Comparing Human and Large Language Model Perceptions of Semantic Similarity: Insights from the SimilEx Dataset

Chiara Alzetta*
Istituto di Linguistica Computazionale
"Antonio Zampolli", CNR

Felice Dell'Orletta**
Istituto di Linguistica Computazionale
"Antonio Zampolli", CNR

Chiara Fazzone†
Istituto di Linguistica Computazionale
"Antonio Zampolli", CNR

Giulia Venturi‡
Istituto di Linguistica Computazionale
"Antonio Zampolli", CNR

This paper investigates how human annotators and Large Language Models (LLMs) assign and justify semantic similarity judgments in a Semantic Textual Similarity (STS) task. To this end, we present a new version of SimilEx, the first Italian dataset containing human similarity judgments and natural language explanations for sentence pairs, extended with LLM-generated scores and explanations, enabling a direct comparison between human and LLM behaviour under parallel annotation conditions. Within this framework, we examine the extent to which humans and LLMs align in their perception of sentence similarity. We explore this question from multiple perspectives, including the relationship between sentence-level stylistic features and similarity scores, the consistency of judgments across annotator types, and the alignment of human and LLM explanations. Our findings show that LLMs tend to express more moderate judgments than humans, resulting in higher agreement. At the same time, stylistic features of the evaluated sentences are related to similarity judgments in both groups. As for explanations, humans typically produce shorter, often nominal constructions, reflecting more individually driven strategies for justifying similarity judgments, whereas LLMs generate more canonical sentence structures whose content is also more consistent across models, suggesting that justification is a more subjective process for humans.

1. Introduction and Motivation

Large language models (LLMs) exhibit impressive linguistic capabilities and achieve outstanding performance across a wide range of natural language processing tasks. This is particularly true for the most recent and ground-breaking models such as GPT-3.5/4 (Achiam et al. 2023), LLaMA-2 (Touvron et al. 2023), and Gemini (Gemini Team

* Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab
E-mail: chiara.alzetta@ilc.cnr.it
** Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab
E-mail: felice.dellorletta@ilc.cnr.it
† Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab
E-mail: chiara.fazzone@ilc.cnr.it
‡ Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa - ItaliaNLP Lab
E-mail: giulia.venturi@ilc.cnr.it

et al. 2023). Despite their remarkable abilities, the opacity of these models remains a significant concern (Belinkov and Glass 2019; Doshi-Velez and Kim 2017). Users often struggle to understand the reasoning behind LLM-generated outputs, creating a trust gap between humans and these artificial agents. This lack of transparency is particularly problematic given that these models frequently produce factual inaccuracies (Maynez et al. 2020; Augenstein et al. 2024) and hallucinations (Ji et al. 2023). Bridging this gap requires mechanisms that allow LLMs to explain their decisions in ways that are both comprehensible and meaningful to humans.

Natural Language Explanations have emerged as a promising approach to address this challenge (Zhao et al. 2024; Wiegreffe and Marasović 2021). By presenting rationales in plain language, explanations may contribute to improving model interpretability, thereby fostering user trust. In addition, the potential of using LLMs to generate natural language explanations, and whether these can effectively complement or even replace human-written justifications, remains an open and actively researched question, especially in real-world applications that require reliable and transparent decision-making (Chen et al. 2025; Huang et al. 2023; Lubos et al. 2024).

Annotated datasets containing natural language explanations are key to addressing this issue, as they enable the development of models capable of generating human-like explanations for their decisions. Therefore, multiple datasets have been created with free-form explanations to be incorporated into the model training process and used as benchmarks at test time, mostly focusing on English (Wiegreffe and Marasović 2021). Some examples are the e-SNLI dataset (Camburu et al. 2018), a version of the Stanford Natural Language Inference (SNLI) dataset (Bowman et al. 2015) enriched with human-annotated explanations; the Common Sense Explanations (CoS-E) (Rajani et al. 2019) and Semi-Structured Explanations for COPA (COPA-SSE) (Brassard et al. 2022) datasets, which include natural language explanations for commonsense reasoning; HateXplain (Mathew et al. 2021), a dataset of social media posts annotated for multiple aspects of hate speech along with textual rationales that support the given categorization, intended for explainable hate speech detection. To the best of our knowledge, the only existing dataset enriched with explanations for Italian is 'e-RTE-3-it' (Zaninello, Brenna, and Magnini 2023), an Italian version of the RTE-3 dataset (Dagan, Glickman, and Magnini 2005) for textual entailment.

Human-written explanations also provide an essential point of comparison for analysing how natural language explanations generated by language models differ from human reasoning in terms of coherence and relevance. For instance, (He et al. 2024) prompt multiple LLMs with a small set of human explanations to generate new ones, which are then evaluated by human judges. Interestingly, their results show that annotators often prefer LLM-generated explanations over human-written ones. Another relevant study, particularly significant for our work, is presented by (Di Bonaventura et al. 2024), who build on the HateXplain dataset and conduct a before-and-after evaluation involving human annotators to assess the usefulness and trustworthiness of LLM-generated explanations in the context of abusive language detection. The authors find that LLM-generated explanations often fail to meet human expectations: after being exposed to these explanations, participants significantly reduce their ratings of both usefulness and trustworthiness. These outcomes motivate us to develop a resource containing both human- and LLM-generated explanations, enabling systematic comparison and analysis of their similarities and differences. We focused on Semantic Textual Similarity (STS), a well-established yet methodologically challenging Natural Language Understanding (NLU) task, which consists in determining the degree of semantic equivalence between two texts (Wang and Dong 2020; Agirre et al. 2013). STS

is a foundational NLU problem relevant to many applications such as summarisation, question answering and conversational systems. Its relevance is reflected in its presence across several editions of the SemEval evaluation campaign, the international initiative for evaluating semantic processing tasks and natural language processing models. Since 2012, STS has been the focus of multiple SemEval tasks, with some variation in task settings and language coverage across editions (see (Cer et al. 2017) for a survey of task editions). In all cases, the datasets distributed to participants were annotated by human annotators, typically recruited through crowdsourcing platforms. Beyond the datasets developed for evaluation purposes, numerous other resources for STS have also been created, although the majority are available only for English (see (Fodor, Deyne, and Suzuki 2025) for a recent overview). However, to the best of our knowledge, no existing STS dataset also includes natural language explanations specifically collected to justify similarity judgments between sentence pairs.

As background, it should also be noted that STS remains an inherently subjective task, even for humans, as highlighted by (Wang et al. 2023) and (Fodor, Deyne, and Suzuki 2025). Similarity judgments may rely on different aspects of text, such as semantic content, event structure, lexical overlap, or stylistic properties, depending on which cues annotators consider most salient. As a result, the perception of similarity may vary across annotators and text types. To explore this issue in greater depth, (Fodor, Deyne, and Suzuki 2025) constructed a dataset in which sentence pairs were systematically varied along specific semantic dimensions, such as argument structure and event modification, in order to examine how these factors influence human similarity ratings. Their findings underscore the multidimensional nature of the task and point to the need for richer forms of annotation that capture not only the judgments themselves but also the underlying reasoning behind them.

Starting from these premises, in this paper, we build on the work of (Alzetta et al. 2024), who introduced the SimilEx dataset[1], which, to the best of our knowledge, is the first Italian dataset containing 2,112 sentence pairs manually annotated for semantic similarity, accompanied by natural language explanations justifying the assigned similarity scores. While the original study focused on analyzing human perceptions of semantic similarity and the rationales provided by annotators, the present work extends the dataset with similarity scores and natural language explanations generated by three LLMs. In line with the acknowledged subjective nature of STS, we did not start from a predefined theoretical notion of similarity intended to guide human and artificial annotators. Accordingly, the annotation guidelines did not impose restrictions on how similarity should be interpreted or assessed. This methodological choice constitutes a key aspect of our contribution, as it yields a resource that enables the examination of how different agents (human and artificial) converge or diverge in their perception of semantic similarity and in the reasoning they make explicit to justify their judgments.

Contributions. In this paper, we make three main contributions:

- We present an extended version of the SimilEx dataset, the first Italian resource featuring human-annotated semantic similarity scores and accompanying natural language explanations, by enriching it with similarity judgments and explanations generated by three state-of-the-art LLMs;

1 The dataset is freely available at `http://www.italianlp.it/resources/`.

- We carry out an in-depth analysis of the subjectivity underlying sentence similarity judgments, with a focus on the agreement and divergence between human and LLM-provided scores;
- We perform a multifaceted analysis of the content and writing style of human- and LLM-generated explanations, providing insights in the strategies used by humans and language models to justify their similarity judgment.

In the remainder of the paper, Section 2 presents the SimilEx dataset, detailing the data collection and annotation process. Section 3 investigates how humans and LLMs perceive similarity by analysing their scores and inter-annotator agreement. We further examine whether the linguistic style of the evaluated sentence pairs influences these judgments in Section 3.3. Section 4 shifts focus to the natural language explanations. Section 4.1 compares writing styles across annotators, while Section 4.2 addresses two research questions: whether similar explanations are provided for the same pairs by different annotators, and how explanation content relates to the assigned similarity scores. Finally, Section 5 concludes the paper and discusses future research directions.

2. The SimilEx Dataset

SimilEx is the first Italian dataset comprising 2,112 sentence pairs annotated for semantic similarity. The dataset was initially introduced in (Alzetta et al. 2024), focusing exclusively on manual human annotations. In this section of the extended version, we provide a comprehensive description of the data acquisition process that led to the construction of the dataset (Section 2.1). Furthermore, we detail the annotation procedure for sentence pairs, distinguishing between annotations performed by human annotators and those generated by Large Language Models, as described in Section 2.2.

2.1 Data Collection

The sentence pairs of SimilEx are sourced from a collection of novels translated into Italian from the late XIX century. This genre has been largely unexplored in the STS studies, which have predominantly focused on more informational textual genres such as newswire, social media posts, and forum discussions. As discussed in what follows, literary texts exhibit distinctive stylistic properties that differentiate them from the genres typically considered in STS datasets. To identify candidate sentence pairs exhibiting semantic similarity, we employed Sentence-BERT (SBERT) (Reimers and Gurevych 2019). SBERT is a modification of BERT (Devlin et al. 2019) made adequate to produce sentence embeddings suitable for STS since they can be compared to evaluate their similarity using cosine similarity, a metric ranging from 0 (which, in this setting, indicates no similarity between the vectors) to 1 (identical vectors, so possibly identical sentences). We included in the SimilEx dataset only pairs obtaining a cosine similarity score ≥ 0.65, resulting in a total of 2,112 sentence pairs. Note that this threshold was empirically selected as a pre-filtering criterion to exclude pairs that are largely unrelated, thereby avoiding the inclusion of sentence pairs that would be uninformative for studying similarity judgments.

Stylistic Properties of Sentences. The textual genre of the sentences (i.e., novels) introduces specific stylistic properties that present potential differences from standard Italian. In

order to characterize the style in which the SimilEx sentences were written, we assessed their linguistic properties using Profiling-UD (Brunato et al. 2020)[2], a web-based tool that captures multiple aspects of sentence structure. The tool extracts around 130 properties representative of the underlying linguistic structure of a sentence, derived from raw, morphosyntactic, and syntactic levels of sentence annotation, all based on the Universal Dependencies (UD) formalism (de Marneffe et al. 2021). These properties have been shown to be highly predictive when used as features by learning models in various classification tasks, such as Automatic Readability and Linguistic Complexity Assessment or Native Language Identification. The analysis of features extracted from SimilEx sentence pairs aims not only to characterise their textual genre, but also to explore potential stylistic differences between paired sentences. [3]

We began our linguistic analysis by computing, for each pair of sentences in the SimilEx dataset, the cosine similarity between the vectors of 130 linguistic traits extracted from each sentence. The resulting average similarity score on the full dataset was 0.61 ($\pm$0.19), suggesting that the sentence pairs in the dataset share several morpho-syntactic traits, but they still exhibit a degree of structural and stylistic variability.

One key feature is **sentence length**, which often correlates with syntactic complexity, elaboration, and narrative style. Within pairs, the average length difference is 17.02 tokens ($\pm$19.55). This value, combined with such a high standard deviation, suggests a large variability of style even between sentences deemed semantically similar. Notably, the average sentence length in SimilEx is 30.18 tokens ($\pm$22.36), which is substantially higher than the average for standard Italian sentences, typically around 20 tokens. This discrepancy reflects the literary origin of the corpus, where sentence structures tend to be more elaborate and stylistically marked.

The notable variability between paired sentences extends, e.g., to two local syntactic phenomena often considered indicative of diverse aspects of linguistic variation: the **distribution of subordinate clauses** and the relative position of nuclear elements (subjects and objects) relative to the verb. Regarding the former, in 64.87% of SimilEx pairs, either both sentences include at least one subordinate clause or neither does. For those pairs where both sentences contain subordinate clauses, the average difference in the number of subordinates is 2.25 ($\pm$1.81). Although this may seem minor, such a difference is quite large considering that each pair consists of two single sentences, making this divergence a non-negligible stylistic factor. Concerning **word order**, a pattern of marked asymmetry also emerges. In 30% of cases (651 pairs), only one of the two sentences contains a post-verbal subject, a non-canonical construction typically associated with stylistic emphasis or topicalization. At the pair level, 39.44% of sentence pairs (833 pairs) feature this construction in at least one sentence, but only 13.26% (280 pairs) display it in both. These findings further underscore the rich stylistic variation present within the dataset and provide a ground for investigating whether and how such variation influences human judgments of similarity.

Lastly, we assessed **lexical overlap**, a key dimension of both stylistic and semantic similarity. We calculated the overlap across different parts of speech (nouns, adjectives, verbs) as well as over all tokens (including punctuation). **Nouns** exhibited the highest overlap, with an average of 13.76%, and only 288 out of 2,112 sentence pairs (13.63%) showing zero noun overlap. In contrast, **adjectives** and **verbs** had much lower aver-

2 The complete set of linguistic characteristics used for the stylistic analysis can be found in Appendix 2.
3 We distribute the values of the linguistic characteristics computed for each SimilEx sentence at `http://www.italianlp.it/resources`.

age overlaps of 1.68% and 3.00%, respectively, with around 80% of pairs showing no overlap in either category. Overall, the global lexical overlap across all tokens averaged just 12.60%, indicating that in SimilEx many sentence pairs exhibit semantic similarity beyond surface-level lexical overlap.

Taken together, these results suggest that SimilEx is characterised by a high degree of intra-pair stylistic variability. This characteristic sets it apart from many existing similarity datasets and makes it particularly valuable for investigating how stylistic differences interact with human perceptions of semantic similarity and the explanatory strategies adopted by human annotators and language models.

2.2 Similarity Annotation

We formulated the annotation task as a Semantic Textual Similarity (STS) task, in which annotators are presented with pairs of sentences and asked to assess the degree to which the two sentences are similar, but without providing any formal definition of similarity.

Sentence pairs of the SimilEx were annotated by two groups of annotators. A first group is composed of **human annotators**. As previously reported in (Alzetta et al. 2024), all human annotators are Italian native speakers, and the sample includes both a pool recruited through an online crowdsourcing platform and two graduate students selected through personal acquaintance. The second group is composed of three **state-of-the-art LLMs** representative of different architectures. In what follows, we provide a description of the background information of both human and artificial annotators and an overview of the annotation guidelines specific to each group of annotators.

2.2.1 Background Information of Annotators

Human Annotators. As mentioned above, the human annotators of SimilEx are recruited through two different recruitment channels. The first and largest group of participants was recruited via the crowdsourcing platform Prolific[4], while the two further annotators, both graduate students, were recruited through personal acquaintance and volunteered to participate in the task.

The Prolific group initially included a broader set of 317 distinct participants. The platform allows annotators to voluntarily share demographic information such as age, gender, and occupation, providing valuable background context. After a preliminary quality check, we excluded 34 annotators identified as unreliable because they either took too short to complete the annotation task, assigned systematically divergent scores compared to the rest of the participants, failed the control questions or submitted blank answers. This resulted in a final pool of participants who took 18 minutes on average to complete the task[5]. Each sentence pair received a minimum of 5 and a maximum of 7 annotations from different participants. The set of Prolific annotators is quite balanced for gender (51% males), and the average age of annotators is 27.05 ($\pm$6.56). Regarding occupation, 50% of participants indicated that they have a full- or part-time job, around 25% declared themselves unemployed, and the remaining 25% preferred not to disclose their occupational status.

The group of graduate student annotators, also gender-balanced, consisted of two individuals (one male, one female), both aged 23.

4 https://www.prolific.com/
5 The compensation is fair according to the platform: 6.30£/hour.

Large Language Models. The group of LLMs used as annotators is composed of three state-of-the-art models representative of diverse architectures and pre-training and fine-tuning strategies. They include:[6]
i) **Anita (Advanced Natural-based interaction for the ITAlian language)** (Polignano, Basile, and Semeraro 2026), a multilingual (English and Italian) model of the LLaMAntino (Basile et al. 2023) Large Language Models family[7]. Specifically, it is an instruction-tuned version of Meta-Llama-3-8b-instruct (Meta AI 2024), a fine-tuned LLaMA 3 model. LLaMA is a decoder-only transformer model designed for high performance on a wide range of natural language tasks. It is trained on publicly available data and optimised for efficiency and multilingual coverage. Anita integrates this foundation with additional fine-tuning and instruction alignment layers to adapt the model for sentence-level inference tasks. It has been fine-tuned using Supervised Fine-Tuning (SFT) techniques on English and Italian datasets to improve its performance. Additionally, it employs Dynamic Preference Optimization (DPO) to align the model's outputs with user preferences, aiming to avoid inappropriate responses and reduce biases. The model leverages QLoRA for efficient fine-tuning, allowing adaptation to the Italian linguistic structure with improved performance and computational efficiency.
ii) **Claude 3.7 Sonnet**[8], developed by Anthropic, and accessed through Anthropic's API. It employs an optimised transformer architecture with enhanced contextual understanding capabilities.
iii) **GPT-4o** (Hurst et al. 2024), developed by OpenAI. We accessed the model via the OpenAI API. GPT-4o is a large-scale transformer trained using a combination of supervised learning and reinforcement learning from human feedback (RLHF). The latter two LLMs are supposed to deliver reliable performance across various domains, including creative writing, programming, and knowledge-intensive applications.

2.2.2 Annotation Guidelines

Both human annotators and LLMs were asked to rate the similarity of each sentence pair in the dataset using a 5-point Likert scale, where 1 is described as *"Completamente diverse"* (Completely different) and 5 as *"Pressoché identiche"* (Almost identical). This general instruction was adapted to suit the specific characteristics of human annotators and LLMs. Some examples of annotation with similarity scores averaged across annotators are shown in Table 1.

Instructions for Human Annotators. Following the standard procedure of the Prolific platform, crowdworkers were presented with a questionnaire containing 30 sentence pairs plus 2 control pairs, and were allowed to complete multiple questionnaires. On the contrary, the two students were presented with the full list of sentence pairs and were requested to annotate a subset of 897 sentence pairs already annotated by Prolific annotators, without having access to the previously assigned scores. For both groups, no formal definition of similarity was provided, only a few examples of highly similar and highly different pairs along with motivations for the similarity scores, as shown in the annotation instructions provided to the annotators reported in full in Appendix 3.

6 In addition to these models, we also experimented with other smaller, non-proprietary models fine-tuned for Italian. However, their performance on the semantic similarity task was consistently weak, even before taking into account their ability to generate explanations. For this reason, we do not include their results in this paper.
7 https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA
8 https://www.anthropic.com/news/claude-3-7-sonnet

Table 1
Pairs of sentences annotated with similarity scores used as control questions (translations in parentheses).

Sentence 1	Sentence 2	Similarity Score
Sì, ma a che è diretta la sua attività? *(Yes, but what is her activity directed toward?)*	Dunque tu dividi la tua sostanza fra le tue figlie senza riservarti nulla? *(So you divide your estate among your daughters without keeping anything for yourself?)*	1
domandò la signora Valentina, sempre con la medesima soavità nello sguardo e nella voce. *(asked Mrs. Valentina, always with the same gentleness in her gaze and voice.)*	domandò Valentina Michailowna sempre con la stessa soavità di voce. *(asked Valentina Michailowna, always with the same gentle tone of voice.)*	5

This represents the main novelty of our approach compared to the methodology used to create datasets for Semantic Textual Similarity tasks, typically organised within the SemEval evaluation campaign (see, among others, (Agirre et al. 2013; Cer et al. 2017)). These datasets are usually built with clear and specific instructions for annotators, who are explicitly asked to evaluate whether paired text portions refer to the same person, action, or event, or to focus their judgment on similarity types such as the same author, time period, or location.

Instructions for LLMs. When devising the guidelines for the three LLMs considered in this study, we took into account their differing capabilities[9]. Anita requires structured prompting and sequential task execution because it is optimised for specific annotation scenarios and depends on clearly defined instructions and examples to perform accurately. It typically needs to carry out different annotation tasks in separate stages, with each task introduced individually and often supported by illustrative examples. In contrast, proprietary models like GPT-4o and Claude 3.7 Sonnet, trained on massive and diverse multimodal datasets, are able to perform multiple annotation tasks in a single pass, even in zero-shot settings. Their robust instruction-following abilities and deeper contextual understanding mean they can interpret complex prompts and handle nuanced, multi-step tasks without the need for fine-tuning or example-based priming. For this reason, we defined two different prompt strategies for the two types of models. Note that prompts are reported in full in Appendix 3.2.

While ANITA completed the similarity rating and the generation of a corresponding natural language explanation in two separate steps, GPT-4o and Claude 3.7 Sonnet accomplished both tasks in response to a single prompt. At each iteration, a new pair of sentences was appended to Anita's prompt and submitted to the model. To better control the model's output and minimise noise, we instructed it to generate only the specific information requested. For the similarity score annotation task, the model

9 It is worth noting that our goal was to compare LLMs as independent annotators under controlled conditions, rather than to estimate within-model output variance. For this reason, each model was run once, mirroring standard human annotation setups where each annotator provides a single judgment per item. Our analysis focuses on variability across annotators (i.e., across models), rather than variability within repeated outputs of the same model. In line with this design choice, all models were used with their default hyperparameters to ensure consistent and reproducible conditions across systems.

was asked to return only a numeric value between 1 and 5. In contrast, we took full advantage of GPT-4o and Claude 3.7 Sonnet's capability to generate structured content. The models were instructed to operate directly on a CSV format, expanding it by adding and populating two new columns: one for similarity scores and one for explanations. Both annotation tasks were guided by the same prompt.

All models were run in zero-shot and few-shot settings without access to the human-annotated data. While GPT-4o and Claude 3.7 Sonnet showed comparable performance in both zero-shot and few-shot configurations, Anita's annotation quality in the zero-shot setting was found to be inadequate. Consequently, we considered only the outputs from the few-shot setting for all models in our analysis.

2.3 Explanations of Similarity

Among the human annotators, only the two graduate students were asked to justify their similarity judgments. As with the rating task, no formal definition of 'explanation' was given. Instead, they were simply instructed to write a brief justification for each score in the form of a single, concise sentence.

Similarly, all LLMs were prompted to generate a natural language explanation corresponding to the similarity score they had assigned. As described in the previous section, GPT-4o and Claude 3.7 Sonnet completed both the scoring and explanation tasks in a single query using a unified prompt. Anita, by contrast, performed the two tasks in two separate steps, each guided by a distinct prompt. As shown in Appendix 3.2, the second prompt, used for the generation of explanation, repeated the sentence pair along with the assigned similarity score. Anita was instructed to begin each explanation with the word "Spiegazione:" (*Explanation*), a constraint that facilitated the automatic extraction of the relevant content using regular expressions, effectively eliminating any extraneous content.

Note that, unlike the similarity judgments, the explanations were provided for a subset of SimilEx, consisting of 897 randomly selected sentence pairs. Examples of these annotations are shown in Table 2.

3. Similarity Perception

The first analysis of SimilEx focuses on the similarity judgments expressed by the human and LLM (Anita, Claude, and GPT-4o) annotators through the scores. Note that all human scores were considered in this analysis, namely those provided by the Prolific annotators as well as by the two students who also wrote sentence pair explanations.

3.1 Preferred Scores

As a first step in analysing similarity judges, we computed the Pearson correlation between the average similarity scores assigned to each sentence pair and the similarity scores produced by SBERT. The results show a weak correlation for both humans ($r = 0.28$, $p < 0.001$) and LLMs ($r = 0.29$, $p < 0.001$), suggesting that neither group bases its judgments solely on the dimensions captured by SBERT representations. By contrast, the similarity scores assigned by human annotators and those produced by LLMs exhibit a strong and statistically significant positive correlation ($r = 0.614$, $p < 0.001$), highlighting a **notable degree of alignment between LLM-generated and human perceptions of similarity**.

Table 2
Example of a SimilEx sentence pair with explanations written by the two students (*Stud1* and *Stud2*) and by the LLMs, followed by the corresponding similarity score (in round parentheses). English translations of sentences are in square brackets.

Pair 19_6	
Sentence 1:	Stavròġin scoppiò in una risata nervosa, innaturale. [*Stavròġin burst into a nervous, unnatural laugh.*]
Sentence 2:	E si udì una risata ironica, malevola. [*And a mocking, malevolent laugh was heard.*]
Explanations	
Stud1:	Entrambi i protagonisti scoppiano a ridere. (4) [*Both characters burst out laughing.*]
Stud2:	Entrambe le frasi descrivono una risata. (3) [*Both sentences describe a laugh.*]
Anita:	Entrambe le frasi descrivono un suono di risata, ma con diverse caratteristiche e aspetti emotivi, in Frase1 si tratta di una risata del personaggio Stavròġin, in Frase2 di una risata in generale, e le differenze nella descrizione fanno scendere il punteggio di similarità. (3) [*Both sentences describe a laugh-like sound, but with different characteristics and emotional tones; in Sentence 1, it is a laugh from the character Stavròġin, while in Sentence 2, it is a laugh in general, and the differences in description lower the similarity score.*]
GPT-4o:	Le frasi descrivono risate che suggeriscono un'emozione negativa, ma specificano diversi tipi di risate. (3) [*The sentences describe laughs that suggest a negative emotion but specify different types of laughter.*]
Claude:	Entrambe le frasi descrivono una risata con connotazione negativa, differendo solo nei dettagli degli aggettivi usati per qualificarla. (4) [*Both sentences describe a laugh with a negative connotation, differing only in the adjectives used to qualify it.*]

To investigate this alignment more closely, we considered the average similarity score assigned to sentence pairs. Human annotators assigned a mean similarity of 2.40 ($\pm$0.98), while LLMs assigned a slightly higher average score of 2.72 ($\pm$0.82). These values indicate that, overall, both groups tend to perceive the sentence pairs as more dissimilar than similar. However, the LLMs not only assign higher similarity scores on average, but they also exhibit a slightly lower variance and a tendency to produce less polarised judgments.

This pattern becomes clearer when considering the full distribution of average scores, shown in Figure 1, which reports scores averaged across all human annotators and across the three LLMs in order to compare humans and LLMs at the level of annotator typology, without focusing on model-specific differences that are analysed separately in Figure 2. Both humans and LLMs tend to rate most pairs below the midpoint of the scale (i.e., score $<$3), but with different proportions: 76.86% of human-annotated pairs fall in this range, compared to 54.40% for LLMs. Conversely, only 6.82% of pairs received a mean score $\geq$4 from humans, while LLMs reached that threshold in 10.09% of cases. Another contrast lies in the intermediate range between scores 3 and 4: 35.51% of LLM-annotated pairs fall here, compared to just 18.80% for human annotators. This suggests that while humans are more likely to opt for low similarity scores, LLMs make more frequent use of middle-scale ratings, possibly reflecting a more cautious or averaged estimation of similarity in borderline cases.

To better understand the scoring behaviour of different LLMs, we analysed the full distribution of judged sentence pairs across bins of average similarity scores, as illustrated in Figure 2. The results reveal **notable differences in how ANITA, GPT-4o,**

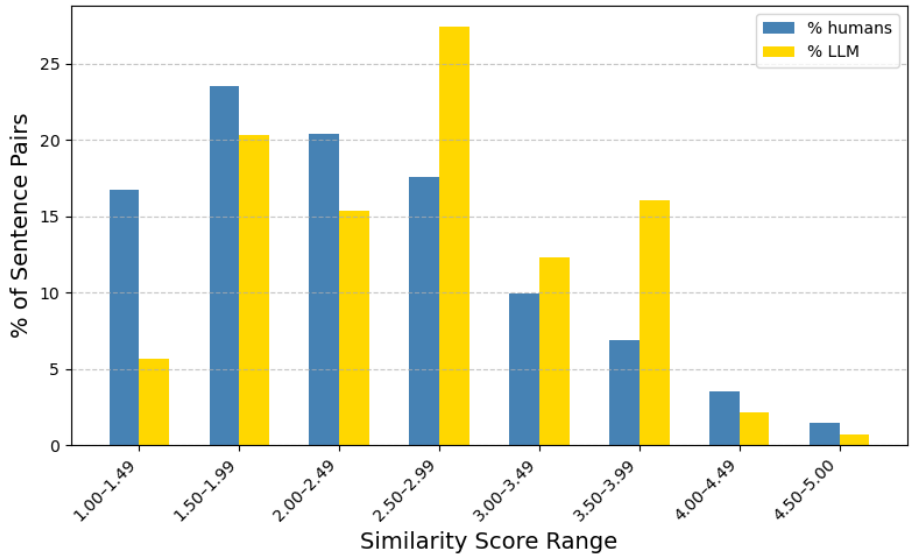

Figure 1
Percentage distribution of mean similarity scores of SimilEx pairs assigned by human and LLM annotators.

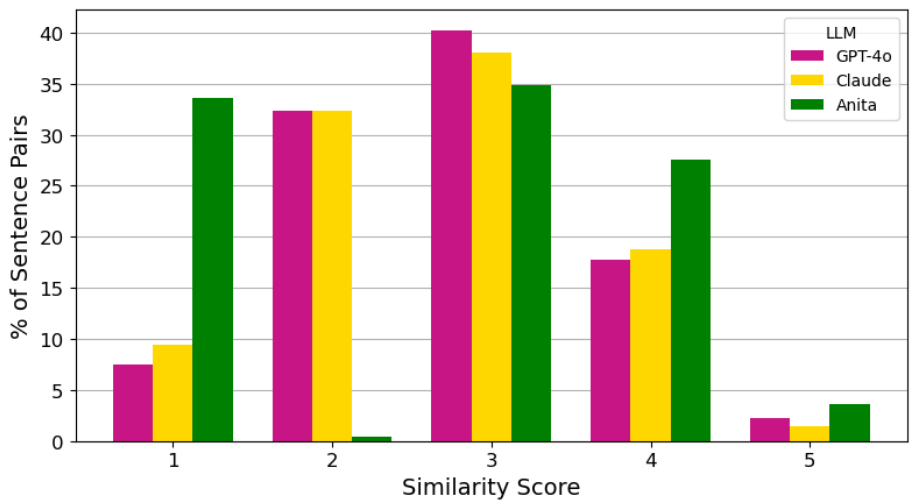

Figure 2
Percentage distribution of similarity scores of SimilEx pairs assigned by each LLM annotator.

and Claude make use of the 1–5 similarity scale. ANITA displays a marked preference for the lowest score (1), assigning it to over 30% of the sentence pairs, a much higher proportion than both GPT-4o and Claude. In contrast, GPT-4o and Claude distribute their scores more evenly across the mid-range of the scale, especially around score 3, which is the most frequently used by all three models. GPT-4o, in particular, assigns a score of 3 to nearly 40% of the pairs, slightly more than Claude and ANITA. Inter-

estingly, ANITA also uses the upper end of the scale (scores 4 and 5) more often than the other models, applying them to around 30% of the sentence pairs. Overall, ANITA exhibits the most balanced distribution of judgments, assigning roughly one-third of the dataset to each of the low, medium, and high similarity bins, while GPT-4o and Claude show a preference to gravitate toward the low-to-mid similarity range, reflecting a more cautious scoring tendency.

3.2 Inter-annotator Agreement

To explore the consistency of the judgments, we examine the inter-annotator agreement (IAA) on the similarity scores using Krippendorff's α coefficient. This metric is suitable when the considered items have a different number of annotations. This meets our scenario, where a sentence pair might be judged by a minimum of 5 to a maximum of 9 human annotators.

Table 3
Agreement scores computed as Krippendorff's α coefficient for different groups of annotators, on all SimilEx pairs and different subsets based on the linguistic properties of the sentences in the pair.

		Sentence length		Subordinates		Tree depth	
	Alpha	**< AVG**	**> AVG**	**Absent**	**Present**	**< AVG**	**> AVG**
Human Annotators	0.352	0.390	0.293	0.533	0.326	0.393	0.32
LLM	0.468	0.481	0.463	0.567	0.457	0.500	0.451
Humans + Anita	0.306	0.342	0.254	0.472	0.283	0.344	0.277
Humans + Claude	0.361	0.401	0.303	0.538	0.335	0.402	0.329
Humans + GPT-4o	0.353	0.393	0.297	0.536	0.325	0.397	0.321

Overall agreement across annotators. As shown in the first column of Table 3, the global IAA between human annotators, computed considering all pairs in the dataset, is 0.352. This level of agreement falls within the *fair* range (Landis and Koch 1977), which is expected given the inherent subjectivity of the task. Nevertheless, it reflects a general tendency for human annotators to converge on many items. In contrast, the agreement among Anita, Claude, and GPT-4o (row *LLM* in Table 3), is higher, reaching a value of 0.468, which corresponds to a *moderate* level of agreement. This suggests that **LLMs tend to produce more consistent similarity scores compared to human annotators**. As suggested by the preferred score distribution analysis (see Figure 2), this result might be due to the models' tendency to assign to pairs the midpoint of the scale (score 3) more frequently than humans. This central score might often be assigned to the same sentence pairs by all models, contributing to the overall increase in agreement.

Alignment of individual LLMs with human judgments. To evaluate which LLM aligns most closely with human judgments, we computed Krippendorff's alpha by integrating each model's similarity scores with those of the human annotators. The results, presented in the last three rows of Table 3, show that for all models, the resulting agreement remains within the *fair* range (0.3–0.4), comparable to the IAA observed among humans alone. Nonetheless, some variation emerges depending on the model considered: Anita yields the lowest agreement with human annotators ($\alpha = 0.306$), while Claude achieves the highest consistency ($\alpha = 0.361$); GPT-4o, while close in performance to Claude, falls slightly lower at 0.353. These results suggest that Claude's similarity judgments are

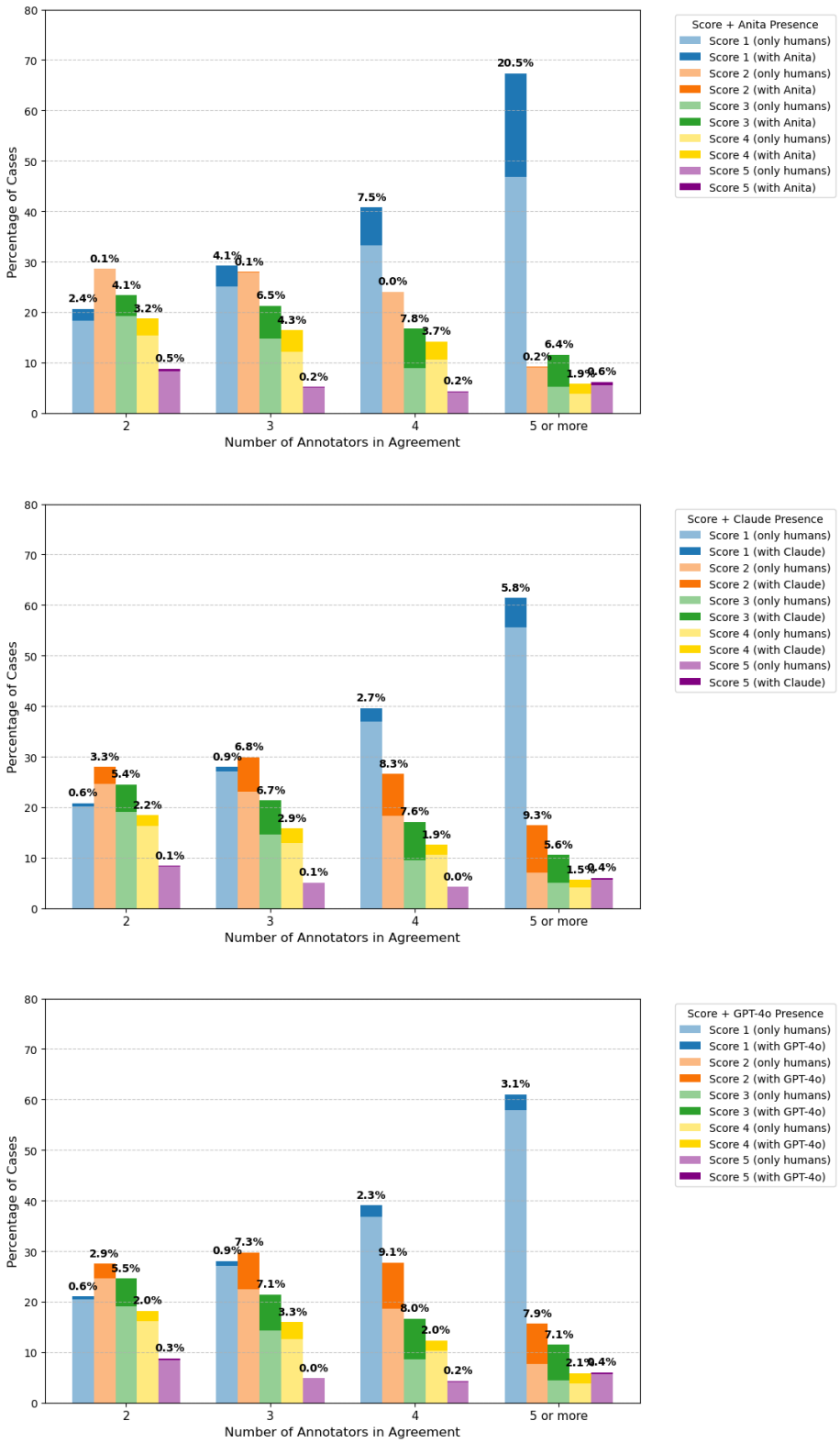

Figure 3
Percentage distribution of similarity scores with respect to the number of annotators (all humans plus one model for each graph) in agreement on the pair.

more in line with those of human annotators, while Anita deviates more substantially. This pattern highlights differences in how each model interprets sentence similarity, potentially influenced by their underlying architecture or training data.

Patterns of agreement and similarity score. To explore this further, we grouped sentence pairs based on the number of annotators (humans and one model at a time) who assigned them the same similarity score. Specifically, we considered groups where 2, 3, 4, or 5 or more annotators agreed. This grouping offers a general sense of whether each sentence pair tended to produce consensus or disagreement among annotators. Figure 3 shows how the sentence pairs are distributed across these levels of agreement. The group sizes vary quite a bit: across all models, about 43% of sentence pairs had only 3 or fewer annotators in agreement. In contrast, around 26% of pairs received identical scores from 5 or more annotators.

When agreement is low on a pair (2-3 annotators in agreement), the assigned scores are more evenly spread across the five-point scale. This suggests that disagreement does not stem from a clear binary split between similar and dissimilar judgments, but rather from a broader range of interpretations: some annotators may view the pair as highly similar, while others see it as very different. The presence of a language model slightly shifts the pattern. Specifically, scores 2 and 3 become somewhat more frequent when GPT-4o or Claude are included. This suggests these models sometimes align with annotators who assign intermediate scores, contributing to modest increases in score variability for ambiguous pairs.

In high agreement cases (i.e., when 5 or more annotators agree), however, the score most frequently assigned is 1, indicating a **common convergence around dissimilarity judgments**. This trend holds even when a language model is added to the annotator pool. As Figure 3 illustrates, score 1 accounts for over 60% of the highly agreed-upon cases. This is especially true when the model is Anita, but the broader trend remains consistent regardless of the considered model: higher agreement is generally associated with lower similarity scores. Interestingly, while Anita shows agreement on two scores in particular (i.e., 1 and 3), Claude and GPT-4o, which show more similar distributions to each other, tend to converge more often on the mid-scale scores (scores 2 and 3).

This pattern is further supported by a **statistically significant negative Pearson correlation between the number of agreeing annotators and the average similarity score of each pair**. The correlation is $r = -0.344$ ($p < 0.001$) for human annotators alone and remains comparably negative with the inclusion of Anita ($r = -0.356$, $p < 0.001$), Claude ($r = -0.314$, $p < 0.001$), and GPT-4o ($r = -0.225$, $p < 0.001$). While all models follow the general trend observed for humans, GPT-4o shows a weaker correlation, indicating a reduced tendency to associate high annotator agreement with judgments of dissimilarity.

3.3 Agreement, Style and Similarity

In Section 2.1, we noted that SimilEx exhibits a high degree of stylistic variability within sentence pairs. This observation led to the hypothesis that the stylistic traits of the sentences might influence the similarity judgments they receive. As a general remark, we found that style minimally affects pairs' similarity: the Pearson correlation between the similarity scores and the distribution of stylistic properties is either non-significant ($p > 0.05$) or extremely low ($r < 0.1$). However, a more in-depth analysis of specific stylistic properties revealed a **nuanced relationship between style and the consistency of judgments**. As above, we carried out this analysis for human annotators, LLMs

alone, and including one LLM at a time in the pool of human annotators. Results of the agreement scores are reported in the three final columns of Table 3.

Sentence length. As a first linguistic feature to monitor, we considered sentence length, a raw yet informative feature reflecting stylistic variation. Since we are dealing with paired sentences, we cannot focus on the absolute length in tokens of each individual sentence, but rather on a synthetic measure that captures the difference in length between the two sentences in a pair. For example, in a pair where one sentence is 12 tokens long and the other is 30 tokens long, we account for their absolute length difference, which is 18 tokens.[10] Contrary to our expectations, the difference between the length of the sentences did not impact the similarity scores assigned by annotators. While the Pearson correlation between sentence length difference and rating variance was negligible for both humans ($r = 0.05$) and LLMs ($r = 0.03$, $p > 0.05$), inter-annotator agreement depicts a more nuanced picture. Specifically, when the length difference was below the average difference in the dataset (i.e., 17 tokens, see Table 3, column *Sentence length*, sub-columns *AVG*), human annotators exhibited a moderate level of agreement ($\alpha = 0.390$), which dropped to fair agreement ($\alpha = 0.293$) when the length difference within the pair was greater. This suggests that sentence length does exert a small but measurable effect on human consistency, likely reflecting increased processing effort or mismatched sentence complexity. In contrast, LLMs showed a higher and more stable agreement across the same split ($\alpha = 0.481$ vs. 0.463), indicating that their judgments are less sensitive to surface-level stylistic variation. The hybrid conditions (humans and model) show a similar pattern to humans, though with slight differences: for example, with GPT-4o, agreement dropped from 0.393 to 0.297 across the two sets, nearly mirroring the human-only shift.

Subordinate clauses. More structurally informative features, such as the presence of subordinate clauses, however, have a clearer and stronger impact on annotation consistency. Among humans, agreement drops sharply from $\alpha = 0.533$ for pairs where neither sentence contains a subordinate clause, to 0.326 when both sentences contain at least one subordinate (see column *Subordinates* in Table 3). A similar decline is seen in all hybrid and LLM-only setups. Specifically, LLMs alone achieve higher overall agreement ($\alpha = 0.567$ without subordinates, 0.457 with subordinates), showing the same trend with a notable drop of over 0.1. Hybrid setups follow suit: for example, agreement with Anita falls from 0.472 to 0.283, with Claude from 0.538 to 0.335, and with GPT-4o from 0.536 to 0.325. These drops, consistently ranging from about 0.1 to nearly 0.2, confirm that the presence of subordinate clauses systematically lowers agreement, regardless of whether judgments are made by humans, models, or a mix of both. This suggests that increased syntactic complexity introduces ambiguity or interpretive divergence that even state-of-the-art models cannot fully resolve.

Syntactic tree depth. Likewise, linguistic features that model the global syntactic structure of a sentence lead to lower agreement. For example, human annotators show higher consistency when the absolute difference between the depth of the syntactic trees of the paired sentences is low, i.e. below the average value of 1.98 (α=0.393), compared to high (α=0.320), a pattern mirrored by all models (see *Tree depth* column in Table 3). These results suggest that while LLMs are generally more robust to surface-level variation

10 We use the absolute difference because the order of the sentences in the pair is not relevant.

like length, both humans and models are similarly affected by complexity related to the overall syntactic structure of the sentences.

4. Similarity Explanation

As explained in Section 2.3, natural language explanations for the assigned similarity scores were provided by two human annotators (i.e. the two undergraduate students, henceforth referred to Stud1 and Stud2) and by the three LLMs for a subset of 897 SimilEx sentence pairs. In this section, we examine the writing style and semantic content of the explanations, with the primary goal of highlighting similarities and differences in the motivations provided by human annotators and LLMs to justify their similarity judgments.

4.1 Linguistic Style of Explanations

We explored the style of explanation relying on the linguistic profiling method described in Section 2.1. By comparing the results of the stylistic analysis, distributed along with the complete dataset (see Appendix 1), for the two students and the three LLMs, several notable differences emerge across multiple linguistic dimensions.

Table 4
Mean (and standard deviation) of a selection of linguistic feature distribution in natural language explanations written by the two human annotators (*Stud1* and *Stud2*) and by the three considered LLMs.

Feature	Stud1	Stud2	Anita	Claude	GPT-4o
sent len.	6.37 (±3.94)	7.74 (±5.07)	40.60 (±9.19)	29.90 (±6.03)	25.70 (±5.91)
word len.	6.51 (±2.14)	6.13 (±2.32)	4.86 (±0.39)	5.21 (±0.52)	5.10 (±0.55)
% ADJ	18.69 (±20.74)	14.41 (±20.36)	8.14 (±4.14)	11.27 (±5.26)	8.13 (±5.12)
% ADV	13.05 (±21.13)	11.98 (±20.82)	3.10 (±3.32)	2.67 (±3.08)	1.98 (±2.89)
% CCONJ	0.84 (±3.12)	0.97 (±3.09)	4.75 (±2.75)	6.20 (±3.75)	7.03 (±4.06)
% PUNCT	0.92 (±3.46)	0.35 (±1.86)	9.66 (±2.92)	11.14 (±3.83)	9.90 (±2.79)
% verbal root	58.31 (±49.33)	62.21 (±48.47)	99.67 (±5.78)	96.88 (±17.40)	98.33 (±12.83)
avg clause len	3.81 (±3.46)	5.44 (±4.62)	13.03 (±5.49)	12.13 (±6.32)	12.52 (±5.82)
max link len	3.11 (±2.24)	3.11 (±2.62)	21.03 (±9.65)	15.78 (±6.32)	12.80 (±5.22)
n. prep chains	0.36 (±0.56)	0.37 (±0.55)	2.45 (±1.31)	1.41 (±1.03)	1.36 (±0.98)
subord prop.	20.95 (±32.21)	12.32 (±26.95)	56.54 (±22.41)	41.51 (±29.98)	29.49 (±29.49)

Table 4 presents a selection of the most distinctive linguistic features. Starting with raw textual properties, we note that **the two students tend to write shorter sentences compared to LLMs**, which generated longer explanations. This is also quite evident by inspecting the examples reported in Table 2, where model-generated explanations are generally longer and more syntactically articulated. Such a broad difference is further reflected in two main stylistic tendencies that differentiate human-written from model-generated explanations, modelled by the distribution of other linguistic features, many of which are closely related to sentence length.

First, the LLM-generated explanations include a higher percentage of coordinating conjunctions (% *CCONJ*) and punctuation marks (% *PUNCT*), reflecting a syntactic preference for longer sentences articulated through coordination and possibly subordination. In contrast, the human-written explanations display a higher amount of modifiers, as indicated by the greater use of adjectives (% *ADJ*) and adverbs (% *ADV*) among all parts of speech. A closer analysis reveals that this trend is mostly due to a recurring

pattern shared by the two human annotators: the sentence "Completamente diverse" (*Completely different*), composed of one adverb and one adjective. This expression is frequently used to justify low similarity scores, accounting for 20.51% of the explanations written by Stud1 and 22.19% of those by Stud2. This observation complements the earlier finding that human annotators tend to assign lower similarity scores overall.

Second, **human explanations are more frequently characterised by a nominal style**. This is evidenced by the lower percentage of sentences with a verbal root, calculated over all syntactic roots. This percentage is notably low when compared to the distribution in the ISDT (Bosco, Montemagni, and Simi 2013), the largest Italian Treebank, where the distribution is 85.73%. LLMs, by contrast, appear to adopt a more 'naturalistic' or conversational style when asked to justify their similarity judgments. As observed above for the explanation "Completamente diverse" (*Completely different*), this is largely due to the recurrence of short expressions exhibiting verb ellipsis in human explanations. For instance, both students frequently use short nominal phrases such as "Ambientazione nello stesso luogo" (*Set in the same place*), "Pressoché identiche" (*Almost identical*), "Stessa professione dei soggetti (*Same profession of the subjects*). Indeed, while the percentage of explanations with verb ellipsis is 35.79% for Stud1 and 34.45% for Stud2, the distribution is markedly lower in the model-generated explanations: 0.67% for Anita, 1% for GPT-4o, and 2.12% for Claude. LLMs, in fact, tend to employ more explicit syntactic structures, often relying on patterns such as "Entrambe le frasi [VERBO]" (*Both sentences* [VERB]), "Entrambe [VERBO]" (*Both* [VERB]), "Le frasi sono completamente diverse:" (*The sentences are completely different:*), and "Le due frasi [VERBO]" (*The two sentences* [VERB]).

Related to this characteristic, we can observe that the average clause length per sentence (*avg clause len*)[11] is significantly shorter in the students' explanations. Additional features further allow modelling differences in the local syntactic structures employed by humans and models. Human-written explanations tend to exhibit shorter syntactic dependency relations (*max link len*), fewer prepositional chains (*n. prep chains*), and a lower number of subordinate clauses (*subord prop.*). Together, these features suggest a syntactic style characterised by simplicity, with fewer intervening words between syntactic heads and dependents, and typically consisting of a single main clause with minimal use of subordination.

4.2 Content of Explanations

The comparative analysis of the content of human- and model-generated explanations was first conducted by calculating the average cosine similarity between SBERT embeddings of explanations, across the subset of 897 SimilEx sentence pairs. The results are shown in Figure 4, which visualises a similarity matrix, reporting all pairwise comparisons based on the average cosine similarity scores computed from SBERT embeddings of each annotator's explanations.

A first observation is that higher cosine similarity scores are found within 'homogeneous' groups, including either two human annotators or the three LLMs. This indicates that **the content of the explanations is more similar within annotators' type (human vs. model) than across types**. Notably, the similarity is higher between the explanations

11 The feature is calculated in terms of the average number of tokens per clause, where the number of clauses corresponds to the ratio between the number of tokens in a sentence and the number of either verbal or copular heads.

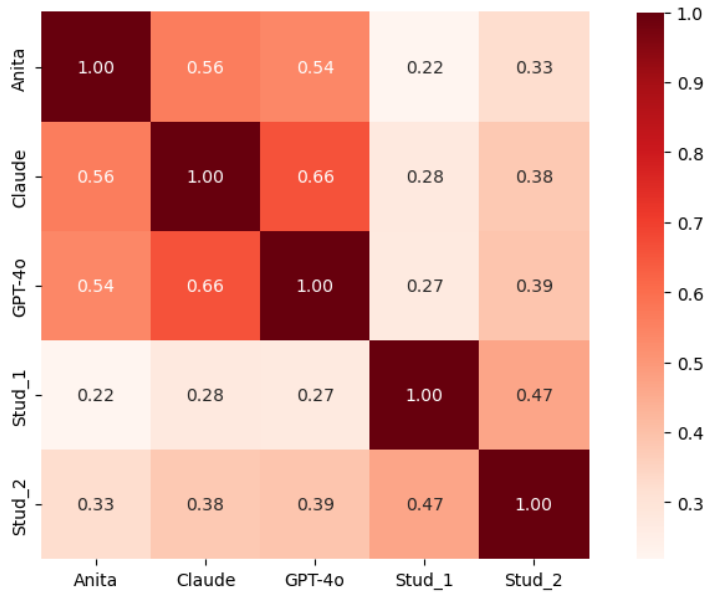

Figure 4
Pairwise average cosine similarity of human- and model-generated explanation computed based on SBERT embeddings.

generated by the LLMs: the highest similarity value is observed between Claude and GPT-4o (0.66), followed by Claude and Anita (0.56), and GPT-4o and Anita (0.54). In contrast, the average similarity between the two human annotators, Stud1 and Stud2, is lower (0.47). These results suggest that **LLMs are more consistent among themselves** in the explanations they provide to justify their similarity judgments. On the other hand, the lower similarity between the two students may be due to a greater variation in the aspects they considered relevant. This supports the hypothesis that the explanation task is more subjective for human annotators, who appear to rely on more diverse and individually-driven reasoning strategies, resulting in less homogeneous content. As proof, consider the examples reported in Tables 5 and 6, where students, unlike the LLMs, focused on diverse aspects of the paired sentences while they assigned either similar (see pair 2_13) or different (see pair 24_23) similarity scores.

The example reported in Table 6 highlights another relevant aspect emerging from the first level of explanation content analysis, i.e. SBERT-based similarity scores tend to be higher when Stud2, rather than Stud1, is involved in pairwise comparisons with LLM-generated explanations. In this case, Stud2 focuses on the same similarity aspect identified by the three LLMs (i.e. that both sentences refer to Prince Andrea, who was summoned by the *generalissimo*).

This result emerges also from the 2-dimensional Principal Component Analysis (PCA) shown in Figure 5, which was conducted on each individual explanation.[12] As shown in the plot, data points cluster into two distinct groups: one representing

12 PCA is a classical data analysis method that reduces the dimensionality of the data while retaining most of its variance by identifying the principal components—axes along which variation is maximal (Jolliffe and Cadima 2016).

Table 5
Example of a SimilEx sentence pair with explanations followed by the corresponding similarity score (in round parentheses). English translations of sentences and explanations are in square brackets.

Pair 2_13	
Sentence 1:	Ma l'amore con un marito inquinato dalla gelosia e da ogni sorta di difetti non era più per lei. [*But love with a husband tainted by jealousy and all sorts of flaws was no longer for her.*]
Sentence 2:	Era forse, semplicemente, un sentimento di gelosia: egli era talmente avvezzo all'amore di lei, che non poteva ammettere che ella potesse amarne un altro. [*Perhaps it was simply a feeling of jealousy: he was so used to her love that he could not accept that she might love someone else.*]
Explanations:	
Stud1:	Parlano di uomini gelosi. (2) [*They talk about jealous men*]
Stud2:	Nel primo caso il focus della frase è la moglie, nella seconda lo è il marito (2) [*In the first case the focus is on the wife, in the second it is on the husband.*]
Anita:	Entrambe le frasi discutono di un rapporto amoroso, focalizzandosi sulla gelosia del marito, con minime differenze nella presentazione dei contenuti. (4) [*Both sentences discuss a romantic relationship, focusing on the husband's jealousy, with minimal differences in how the content is presented.*]
GPT-4o:	Entrambe le frasi trattano il tema della gelosia nel contesto di una relazione, ma da prospettive differenti: una il rifiuto dell'amore, l'altra la percezione di gelosia dell'uomo. (3) [*Both sentences deal with the theme of jealousy in the context of a relationship, but from different perspectives: one about the rejection of love, the other about the man's perception of jealousy.*]
Claude:	Entrambe le frasi trattano il tema della gelosia in una relazione amorosa, ma da prospettive diverse: la prima dal punto di vista di una donna insofferente, la seconda da quello di un uomo geloso. (3) [*Both sentences deal with the theme of jealousy in a romantic relationship, but from different perspectives: the first from the point of view of an exasperated woman, the second from that of a jealous man.*]

the content of the explanations written by the two human annotators, and the other representing the model-generated explanations. Notably, these two clusters do not overlap, reinforcing the distinction observed in the similarity analysis. Within this distribution, the data points corresponding to Stud2 are located closer to the cloud of model-generated explanations, while those from Stud1 are more distant, further supporting the previously observed differences between the two human annotators.

Explanation similarity and similarity judgments. The content analyses presented above were further extended to address a final research question: whether a relationship exists between the similarity of the explanations, in terms of their content, and the similarity perception scores assigned to SimilEx sentence pairs by human annotators and LLMs. To this end, we computed the Pearson correlation between SBERT scores and the absolute difference in the similarity judgments, considering all pairs of annotators. Interestingly, the trend varies depending on whether the pair includes human or LLM annotators. For the two human annotators, we observe a moderate negative relationship, equal to $r = -0.459$ ($p < 0.001$). This indicates that the more semantically similar the explanations are, the smaller the difference in the students' similarity judgments. Specifically, students' explanations tend to be more similar when the similarity scores assigned by both of them are lower (i.e. 1 or 2), as in the example reported in Table 7. When LLMs are involved, a negative correlation still emerges, though it is weaker:

Table 6
Example of a SimilEx sentence pair with explanations followed by the corresponding similarity score (in round parentheses). English translations of sentences and explanations are in square brackets.

Pair 24_23	
Sentence 1:	L'aiutante, accostatosi al principe Andrea, che era sempre sulla scala, lo invitò a colazione; ma di lì a mezz'ora, il principe fu di nuovo chiamato presso il generalissimo. [*The aide, having approached Prince Andrei, who was still on the stairs, invited him to breakfast; but half an hour later, the prince was summoned again to the generalissimo.*]
Sentence 2:	Il principe Andrea era di servizio in quel giorno presso il generalissimo. [*Prince Andrei was on duty that day with the generalissimo.*]
Explanations:	
Stud1:	Pressochè identiche (5) [*Almost identical.*]
Stud2:	Entrambe le frasi parlano del principe Andrea convocato dal generalissimo. (4) [*Both sentences talk about Prince Andrei being summoned by the generalissimo.*]
Anita:	Le due frasi si riferiscono al principe Andrea, ma solo la seconda fornisce informazione sulla sua occupazione del giorno, mentre la prima fornisce una sequenza di eventi in cui il principe Andrea è coinvolto, con un riferimento a un dettaglio di tempo e luogo. (3) [*The two sentences refer to Prince Andrei, but only the second provides information about his duty on that day, while the first describes a sequence of events involving the prince, with a reference to time and place.*]
GPT-4o:	Entrambe le frasi parlano del principe Andrea e del suo servizio presso il generalissimo, sebbene la prima offra più dettagli sugli eventi del giorno. (4) [*Both sentences refer to Prince Andrei and his service with the generalissimo, although the first provides more details about the day's events.*]
Claude:	Entrambe le frasi riguardano il principe Andrea e il suo rapporto con il generalissimo, ma la prima contiene dettagli specifici di un episodio mentre la seconda è una dichiarazione generale. (3) [*Both sentences concern Prince Andrei and his relationship with the generalissimo, but the first contains specific details of an episode, while the second is a general statement.*]

$r = -0.25$ ($p < 0.001$) between Anita and GPT-4o, and $r = -0.17$ ($p < 0.001$) between Anita and Claude. In contrast, no statistically significant correlation is observed between Claude and GPT-4o. This pattern appears to be confirmed by the examples reported in Tables 2, 5, 6 and 7, which show that, regardless of the similarity score assigned, the LLMs tend to provide explanations that are not only quite similar to each other, often focusing on the same aspects of the sentence pair, but also relatively superficial. In particular, their explanations frequently resemble brief summaries of the two sentences, often repeating several words from the sentence pair itself. This behaviour contrasts with that of human annotators, whose explanations more concisely reflect the individual motivations behind their judgments.

5. Conclusion and Future Work

In this paper, we extend the study presented in (Alzetta et al. 2024) by investigating how human annotators and Large Language Models assign and justify semantic similarity judgments in a Semantic Textual Similarity (STS) task. To this end, we introduce the results of a new annotation campaign aimed at collecting both similarity scores and

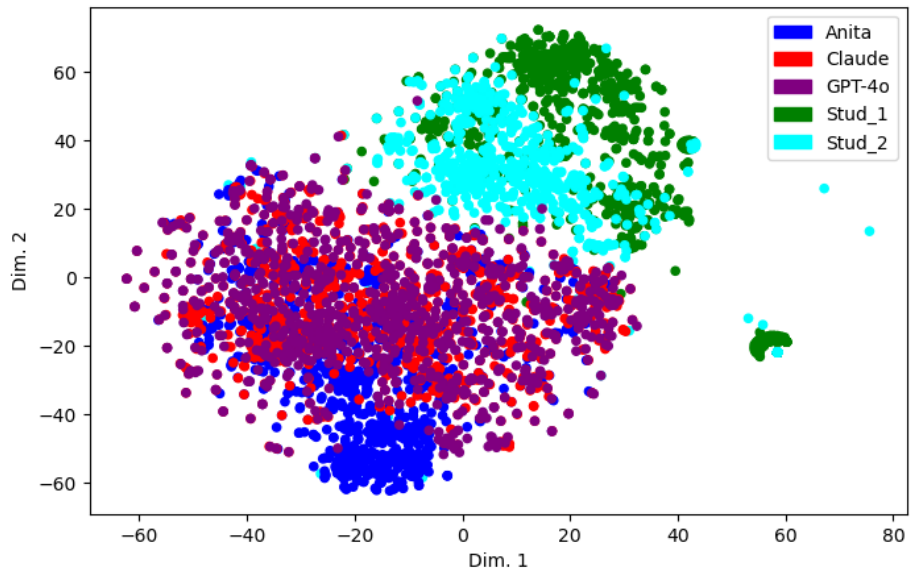

Figure 5
2-dimensional PCA projection of the SBERT embeddings representing human- and model-generated explanations.

natural language explanations (i.e., justifications for the assigned scores) produced by three LLMs: Anita, Claude, and GPT-4o.

A key contribution of this work lies in the collected human- and LLM-generated annotations that are directly comparable. This comparability is ensured by keeping the annotation guidelines as parallel as possible, with the same instructions provided to both human annotators and LLMs. At the same time, the extent to which LLMs adhere to these shared instructions varies, reflecting differences in their underlying optimisation strategies (as discussed in Section 2.2.2).

Taken together, our findings provide a comprehensive account of how human annotators and LLMs compare in assigning and justifying semantic similarity judgments, while the following sections offer a more detailed analysis of these aspects. This study represents, to the best of our knowledge, the first direct comparison between human and LLM annotations for Italian in an STS task enriched with natural language explanations and based on a literary text genre that has so far received limited attention in STS research.

LLMs vs. human annotators assign higher similarity scores. Our comparative analysis reveals that, despite a moderately high positive correlation between the similarity scores assigned by human annotators and LLMs ($r = 0.614$, $p < 0.001$), LLMs tend to assign overall higher similarity scores, with a tendency toward the middle score. In contrast, human annotators more frequently assign lower scores. Differences also emerge across the three LLMs. Anita shows a larger percentage of SimilEx sentence pairs rated as either very dissimilar (similarity score=1) or highly similar (similarity score = 4 or 5). Claude and GPT-4o, on the other hand, adopt a more neutral scoring behaviour, favouring lower-middle and middle-range scores (similarity scores=2 or 3). This distinct pattern in Anita's behaviour recurs in other comparative analyses conducted throughout the study.

Table 7
Example of a SimilEx sentence pair with explanations followed by the corresponding similarity score (in round parentheses). English translations of sentences and explanations are in square brackets.

Pair 12_32	
Sentence 1:	Era un cane abbastanza grosso, nero con macchie bianche, da cortile, non vecchissimo, dagli occhi intelligenti e dalla coda fioccosa. [*It was a fairly large dog, black with white spots, a yard dog, not very old, with intelligent eyes and a bushy tail.*]
Sentence 2:	Infatti, il mio cane stava immobile davanti a una macchia di querce all'estremità della stretta gola che sboccava sulla strada. [*In fact, my dog was standing still in front of a patch of oaks at the end of the narrow gorge that opened onto the road.*]
Explanations:	
Stud1:	Menzionano cani (2) [*They mention dogs.*]
Stud2:	Entrambe le frasi parlano di cani (2) [*Both sentences are about dogs*]
Anita:	La similarità è bassa perché le due frasi non trattano lo stesso argomento, la prima descrive un cane, la seconda un cane in un contesto diverso, descrivendo il luogo in cui si trova. (1) [*The similarity is low because the two sentences do not deal with the same topic. The first describes a dog, the second a dog in a different context, describing the place where it is.*]
GPT-4o:	Entrambe le frasi parlano di un cane, ma la prima descrive l'aspetto e le caratteristiche del cane, mentre la seconda si concentra su un'azione specifica in un contesto diverso. (2) [*Both sentences talk about a dog, but the first describes the dog's appearance and characteristics, while the second focuses on a specific action in a different context.*]
Claude:	Entrambe le frasi parlano di un cane, ma la prima ne descrive l'aspetto fisico mentre la seconda ne descrive la posizione in un contesto esterno. (3) [*Both sentences are about a dog, but the first describes its physical appearance, while the second describes its position in an external context.*]

LLMs vs. human annotators show a higher inter-annotator agreement. Possibly as a consequence of their shared tendency toward middle-range scores, LLMs exhibit higher inter-agreement than human annotators. Specifically, while agreement among the human participants falls within the *fair* range (Krippendorff's $\alpha = 0.352$), the agreement among the LLMs reaches a *moderate* level ($\alpha = 0.468$). Complementary results are observed when computing agreement scores for mixed pools composed of the two human annotators and one LLM: agreement is consistently lowest when Anita is included in the pool. Further differences between human and LLM annotators emerged when examining the relationship between inter-annotator agreement and *i)* the assigned similarity scores and *ii)* the linguistic features of the SimilEx sentence pairs. The first analysis revealed that high agreement, whether among humans alone or in mixed pools, tends to occur for lower similarity scores, especially score 1, suggesting a shared perception of dissimilarity. While all mixed pools follow this general trend, those including GPT-4o show a weaker correlation, indicating a reduced tendency to align with the human pattern of agreement on dissimilar judgments. Different patterns were observed in relation to linguistic features. In particular, agreement among human annotators decreases when the length difference between the paired sentences increases, while agreement between LLMs is less affected by such variation. However, both humans and LLMs exhibit lower agreement when sentence pairs contain more complex structural features, such as subordinate clauses, indicating that syntactic complexity poses challenges to consistent similarity assessment across all annotators.

LLMs vs. human annotators write longer and more similar natural language explanations. When prompted to justify their similarity scores through natural language explanations, LLMs tend to generate longer and more elaborate motivations compared to those written by human annotators. In terms of writing style, human explanations are often shorter in terms of number of tokens and exhibit a nominal structure, frequently omitting verbs and employing minimal subordination. In contrast, LLMs generally generate explanations with full verbal sentence structures and more canonical syntax. A further key difference lies in the content of the explanations. Interestingly, LLM-generated explanations tend to be more similar to each other than those produced by human annotators. This suggests that score justification is a more subjective process for humans, who vary in the aspects they consider relevant, whereas LLMs appear to consistently focus on similar aspects of the SimilEx sentence pairs when formulating their motivations.

The findings from this study open several promising avenues for future research. One interesting direction involves expanding the sentence pairs to a multilingual setting. In particular, a future study could involve constructing a multilingual dataset from the same late 19th-century novels translated into multiple languages beyond Italian. Such a resource could be annotated by various LLMs, enabling a comparative analysis of both the similarity scores assigned across different languages and the accompanying natural language explanations. While large-scale multilingual annotation may be challenging to carry out with human annotators, LLMs offer the possibility to explore these aspects more systematically. This would provide a useful resource for analysing how LLMs assign and justify similarity judgments across languages.

Beyond its primary role as a resource for comparing human and LLM behaviour in assigning and justifying semantic similarity judgments, the SimilEx extension also enables the exploration of hybrid annotation settings, where LLM-generated explanations are evaluated by human annotators to assess their reliability and perceived adequacy. In this perspective, it can be used to investigate whether LLM-produced scores and explanations can support annotation processes in a cost-effective way, possibly complementing human annotations, an approach already explored for other natural language inference tasks, such as (Chen et al. 2025). Additionally, the LLM-generated explanations we collected can support future studies aimed at categorizing explanation content, as proposed by (Hong et al. 2025). Such analyses may enable fine-grained comparisons of the specific aspects of sentence pairs that humans and LLMs focus on to justify their similarity judgments.

Appendix

1. Data Distribution

The complete SimilEx dataset is freely available under the CC BY-NC-SA license at `http://www.italianlp.it/resources/` along with the results of the stylistic analysis of both paired sentences and the natural language explanations provided by the two students.

Specifically, on the dedicated page, you can find the SimilEx dataset, which includes the similarity scores assigned by each annotator for the full set of 2,112 sentence pairs, as well as the corresponding natural language explanations for the subset of 897 pairs.

2. Linguistic Features

The set of linguistic features derived by Profiling-UD is extracted from different levels of linguistic annotation and captures a wide number of linguistic phenomena. The features can be grouped as follows:

- **Raw text**
  - Number of tokens in sentence;
  - Average characters per token.
- **Morphosyntactic information**
  - Distibution of UD POS;
  - Lexical density.
- **Inflectional morphology**
  - Distribution of lexical verbs and auxiliaries for inflectional categories (tense, mood, person, number).
- **Verbal Predicate Structure**
  - Distribution of verbal heads and verbal roots;
  - Average verb arity and distribution of verbs by arity.
- **Global and Local Parsed Tree Structures**
  - Average depth of the whole syntactic trees;
  - Average length of dependency links and of the longest link;
  - Average length of prepositional chains and distribution by depth;
  Average clause length.
- **Relative order of elements**
  - Distribution of subjects and objects in post- and pre-verbal position.
- **Syntactic Relations**
  Distribution of dependency relations.
- **Use of Subordination**
  - Distribution of subordinate and principal clauses;
  - Average length of subordination chains and distribution by depth;
  - Distribution of subordinates in post- and pre-principal clause position.

3. Annotation Instructions

3.1 Instructions for Humans in Italian

Stai per svolgere un questionario nel quale ti verrà chiesto di valutare se due frasi sono fra di loro simili o diverse. Per farlo, ti mostreremo delle coppie di frasi estratte da romanzi e ti chiederemo di assegnare ad ogni coppia un punteggio compreso fra 1 e 5. Usa 1 per dire che le due frasi sono fra loro completamente diverse; Usa 5 per dire che sono pressoché uguali. Gli altri punteggi ti serviranno per valutare i casi intermedi.
Due frasi possono dirsi uguali o diverse sulla base di diversi elementi. Ecco alcuni esempi per aiutarti nella valutazione.

Coppie di frasi diverse (punteggio 1).

Esempio 1:

a) Io desidererei tanto non sentire così intensamente e non prendermi tanto a cuore tutto

quello che succede.
b) Sì, non sono in me, sono tutta nell'aspettativa e vedo tutto un po' troppo facile.
Esempio 2:
a) Anche il vecchio principe t'è affezionato.
b) - Non mi sembra di averveli chiesti, - scattò il principe irritatissimo.
Esempio 3:
a) Il the veramente era del color della birra, ma io ne bevvi un bicchiere.
b) Ma non passò neanche un minuto, che la birra gli diede alla testa e per la schiena gli corse un leggero e perfin piacevole brivido.

Fai particolare attenzione agli esempi 2 e 3: anche se le frasi hanno delle parole in comune (come 'principe' e 'birra' negli esempi) non è detto che siano uguali!

Coppie di frasi molto simili (punteggio 5).
Esempio 1:
a) Signori della giuria, la psicologia è a doppio taglio e anche noi siamo in grado di comprenderla.
b) Vedete allora, signori della giuria, dal momento che la psicologia è un'arma a doppio taglio, permettetemi di occuparmi del secondo taglio e vediamo che cosa viene fuori.
Esempio 2:
a) "Un rettile divorerà l'altro", aveva detto il giorno prima Ivan, parlando con rabbia del padre e del fratello.
b) "Un rettile divorerà l'altro, quella è la fine che faranno!".
Esempio 3:
a) Ma una volta deciso, continuò con la sua voce stridula, senza timori, senza esitazioni e sottolineando alcune parole.
b) Parlava rapido, senza fermarsi un momento, senza la minima esitazione, quasi rimproverasse a sè stesso di aver tanto indugiato a mettere Marianna a parte di tutti i suoi segreti, quasi scusandosi presso di lei.

Gli esempi 1 e 2 riportano frasi che non solo contengono molte parole in comune ma sono simili anche per quanto riguarda la scena descritta. Nel terzo esempio, entrambe le frasi descrivono una persona intenta a parlare in modo svelto e deciso. Possiamo dire che in questi esempi l'alta similarità fra le frasi è data dal fatto che, ad eccezione di alcuni dettagli, esse descrivono scene o immagini molto simili, anche se si svolgono in contesti diverse.

3.1.1 Translation of Human Instructions into English

You are about to take a questionnaire in which you will be asked to assess whether two sentences are similar or different to each other. To do this, we will show you pairs of sentences extracted from novels and ask you to give each pair a score between 1 and 5. Use 1 to say that the two sentences are completely different from each other; Use 5 to say that they are almost the same. The other scores will be used to evaluate the intermediate cases.
Two sentences can be equal or different based on several elements. Here are some examples to help you in your evaluation.

Pairs of different sentences (score 1)

Examples: Please refer to the above section to see the original examples in Italian.

Pay particular attention to examples 2 and 3: although the sentences have words in common (like 'prince' and 'beer' in the examples) they are not necessarily the same!

Pairs of very similar sentences (score 5)

Examples: Please refer to the above section to see the original examples in Italian.

Examples 1 and 2 show sentences that not only contain many words in common but are also similar in terms of the scene described. In the third example, both sentences describe a person speaking quickly and decisively. We can say that the high similarity between the sentences in these examples is due to the fact that, except for a few details, they describe very similar scenes or images, even though they take place in different contexts.

3.2 Instructions for Large Language Models in Italian

First prompt for Anita - Similarity Scores: Assegna un punteggio di similarità su una scala da 1 (completamente diverse) a 5 (pressoché identiche) a questa coppia di frasi. Questi sono degli esempi:
-
Frase1: — domandò la signora Valentina, sempre con la medesima soavità nello sguardo e nella voce.
Frase2: — domandò Valentina Michailowna sempre con la stessa soavità di voce.
Similarità: 5
-
Frase1: Ivan non si lascia irretire nemmeno da migliaia di rubli.
Frase2: E le migliaia di rubli che passano per le sue mani!
Similarità: 2
-
Scrivi soltanto il punteggio numerico (minimo 1, massimo 5).

Second prompt for Anita - Explanations: Scrivi una spiegazione breve per il punteggio di similarità assegnato a questa coppia di frasi. Questi sono degli esempi:
-
Frase1: — domandò la signora Valentina, sempre con la medesima soavità nello sguardo e nella voce.
Frase2: — domandò Valentina Michailowna sempre con la stessa soavità di voce.
Similarità: 5
Spiegazione: Entrambe le frasi si riferiscono a una domanda fatta da Valentina, sottolineando la soavità della sua voce, con minime variazioni nei dettagli.
-
Frase1: Ivan non si lascia irretire nemmeno da migliaia di rubli.
Frase2: E le migliaia di rubli che passano per le sue mani!
Similarità: 2
Spiegazione: Entrambe fanno riferimento a migliaia di rubli, ma una si concentra sull'indifferenza mentre l'altra sulla quantità di denaro gestita.
-
Scrivi soltanto la spiegazione per il punteggio di similarità. Introduci la spiegazione con "Spiegazione:".

Single prompt for GPT-4o and Claude 3.7 Sonnet - Similarity Scores and Explanations: Questo file contiene 3 colonne: ir_ID, Sentence_1 e Sentence_2, dove ir_ID contiene l'ID della coppia di frasi. Assegna a ciascuna coppia di frasi un punteggio

di similarità da 1 (completamente diverse) a 5 (pressoché identiche), e una spiegazione breve tra virgolette ("") per il punteggio assegnato. Aggiungi due colonne: Similarità e Spiegazione. L'output deve essere in formato CSV, con i campi delle diverse colonne separati da virgola. Questi sono degli esempi:
-
1_2 , "— domandò la signora Valentina, sempre con la medesima soavità nello sguardo e nella voce", "— domandò Valentina Michailowna sempre con la stessa soavità di voce", 5, "Entrambe le frasi si riferiscono a una domanda fatta da Valentina, sottolineando la soavità della sua voce, con minime variazioni nei dettagli".
-
1_9, "Ivan non si lascia irretire nemmeno da migliaia di rubli.", "E le migliaia di rubli che passano per le sue mani!", 2, "Entrambe fanno riferimento a migliaia di rubli, ma una si concentra sull'indifferenza mentre l'altra sulla quantità di denaro gestita."
-

3.2.1 Translation of LLM Instructions into English

First prompt for Anita - Similarity Scores: Assign a similarity score on a scale from 1 (completely different) to 5 (nearly identical) to this pair of sentences.
Here are some examples:
Please refer to the above section to see the original examples in Italian.
Write only the numerical score (minimum 1, maximum 5).

Second prompt for Anita - Explanations: Write a brief explanation for the similarity score assigned to this pair of sentences.
Here are some examples:
Please refer to the above section to see the original examples in Italian.
Similarity: 5
Explanation: Both sentences refer to a question asked by Valentina, emphasizing the softness of her voice, with minor variations in detail.
-
Please refer to the above section to see the original examples in Italian.
Similarity: 2
Explanation: Both refer to thousands of rubles, but one focuses on indifference while the other on the amount of money handled.
-
Write only the explanation for the similarity score. Introduce the explanation with "Spiegazione:".

Single prompt for GPT-4o and Claude 3.7 Sonnet - Similarity Scores and Explanations: This file contains 3 columns: ir_ID, Sentence_1 and Sentence_2, where ir_ID contains the ID of the sentence pair.
Assign each pair of sentences a similarity score from 1 (completely different) to 5 (nearly identical), and a brief explanation in quotation marks ("") for the assigned score. Add two columns: Similarity and Explanation. The output must be in CSV format, with the fields of the different columns separated by commas. Here are some examples:
-
1_2, Please refer to the above section to see the original sentences in Italian, 5, "Both sentences refer to a question asked by Valentina, emphasizing the softness of her voice, with minor variations in detail."
-

1_9, Please refer to the above section to see the original sentences in Italian, 2, "Both refer to thousands of rubles, but one focuses on indifference while the other on the amount of money handled."
-

4. Acknowledgments

This paper is supported by the PRIN 2022 PNRR Project EKEEL - Empowering Knowledge Extraction to Empower Learners (P20227PEPK) funded by the European Union – NextGenerationEU, the project CHANGES - Cultural Heritage Innovation for Next-Gen Sustainable Society (PE00000020) project under the NRRP MUR program funded by the NextGenerationEU and by the PRIN 2022 Project TEAMING-UP - Teaming up with Social Artificial Agents (20177FX2A7) funded by the Italian Ministry of University and Research.

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agirre, Eneko, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In Mona Diab, Tim Baldwin, and Marco Baroni, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Alzetta, Chiara, Felice Dell'orletta, Chiara Fazzone, and Giulia Venturi. 2024. SimilEx: The first Italian dataset for sentence similarity with natural language explanations. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 20–28, Pisa, Italy, December. CEUR Workshop Proceedings.

Augenstein, Isabelle, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Basile, Pierpaolo, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*.

Belinkov, Yonatan and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Bosco, Cristina, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper, editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Brassard, Ana, Benjamin Heinzerling, Pride Kavumba, and Kentaro Inui. 2022. COPA-SSE: Semi-structured explanations for commonsense reasoning. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3994–4000, Marseille, France, June.

Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In Nicoletta Calzolari,

Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France, May. European Language Resources Association.

Camburu, Oana-Maria, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Cer, Daniel, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August. Association for Computational Linguistics.

Chen, Beiduo, Siyao Peng, Anna Korhonen, and Barbara Plank. 2025. A rose by any other name: LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10777–10802, Vienna, Austria, July. Association for Computational Linguistics.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognizing Textual Entailment, First Pascal Machine Learning Challenges Workshop (MLCW 2005)*, Southampton, UK, April. Springer.

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June. Association for Computational Linguistics.

Di Bonaventura, Chiara, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara Mcgillivray. 2024. Is explanation all you need? an expert survey on LLM-generated explanations for abusive language detection. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 280–288, Pisa, Italy, December. CEUR Workshop Proceedings.

Doshi-Velez, Finale and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *https://arxiv.org/abs/1702.08608*.

Fodor, James, Simon De Deyne, and Shinsuke Suzuki. 2025. Compositionality and sentence meaning: Comparing semantic parsing and transformers on a challenging sentence similarity dataset. *Computational Linguistics*, 51(1):139–190.

Gemini Team, Google, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

He, Xuanli, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. 2024. Using natural language explanations to improve robustness of in-context learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13477–13499, Bangkok, Thailand, August. Association for Computational Linguistics.

Hong, Pingjun, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2025. LiTEx: A linguistic taxonomy of explanations for understanding within-label variation in natural language inference. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34065–34085, Suzhou, China, November. Association for Computational Linguistics.

Huang, Shiyuan, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated

self-explanations. *arXiv preprint arXiv:2310.11207*.
Hurst, Aaron, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, Alexander J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
Jolliffe, Ian T. and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.
Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.
Lubos, Sebastian, Thi Ngoc Trang Tran, Alexander Felfernig, Seda Polat Erdeniz, and Viet-Man Le. 2024. Llm-generated explanations for recommender systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '24, page 276–285, New York, NY, USA, July. Association for Computing Machinery.
Mathew, Binny, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, Online, May. AAAI Press, Palo Alto, California USA.
Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July. Association for Computational Linguistics.
Meta AI. 2024. Llama 3 model card. `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`. Accessed: 2026-05-05.
Polignano, Marco, Pierpaolo Basile, and Giovanni Semeraro. 2026. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Scientific Reports*.
Rajani, Nazneen Fatema, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy, July. Association for Computational Linguistics.
Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, Hong Kong, China, November. Association for Computational Linguistics.
Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
Wang, Jiapeng and Yihong Dong. 2020. Measurement of text similarity: a survey. *Information*, 11(9):421.
Wang, Yuxia, Shimin Tao, Ning Xie, Hao Yang, Timothy Baldwin, and Karin Verspoor. 2023. Collective Human Opinions in Semantic Textual Similarity. *Transactions of the Association for Computational Linguistics*, 11:997–1013, 08.
Wiegreffe, Sarah and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. *arXiv preprint arXiv:2102.12060*.
Zaninello, Andrea, Sofia Brenna, and Bernardo Magnini. 2023. Textual entailment with natural language explanations: The italian e-RTE-3 Dataset. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, Venice, Italy, November 30 - December 2nd.
Zhao, Haiyan, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.