	68.4	*45.4*	*70.5*	*55.2*
GPT-4 (T)	65.5	91.4	76.3	*53.9*	*73.2*	*62.1*
Human*	90.7	88.8	**89.4**	82.2	81.7	**81.1**

Table 15: Precision, recall and F1 scores for new information detection on the English-Chinese test set.

After selecting paragraph pairs, we further filtered the data according to length and paragraph similarity score. We only kept those with English paragraphs containing between 86 and 1000 characters, (Spanish, Hindi) paragraphs containing between 120 and 1000 characters, and a similarity score between .63 and .95. For Chinese-English paragraphs, we removed pairs where the Chinese paragraphs had over 250 characters.

Annotation Process We recruited workers from Upwork, selecting those who were either bilingual or fluent in either language, and who had translation experience between the languages of interest. To ensure quality control, workers had to pass a qualification round consisting of 14 paragraph pairs

(i.e., 7 pairs, but annotating both directions). These qualification rounds also served to give feedback to the annotators. Four annotators were chosen for English-Spanish and English-Chinese, and three annotators were chosen for English-Hindi. Annotators were paid $300 for every 140 paragraph pairs (70 paragraph pairs, in both directions), at an estimated hourly rate of $10-$25. Annotators were hired from Argentina, Colombia, India, China and the US. All annotators had at least an undergraduate degree, and seven had post-graduate degrees.

Annotators were presented with each (Spanish, Hindi, Chinese) paragraph first and asked to annotate the related English paragraph; then the order of the paragraphs were flipped and they were asked to annotate the (Spanish, Hindi, Chinese) paragraph. Annotators were able to reject (toss out) paragraphs that were too dissimilar, for cases where the entire paragraph was new (in each direction) or when the paragraphs had superficial similarities but were about completely different subjects. They were also given the option to leave a comment for each paragraph pair in order to leave feedback or outline their thought process. The instructions given to annotators are in Appendix G. Prodigy (Montani and Honnibal) was used for the annotation interface, shown in Appendix E.

C Implementation Details

Alignment We use SimAlign (Jalili Sabet et al., 2020), an MT aligner based on comparing cosine similarities of mBERT embeddings. SimAlign was chosen because its performance is comparable to

the best supervised aligners such as fastalign/IBM2 (Dyer et al., 2013), efmaral/eflomal (Östling and Tiedemann, 2016) and Giza++/IBM4 (Och and Ney, 2003).[12] We use the *argmax* method of SimAlign, and tune the null threshold τ on our dev set in order to maximize F1. For ES-EN and EN-ES we used $\tau = 0.9997$, for HI-EN and EN-HI we used $\tau = 0.99979$, and for ZH-EN and EN-ZH we used $\tau = 0.99976$. Evaluations with Alignment were done on a laptop with 32 GB of RAM and no GPUs.

SLR-NLI We retrained the SLR-NLI BERT-based model.[13] We sentence-segment the target paragraph and run SLR-NLI on each (paragraph, sentence) pair. When the target paragraph is non-English, any predicted spans are mapped back to the source paragraph using an MT aligner (SimAlign with *itermax*). We used SLR-NLI with combinations of 2 consecutive spans.[14] The threshold for selecting neutral and contradiction spans was tuned on the development set. We used thresholds of 0.15 for ES-EN and EN-ES, 0.20 for HI-EN, 0.10 for EN-HI, 0.15 for ZH-EN, and 0.25 for EN-ZH. Evaluations with SLR-NLI were performed on a laptop with 32 GB of RAM and no GPUs.

NLI Attribution We used a BERT-based (Devlin et al., 2019) NLI model trained on MNLI[15] (Williams et al., 2018). For our attribution method, we use integrated gradients (Sundararajan et al., 2017). We chose this model and attribution method after preliminary experiments comparing three different attribution methods (Saliency, InputXGradients and Integrated Gradients) and two different models (BERT and deBERTa[16]) on the dev set of ES-EN.

NLI Attribution experiments were done on one NVidia Titan RTX GPU. Thresholds for selecting tokens based on their attribution scores were tuned on the development set. We used thresholds of 0.03052 for ES-EN, 0.02263 for EN-ES, .02260 for HI-EN and EN-HI, 0.02260 for ZH-EN, and 0.02470 for EN-ZH.

Intuitively, spans which contain new information not present in the source paragraph should cause NLI models to classify the hypothesis as *neutral* or *contradiction*. SLR-NLI is designed explicitly to find these spans, while attribution methods may surface tokens which are neutral with higher attribution scores. Since both SLR-NLI and the token attribution model are monolingual (English) models, for both methods we first translate either the source (for *-EN pairs) or the target (for EN-* pairs) paragraph to English using Google Translate.[17] When the language of the source paragraph is non-English, we use its translation as the premise, and when the target language is non-English, we translate the target paragraph to English to use as the hypothesis[18] for the NLI model, and any localized spans must be mapped via MT alignment back to the tokens of the target paragraph. For NLI Attribution, we follow Zaman and Belinkov (2022) and do this by summing the attribution scores for all translation tokens which map onto a target paragraph token.

LLMs When testing **GPT-3.5-turbo** and **GPT-4**, we specifically used gpt-3.5-turbo-0613 and gpt-4-0613, since these models will not be updated.[19] We used the 7B version of **Llama-2-chat** (Touvron et al., 2023), the 7.1B version of **BLOOMZ** (Muennighoff et al., 2023) and the 7.5B version of **XGLM** (Lin et al., 2022). BLOOMZ is an instruction-tuned model, while XGLM is a non-instruction tuned autoregressive LM. We used prompts that specify the annotation task in depth, given in Appendix F.

We obtained similar performance for the GPT models whether we presented the data as (paragraph, paragraph) pairs or (paragraph, sentence) pairs, so we only report the paragraph-level version. However, sentence segmenting and running the LLMs for (paragraph, sentence) pairs was slightly more effective for the smaller LMs, so we report the sentence-level version for these. For GPT-3.5-turbo and GPT-4 we used a temperature of 0.7 and top-p of 1, while for the smaller language models we used greedy decoding.

For the BLOOMZ, XGLM and Llama-2-chat experiments, we used a NVIDIA RTX A6000 GPU.

---

[12]Giza++, fastalign and efmaral refer to different implementations of the original systems.

[13]Without e-SNLI supervision, using the defaults in `https://github.com/joestacey/snli_logic`

[14]See the discussion in Section 2.2 of Stacey et al. (2022).

[15]`https://huggingface.co/gchhablani/bert-base-cased-finetuned-mnli`

[16]Specifically deBERTa trained on a mix of NLI datasets: `https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c`

[17]Future work can consider inherently cross-language NLI models.

[18]Or, more precisely, each sentence of the translated target paragraph is a hypothesis; we run each method over all (paragraph, sentence) pairs and aggregate the results.

[19]`https://platform.openai.com/docs/models/continuous-model-upgrades`, last accessed Oct. 15, 2023.

Llama-2-chat took under 30 minutes per experiment, while XGLM and BLOOMZ took under 1 hour per experiment.

D   Dataset Examples

Figure 5 shows three examples from our dataset.

ES: Los estados y fases del sueño humano se definen según los patrones característicos que se observan mediante el electroencefalograma (EEG), el electrooculograma (EOG, una medición de los movimientos oculares) y el electromiograma de superficie (EMG, movimiento de los músculos esqueléticos). El registro de estos parámetros electrofisiológicos para definir los estados de sueño y de vigilia se denomina polisomnografía.

EN: Key physiological methods for monitoring and measuring changes during sleep include electroencephalography (EEG) of brain waves, electrooculography (EOG) of eye movements, and electromyography (EMG) of skeletal muscle activity. Simultaneous collection of these measurements is called polysomnography, and can be performed in a specialized sleep laboratory. Sleep researchers also use simplified electrocardiography (EKG) for cardiac activity and actigraphy for motor movements .

ES: En la década de 1820, las tensiones sectarias en Inglaterra se habían aliviado y algunos escritores británicos comenzaron a preocuparse, pues la Navidad estaba en vías de desaparición. Dado que imaginaban la Navidad como un tiempo de celebración sincero, hicieron esfuerzos para revivir la fiesta. El libro de Charles Dickens Un cuento de Navidad, publicado en 1843, desempeñó un importante papel en la reinvención de la fiesta de Navidad, haciendo hincapié en la familia, la buena voluntad, la compasión y la celebración familiar.

EN: In the early-19th century, writers imagined Tudor Christmas as a time of heartfelt celebration. In 1843, Charles Dickens wrote the novel A Christmas Carol, which helped revive the "spirit" of Christmas and seasonal merriment. Its instant popularity played a major role in portraying Christmas as a holiday emphasizing family, goodwill, and compassion.

ES: La película se reestrenó en formato 3D el 4 de abril de 2012, seis días antes de la fecha del centenario de la partida del Titanic de Inglaterra y un mes antes del centésimo aniversario de Paramount Pictures, la otra casa productora de la película. Junto con la recaudación del reestreno, la recaudación total de la película suma 2185372302 dólares.

EN: The 3D version of Titanic premiered at the Royal Albert Hall in London on March 27 , 2012 , with James Cameron and Kate Winslet in attendance , and entered general release on April 4 , 2012 , six days shy of the centenary of RMS Titanic embarking on her maiden voyage.

Figure 5: Three examples of paragraph pairs from the ES-EN portion of X-PARADE annotated with spans for *new information* (blue) and *inferable* (green).
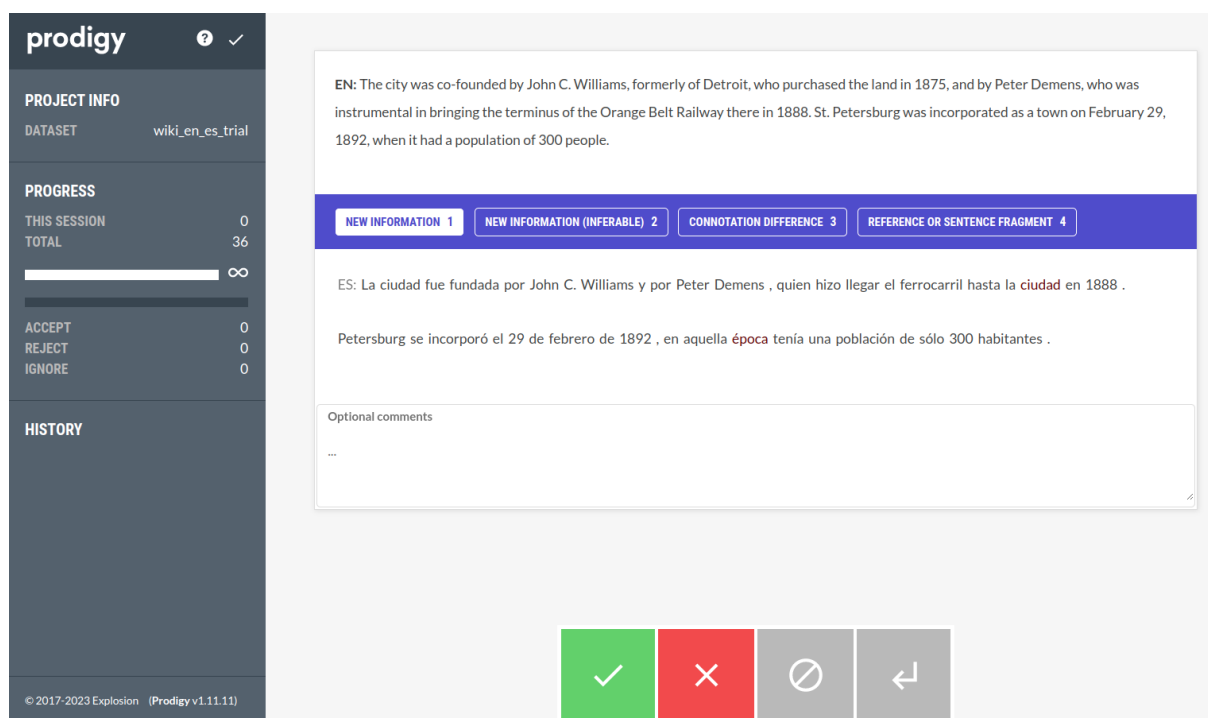
E   Annotation Interface

Figure 6: Screenshot of the interface used to annotate the dataset. We highlighted in red tokens based on the output of a word aligner in order to enable annotators to more easily spot differences. However, annotators were cautioned that these highlights were merely suggestions.

F Prompt for LLMs

The prompts used for the LLMs in our experiments are shown below in Figures 7 and 8:

You are an expert annotator, fluent in English and Spanish, and with extensive translation experience. This annotation task is to identify pieces of content that differ in paragraph pairs across different languages. You will be given two paragraphs: one in Spanish and one English. These are not necessarily translations of each other. For the second paragraph, find all spans of text corresponding to new information. This is content which is not given in the other one and which cannot be inferred using reasoning or background knowledge. Do not swap names with pronouns or modify the text, i.e., copy the spans verbatim. Format your span selections in json format. Annotate only the second paragraph.

First paragraph. ES: La ciudad fue fundada por John C. Williams y por Peter Demens, quien hizo llegar el ferrocarril hasta la ciudad en 1888. Petersburg se incorporó el 29 de febrero de 1892, en aquella época tenía una población de sólo 300 habitantes.

Second paragraph. EN: The city was co-founded by John C. Williams, formerly of Detroit, who purchased the land in 1875, and by Peter Demens, who was instrumental in bringing the terminus of the Orange Belt Railway there in 1888. St. Petersburg was incorporated as a town on February 29, 1892, when it had a population of 300 people.

OUTPUT in json:
{
"New Information": [
"formerly of Detroit",
"who purchased the land in 1875",
"the terminus",
"Orange Belt Railway",
"St.",
]}

Figure 7: One-shot prompt used for the evaluating GPT-3.5-turbo and GPT-4 on the ES-EN portion of the dataset. For the other language directions we translated the output spans and source and target paragraphs appropriately. For the smaller LMs, we had it output a list rather than a json, since they struggled in producing valid json.

You are an expert annotator, fluent in English and Spanish, and with extensive translation experience. This annotation task is to identify pieces of content that differ in paragraph pairs across different languages. You will be given two paragraphs: one in Spanish and one English. These are not necessarily translations of each other. For the second paragraph, please select all spans which fall under the following mutually exclusive categories:

- "New Information": Content in the paragraph which is not given in the first one and which cannot be inferred using reasoning or background knowledge.
- "New Information (Inferable)": New content in the second paragraph that is not present in the first paragraph, but which can reasonably be inferred from it. Inferences can make use of information in the first paragraph, background knowledge or commonsense reasoning.

Anything not labeled as one of these classes is taken to have the same information content as in first other paragraph. Do not swap names with pronouns or modify the text, i.e., copy the spans verbatim. Format your span selections in json format. Annotate only the second paragraph.

First paragraph. ES: La ciudad fue fundada por John C. Williams y por Peter Demens, quien hizo llegar el ferrocarril hasta la ciudad en 1888. Petersburg se incorporó el 29 de febrero de 1892, en aquella época tenía una población de sólo 300 habitantes.

Second paragraph. EN: The city was co-founded by John C. Williams, formerly of Detroit, who purchased the land in 1875, and by Peter Demens, who was instrumental in bringing the terminus of the Orange Belt Railway there in 1888. St. Petersburg was incorporated as a town on February 29, 1892, when it had a population of 300 people.

OUTPUT in json:
{
"New Information": [
"formerly of Detroit",
"who purchased the land in 1875",
"the terminus",
"Orange Belt Railway",
"St."
],
"New Information (Inferable)": [
"who was instrumental",
"as a town"]
}

Figure 8: One-shot prompt used for evaluating GPT-4 on three-way (new, same, inferable) classification task on the ES-EN portion of the dataset. For the other language directions we translated the output spans and source and target paragraphs appropriately.

G   Annotator Instructions

Task Description

Thank you for participating in this task!

This annotation task is to identify pieces of content that differ in paragraph pairs across different languages. You will be given two paragraphs: one in English, and one in another language. These are not necessarily translations of each other. For each pair of paragraphs, please do the following:

- Read the first paragraph carefully.
- Read the second paragraph, and select spans (contiguous word sequences) in the second paragraph that differ in meaning with respect to the first paragraph shown. Spans may indicate differences falling into one of the following four categories (with examples given in English for simplicity):

1. New Information - Content in one paragraph which is not given in the other one and which cannot be inferred (using reasoning or background knowledge).

- This could be content that is **added** (e.g., "Charles Dickens was born on 7 February 1812 _in Portsea Island_" vs "Charles Dickens was born on 7 February 1812") or **changed** (e.g., "Saint Patrick's Day is a religious and cultural _holiday_" vs "Saint Patrick's Day is a religious and cultural _festival_". In this example, "festival" is labeled as "new information" because not every holiday is a festival. **Note that we are particularly interested in fine-grained meaning changes like this.**

2. New Information (Inferable) -New content in one paragraph that is not present in the other, but which can reasonably be inferred from it. Inferences can make use of information in the paragraph, background knowledge or commonsense reasoning.

- **An example of background knowledge:**
    - PARAGRAPH 1: "Michael Jackson was heavily influenced by funk, disco and gospel."
    - PARAGRAPH 2: "_The King of Pop_ was heavily influenced by funk, disco and gospel."

    Here, since it is common knowledge that Michael Jackson is known as The King of Pop, but "The King of Pop" does not appear explicitly in the first paragraph, "The King of Pop" should be labeled as "new information (inferable)"

- **Examples of reasoning:**
  - "70% of the students passed the exam" vs "*30% of the students failed the exam*"
  - "Emma burst into tears" vs "*Emma cried*"

One way to think about the "new information (inferable)" label is: the information given is different, but if I *know* the information given in the first paragraph, then I would be able to infer (without extensive web searches or deep subject matter expert knowledge) the information in the second paragraph.

3. Connotation Difference - The words or expressions express the same thing exactly, but have different connotations (i.e., different associations or attitudes expressed even though the literal meaning is the same, such as "the *slender* man" vs "the *scrawny* man").

4. Reference - The span corresponds to a reference, which interrupts the flow of the text. For example, consider the following paragraph:

> "Though the rebels lacked military training, they displayed skilful use of available local materials and unusual tactics against the disciplined Roman armies. *Frontinus, Stratagems, Book I, 5:20–22 and Book VII:6.* They spent the winter of 73–72 BC training, arming and equipping their new recruits."

It is clear that "Frontinus, Stratagems, Book I, 5:20–22 and Book VII:6." is not part of the main text and is actually a reference being cited.

Annotation Instructions

First you will be presented with the English paragraph and asked to annotate the paragraph in the other language. Then the paragraphs will be swapped and you will annotate information in the English paragraph which is not in the other one. To annotate: (1) first select the button for the desired label, and (2) drag the cursor over a span to label it. If some span is labeled by accident, you can remove the label by hovering over it and clicking, as shown below:
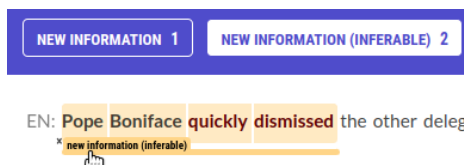
After selecting the spans, click the **green check mark (✓)** button to progress to the next paragraph pair. To save your annotations, please click the **save/floppy disk (💾)** icon in the top left (or Ctrl+S keys) to save your work.

Notice some words are shown in red. These words are more likely (but not guaranteed) to not correspond to any word in the other paragraph; we hope this helps direct your attention to the "most different" parts of the content, and to make it easier to annotate.

Two more buttons are available to you:

- **"Ignore"** (⊘) can be used to move on to the next example without annotating. Please use this sparingly (less than 5% of the time), and only in cases where there is something really wrong with the example. For example, this could happen if the paragraphs are just snippets of math equations that you can't make sense of.
- **"Undo"** (↵) can be used if you need to go back to edit the previously annotated paragraph pair.

Finally, every paragraph pair also has a box with "Optional comments". You do **not** need to fill this, but it can be used to give feedback or ask questions if the need arises.

Special Cases

Some paragraph pairs are may have some superficial similarity, but not actually be about the same thing or event. If the paragraphs are not about the same thing, you can click the red ✕ button to move on to the next example. **Only use this is *the entire* paragraph is completely different.** If even one sentence has some information overlapping, it should be labeled.

Some paragraphs are so close they can be considered perfect translations. In that case nothing needs to be labeled; please do not label anything and just click the green check mark button to progress to the next example.

FAQ

- **What if I want to indicate a "deletion" in the paragraph being annotated? (content that occurs in the first paragraph but not in the second)**
  - A: This is equivalent to there being new content in the first paragraph. You can annotate this directly as "new information" in the paragraph once the paragraphs flip.

- **What constitutes "background knowledge"?**

- A: When deciding if something is "new information (inferable)" you can call upon background knowledge. This is anything that you know and you think is common knowledge, like referring to President Biden as "Joe Bien".

- **I suspect something is not factually correct. What do I do?**
  - A: This task does not concern factual correctness. You do not need to look up whether something is actually true or not, but only how the meanings of the two paragraphs compare to each other.

- **Should I annotate differences in grammar?**
  - A: Only in cases where this would trigger a change in meaning. For example, differences in tense or gender which are clearly typos should not be marked. Similarly, differences in tense may not necessarily indicate differences in when events occur.

- **How should pronouns be handled? (e.g., one paragraph uses "he", while the other refers to a specific person by name).**
  - A: If it is clear that these refer to the same entity, then do not label it.

H Examples of Inferable Span Disagreement

Source paragraph (truncated)	Target paragraph (truncated)	Inferable span?
Las liebres son solitarias, aunque no les importa en absoluto la presencia de otras liebres en los alrededores. **Tan solo se producen peleas durante la época de celo** (variable según especies), que pueden llegar a ser hasta cierto punto cómicas en algunas especies. **Las liebres europeas de sexo masculino apenas comen durante este período (primavera)**, y pasan el día luchando con sus rivales...	Normally a shy animal, the European brown hare changes its behavior in spring, when it can be seen in daytime chasing other hares...	No, Yes
Tercera edad o senectud es un término antroposocial que hace referencia a **las últimas décadas de la vida**, en la que uno **se aproxima a la edad máxima que el ser humano puede vivir**.	Old age is the range of ages nearing and surpassing the life expectancy of human beings; it is the end of the human life cycle...	No, Yes, No, Yes
El Templo de Júpiter del Capitolino fue comenzado por Tarquinio Prisco y completado por el último rey de Roma, Tarquinio el Soberbio, aunque fue inaugurado, según una tradición registrada por los historiadores, el 13 de septiembre, al comienzo de la época republicana...	The building was supposedly begun by king Tarquinius Priscus, completed by the last king (Tarquinius Superbus) and inaugurated.	Yes
La Semana Santa, y la Pascua en particular, está ligada a través de la última cena y la crucifixión de Jesús a la Pésaj (Pascua Judía) y **al Éxodo del pueblo hebreo narrado en el Antiguo Testamento**...	Easter is linked to Passover and the Exodus from Egypt recorded in the Old Testament through the Last Supper, sufferings, and crucifixion of Jesus that preceded the resurrection.	Yes
Luego de ello, Affleck **comenzó a salir con** Jennifer Lopez en julio **de 2002 tras protagonizar juntos Gigli** (2003). También trabajaron en Jersey Girl (2004) y en el videoclip de «Jenny from the Block». Su relación, la cual fue apodada «Bennifer» y considerada como una superpareja, atrajo una atención masiva por parte de los medios y se generaron una gran cantidad de rumores sobre aspectos personales de ambos. La pareja se comprometió en noviembre de 2002 y tenían una boda prevista para el 14 de septiembre de 2003, pero fue pospuesta apenas cuatro días antes del evento a causa del acoso de los paparazzi. Finalmente **se separaron en enero de 2004** en buenos términos.	Affleck first dated Jennifer Lopez from 2002 to 2004. They became friends on the set of Gigli in December 2001, having previously encountered each other at industry parties. They began a romantic relationship in July 2002 when Lopez filed for divorce from her second husband, Cris Judd.	Yes, Yes, Yes
La pelvis es la región anatómica inferior del tronco. Siendo una cavidad, la pelvis es un embudo osteomuscular que se estrecha hacia abajo, **limitado por** el hueso sacro, el cóccix y los coxales (que forman la cintura pélvica) y **los músculos de la pared abdominal inferior y del perineo**. Limita un espacio **llamado cavidad pélvica, en donde se encuentran órganos importantes**...	The pelvic region of the trunk is the lower part of the trunk, between the abdomen and the thighs. It includes several structures: the bony pelvis, the pelvic cavity, the pelvic floor, and the perineum.	Yes, Yes
Ese mismo año conoció a uno de los grandes amores de su vida, Charles-Joseph Lamoral, **príncipe** de Ligne.	Early in Bernhardt 's career, she had an affair with a Belgian nobleman, Charles-Joseph Eugène Henri Georges Lamoral de Ligne (1837–1914), son of Eugène, 8th Prince of Ligne, with whom she bore her only child, Maurice Bernhardt (1864–1928).	Yes
El 19 de abril, Malcolm X concluyó el Hajj, dando las siete vueltas alrededor de la Kaaba, bebiendo del Pozo de Zamzam y corriendo siete veces a través de las colinas de Al-Safa y Al-Marwah. Según su autobiografía, este viaje le permitió ver a los musulmanes **de diferentes razas** que interaccionan como iguales **y llegó a creer** que el islam puede superar los problemas raciales.	MalcolmX later said that seeing Muslims of "all colors, from blue-eyed blonds to Black-skinned Africans," interacting as equals led him to see Islam as a means by which racial problems could be overcome.	Yes, No, No
Las ovejas han tenido una fuerte presencia en la cultura de muchos países, especialmente en las zonas donde constituyen el tipo más común de ganado. En la literatura, especialmente en las fábulas, son las representantes típicas de la bondad, **mansedumbre** y las pocas luces, en contraposición con el lobo o el zorro...	In the English language, to call someone a sheep or ovine may allude that they are timid and easily led.	Yes
Después de **la conquista** del Imperio aqueménida, a manos de Alejandro Magno, y más tarde, tras la caída de los partos, el Imperio sasánida gobernó **el norte y el sur** del golfo, manteniendo la Ruta de la Seda.	Following the fall of Achaemenid Empire, and after the fall of the Parthian Empire, the Sassanid Empire ruled the northern half and at times the southern half of the Persian Gulf...	Yes, No, No, Maybe
El kriya yoga es **la forma práctica** de las doctrinas del yoga, la unión **con Dios** mediante la devoción activa y la realización correcta de los deberes diarios.	The "science" of Kriya Yoga is the foundation of Yogananda's teachings. An ancient spiritual practice, Kriya Yoga is union (yoga) with the Infinite.	Maybe, Yes, No

Table 16: Examples of spans labeled as inferable (green) in the ES-EN portion of X-PARADE where not all annotators agreed on the span label. The right column shows, for each span, whether we judge the span to be inferable (*Yes*), not inferable (*No*, shown in blue in cases where *new* is a more appropriate label), and *Maybe* for cases where our the answer depends on how much domain-specific background knowledge one draws from to make the inference. The most relevant parts of the Spanish paragraph for each judgement are shown in **bold**.

Source paragraph (truncated)	Target paragraph (truncated)	Inferable span?
Rachel Louise Carson (27 de mayo de 1907 - 14 de abril de 1964) fue una bióloga marina y conservacionista estadounidense que, a través de **la publicación de** Primavera silenciosa en 1962 y otros escritos, **contribuyó a la puesta en marcha de la moderna conciencia ambiental**.	Rachel Louise Carson (May 27, 1907 – April 14, 1964) was an American marine biologist, writer, and conservationist whose influential book Silent Spring (1962) and other writings are credited with advancing the global environmental movement.	Yes, Yes, Yes, Yes
La búsqueda de **los rasgos** de líderes han sido una constante en todas las culturas durante siglos. Escrituras filosóficas como la República de Platón o las Vidas de Plutarco han explorado una pregunta básica: «¿Qué cualidades distinguen a un líder?».	The search for the characteristics or traits of leaders has continued for centuries. Philosophical writings from Plato's does not use the word "leadership"...	No
En **la década de 1820**, las tensiones sectarias en Inglaterra se habían aliviado y algunos escritores británicos comenzaron a preocuparse, pues la Navidad estaba en vías de desaparición. Dado que imaginaban la Navidad como un tiempo de celebración sincero, hicieron esfuerzos para revivir la fiesta. **El libro de Charles Dickens Un cuento de Navidad, publicado en 1843**, **desempeñó un importante papel** en la reinvención de la fiesta de Navidad, haciendo hincapié en la familia, la buena voluntad, la compasión y la celebración familiar.	In the early-19th century, writers imagined Tudor Christmas as a time of heartfelt celebration. In 1843, Charles Dickens wrote the novel A Christmas Carol, which helped revive the "spirit" of Christmas and seasonal merriment.	Yes, Yes, Yes, Yes
La longitud es una medida de una dimensión (lineal; por ejemplo la distancia en m), mientras que el área es una medida de dos dimensiones (al cuadrado; por ejemplo m²), y el volumen es una medida de tres dimensiones (cúbica; por ejemplo m³).	Length is the measure of one spatial dimension, whereas area is a measure of two dimensions (length squared) and volume is a measure of three dimensions (length cubed).	Yes, Yes, Yes
Es **venerado** como santo por la Iglesia evangélica luterana en Estados Unidos (Calendario de Santos Luterano) **y la Iglesia anglicana**. Su festividad se conmemora el 31 de marzo.	Donne is remembered in the Calendar of Saints of the Church of England, the Episcopal Church liturgical calendar and the Calendar of Saints of the Evangelical Lutheran Church in America for his life as both poet and priest.	Yes, No
El ácido láctico, o su forma ionizada, el lactato (del lat. lac, lactis, leche), también conocido por su nomenclatura oficial ácido 2-hidroxi-propanoico o ácido α-hidroxi-propanoico, es un compuesto químico que desempeña importantes roles en varios procesos bioquímicos, como la fermentación láctica. Es un ácido carboxílico, **con un grupo hidroxilo en el carbono adyacente al grupo carboxilo**, lo que lo convierte en un ácido α-hidroxílico (AHA) de fórmula H3C-CH(OH)-COOH (). **En solución puede perder el hidrógeno unido al grupo carboxilo y convertirse en el anión lactato.**	Production includes both artificial synthesis as well as natural sources. Lactic acid is an alpha-hydroxy acid (AHA) due to the presence of a hydroxyl group adjacent to the carboxyl group. It is used as a synthetic intermediate in many organic synthesis industries and in various biochemical industries. The conjugate base of lactic acid is called lactate (or the lactate anion).	No, Maybe

Table 17: Examples of spans labeled as inferable (green) in the ES-EN portion of X-PARADE where not all annotators agreed on the span label. The right column shows, for each span, whether we judge the span to be inferable (*Yes*), not inferable (*No*, shown in blue in cases where *new* is a more appropriate label), and *Maybe* for cases where our the answer depends on how much domain-specific background knowledge one draws from to make the inference. The most relevant parts of the Spanish paragraph for each judgement are shown in **bold**.